Driver cognitive workload - A comprehensive multi-measure approach

vorgelegt von
M.Sc.
Stefan Ruff
geb. in Torgau

von der Fakultät V - Verkehrs- und Maschinensysteme
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr.-Ing. Henning Jürgen Meyer
Gutachter 1: Prof. Dr.-Ing. Matthias Rötting
Gutachter 2: Prof. Dr. Mark Vollrath

Tag der wissenschaftlichen Aussprache: 28.08.2017

Berlin 2017

This thesis is dedicated to my grandmother Gitta Hintersdorf

♡

Acknowledgements

First and foremost I am grateful to my advisor Prof. Dr. -Ing. Matthias Rötting for the continuous support of my Ph.D study, for encouraging my research, for allowing me to grow as a research scientist and for providing all resources needed. I also would like to thank Prof. Dr. Mark Vollrath and his team at Technische Universität Braunschweig for their support in the second experiment of this thesis. Without the opportunity to conduct my experiment and the welcoming atmosphere at the Institut für Psychologie Ingenieur- und Verkehrspsychologie I would never have accomplished my goals.

I would also like to thank Prof. Dr.-Ing. Henning Jürgen Meyer for chairing the defense commitee.

My sincere thanks also go to Mario Lasch for his willingness to help and his amazing ability to find a solution to almost every technical problem. I also want to thank Oliver Hammerschmidt, Oliver Kroll, Otto Lutz, Lena Lüneburg, Ronja Gerdes and Doris Sonntag for their help before, during and after the experiments and all the people that participated in my experiments.

A special thanks goes to my fellow colleagues Antje Venjakob, Katja Karrer-Gauß, Felix Siebert and Stefan Damke for the stimulating discussions and for all the fun we have had in the last years.

Last but not the least, I would like to thank my family for their patience and support, Gabriele for encouraging me to walk this path in the first place, my friends Felix, Romy, Basti, Helene, Stefko, Roxana and all the others for cheering me up when needed and my business partners Antje and Klaas for their patience.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Berlin, den

Stefan Ruff

Abstract

Using in-car speech interaction systems has become increasingly popular. However, the cognitive workload resulting from this and its consequences for the driver are not yet fully understand. This thesis therefore aims to better understand the consequences of driver cognitive workload regarding performance, subjective evaluation and physiological changes and to identify suitable non-intrusive physiological measures that can be combined for the classification of a driver's cognitive workload. Only when consequences of workload are understood and quantified, it is possible to develop assistance systems that assess suboptimal driver states and, if necessary, initiate countermeasures in order to prevent dangerous situations. To this end two experiments were conducted that assessed and validated the sensitivity of multiple physiological measures to different levels of cognitive workload.

In the first experiment, thirty participants performed a lane change task (LCT), and parallelly to that, completed a paced auditory serial addition test (PASAT) in three conditions of varying degrees of difficulty. Results regarding performance and subjective measures confirmed that three different levels of cognitive workload were induced. Most of the ocular, cardiovascular and speech signal parameters exhibited sensitivity to the cognitive workload levels but lacked differential sensitivity at higher levels of cognitive driver workload. Only pupil size differentiated between all three levels of driver cognitive workload. In addition, eye-fixation parameters revealed strong dependencies of visual aspects of the driving task, so that their usefulness as indicators of cognitive workload is limited with respect to generalization.

The second experiment aimed at validating the results of the first experiment in a more realistic driving scenario. Ninety participants took part in a driving simulator study and were randomly assigned to one of three workload conditions. An auditory-verbal n-back task with three levels

(0-back, 1-back, 2-back) was used to manipulate workload. Again, performance as well as subjective workload measures confirmed that three workload levels were induced. Pupil size, horizontal fixation dispersion, heart rate, nose tip temperature, voice fundamental frequency and intensity were sensitive to the different workload levels, with pupil size and heart rate showing the largest effects. Pupil size and voice intensity revealed differential sensitivity between all three levels of workload, whereas the other measures failed to differentiate between medium and high levels of workload. The results indicate that physiological parameters in general are suitable to measure cognitive workload in the driving context. They also suggest that no single parameter is sufficient for a reliable classification of different levels of cognitive driver workload, as each parameter only reflects parts of the responses of the autonomous nervous system.

For this reason, a data-driven modelling approach was used to combine the most promising physiological measures for the classification of drivers' cognitive workload on the basis of the data gathered in the second experiment. The best performing model combined heart rate and pupil size and resulted in 92% classification accuracy for the differentiation between low and a combination of medium and high workload and 76% classification accuracy in the differentiation between all three workload levels.

Several challenges and methodological limitations, such as robust data collection and the demarcation of cognitive driver workload from other psychological states, have to be addressed to enable the use of physiological measures for reliable driver cognitive workload assessment in real-world driving. However, overall the results of this thesis suggest that physiological measures have great potential to be used for drivers' cognitive workload assessment.

Zusammenfassung

Die Verwendung von Sprachinteraktionssystemen im Fahrzeug wird immer beliebter. Allerdings sind die damit verbundene kognitive Beanspruchung, sowie die daraus resultierenden Konsequenzen für FahrerInnen, noch nicht vollständig verstanden. Nur wenn diese verstanden und quantifiziert werden können, ist es möglich Assistenzsysteme zu entwickeln, die suboptimale Fahrerzustände erfassen und gegebenenfalls Gegenmaßnahmen einleiten, um schwerwiegende Folgen zu verhindern. Deshalb zielt diese Dissertation darauf ab, die Konsequenzen der kognitiven Beanspruchung von FahrerInnen in Bezug auf die Fahrleistung, die subjektiv erlebte Beanspruchung und physiologische Veränderungen besser zu verstehen und geeignete, nicht-invasiv messbare physiologische Parameter zu identifizieren, die für die Klassifizierung der kognitiven Beanspruchung von FahrerInnen kombiniert werden können. Zu diesem Zweck wurden zwei Experimente durchgeführt, die die Sensitivität verschiedener physiologischer Parameter für unterschiedliche Niveaus kognitiver FahrerInnenbeanspruchung evaluieren und validieren.

Im ersten Experiment führten dreißig TeilnehmerInnen eine Spurwechselaufgabe (Lane Change Task - LCT) im Fahrsimulator aus. Parallel dazu wurde eine extern getaktete serielle Additionsaufgabe (PASAT) in drei unterschiedlichen Schwierigkeitsgraden bearbeitet. Die Ergebnisse hinsichtlich der Leistung und der subjektiv erlebten Beanspruchung bestätigten, dass durch die experimentelle Manipulation drei verschiedene kognitive Beanspruchungsniveaus induziert wurden. Die meisten der okularen, kardiovaskulären und Sprachsignalparameter wiesen eine generelle Sensitivität für die verschiedenen kognitiven Beanspruchungsniveaus auf. Allerdings zeigte sich, dass die Sensitivität bezüglich Unterschieden in höheren Beanspruchungsregionen limitiert ist. Nur die Pupillengröße unterschied zwischen allen drei Niveaus der kognitiven Beanspruchung der FahrerInnen. Darüber hinaus zeigten Fixationsparameter der Augenbewegung starke Abhängigkeiten von visuellen Aspekten der Fahraufgabe, so dass ihre

Nützlichkeit als Indikatoren kognitiver Beanspruchung in Bezug auf die Verallgemeinerbarkeit begrenzt ist.

Das zweite Experiment zielte darauf ab, die Ergebnisse des ersten Experiments in einem realistischeren Fahrszenario zu validieren. Neunzig TeilnehmerInnen nahmen an einer Fahrsimulatorstudie teil und wurden zufällig einer von drei kognitiven Beanspruchungsbedingungen zugeordnet. Eine auditiv-verbale n-back Aufgabe mit drei Schwierigkeitsstufen (0-back, 1-back, 2-back) wurde verwendet, um die Beanspruchung zu manipulieren. Wiederum bestätigten sowohl die Leistungsmaße wie auch Maße der subjektiv erlebten Beanspruchung, dass drei kognitive Beanspruchungsniveaus induziert wurden. Pupillengröße, horizontale Fixationsstreuung, Herzrate, Nasenspitzentemperatur, Grundfrequenz des Sprachsignals und Sprachintensität waren sensitiv für Unterschiede in der kognitiven Beanspruchung, wobei die Pupillengröße und die Herzrate die größten Effekte zeigten. Die Pupillengröße und die Sprachintensität diskriminierten dabei zwischen allen drei Beanspruchungsniveaus, während die anderen physiologischen Parameter nicht sensitiv für Unterschiede zwischen mittlerer und hoher Beanspruchung waren. Die Ergebnisse zeigen, dass physiologische Parameter generell dazu geeignet sind, die kognitive Beanspruchung im Fahrkontext zu messen, verdeutlichen aber gleichzeitig, dass kein einzelner Parameter für eine zuverlässige Klassifizierung der verschiedenen Niveaus der kognitiven Beanspruchung ausreicht, da jeder Parameter nur Teile der Reaktion des autonomen Nervensystems abbildet.

Aus diesem Grund wurde ein datengetriebener Modellierungsansatz verwendet, um die vielversprechendsten physiologischen Parameter zur Klassifizierung der kognitiven Beanspruchung von Fahrern zu kombinieren. Das beste Modell auf Basis der im zweiten Experiment gesammelten Daten, kombinierte Herzrate und Pupillengröße und erzielte eine Klassifikationsgenauigkeit von 92% für die Differenzierung zwischen niedriger kognitiver Beanspruchung und der Kombination von mittlerer und hoher kognitiver Beanspruchung. Die Klassifikation von allen drei Beanspruchungsniveaus resultierte in einer Klassifikationsgenauigkeit von 76%.

Insgesamt deuten die Ergebnisse dieser Dissertation darauf hin, dass physiologische Parameter großes Potenzial für die Erfassung des kognitiven

Beanspruchungsniveaus von FahrerInnen haben. Allerdings müssen zunächst Herausforderungen, wie die robuste Datenerfassung und die Abgrenzung der kognitiven Beanspruchung von anderen psychologischen Zuständen überwunden werden, um einen Einsatz im realen Fahrkontext zu ermöglichen.

Contents

List of Figures

Chapter 1

Introduction

Imagine you are driving a car in the late 80s. Think about what in-car technology could possibly withdraw your attention from the driving task and compare it with what the situation is like today. About 30 years ago, the only piece of technology in a car that could divert drivers from their actual task was the radio. Since those days our society has radically evolved to a high-tech information society. Not only that computers, smartphones and robots have become integral parts of the world, but also the technological possibilities, enabled through increasingly high computing power, impact our daily lives in a way that would not have been thinkable before. To better understand the dimension of this change Dr. Michio Kaku, a high-profil futurist and physicist, states that a modern cell phone *"... has more computer power than all of NASA back in 1969, when it placed two astronauts on the moon."*(Kaku, 2012, p.10). These developments also massively changed the way we drive today. Comparing, for example, the number of injured or killed traffic victims in Germany from the early 90s (1991: 516835) and today (2015: 396891) it is evident that driving a car, in general, has become very safe over the past decades even though the number of crashes involving motorized vehicles increased from 2.3 million to 2.5 million a year in the same time span (Destatis, 2017). The decreasing risk of getting injured in a traffic accident, despite the fact that the number of accidents increased can be explained by safety enhancing technological advances in the automotive industry. Due to the introduction of safety-related technologies to the mass market, e.g. airbags in 1980, ABS (Anti-lock Brake System) in 1985, ESC (Electronic Stability Control) in 1995 and ACC (Adaptive Cruise Control) in 1999, drivers get more and more assistance to avoid critical situations or to at least lower the impact of a crash. Every year the number of new advanced driver assistance systems (ADAS) like collision avoidance, lane departure warnings and blind spot monitoring increases. According to Statista

(2016a) the production volume of ADAS steadily increases accompanied by an annual revenue growth rate (2016-2020) of 33.97 % for the safety and driving assistance hardware industry. Besides those safety-enhancing technologies, a trend to more and more information and communication technologies available in the car apparently satisfy the drivers' growing need to stay connected while driving. Infotainment systems and smartphones changed the nature of driving widely, or better: the tasks a driver manages while driving. Especially smartphones have become indispensable in our daily lives. The smartphone penetration rate in North America and Western Europe reached 70% in 2015 and is expected to rise up to 80 % by the end of the decade (Statista, 2016b). A recent study on general smartphone usage by Montag et al. (2015) found a mean length of smart phone activity a day of 161.95 minutes. They concluded that smartphones highly disrupt our work life and social activities. With this in mind it is not surprising that smartphones are also extensively used while driving. According to latest surveys in the U.S. 61% of drivers use their smartphones to read, send or reply to text messages, 33% to read, send or reply to Emails, 28% to surf the internet and 27% to read or post on Facebook while driving (Statista, 2015). Therefore, in car secondary-tasks are no longer limited to tune to a radio station or talk to another passenger, but involve extensive amounts of attentional resources multiple times per drive. An interview study conducted in Germany by Huemer and Vollrath (2011) found that 80% of the 289 drivers they interviewed, engaged in one to three secondary-tasks in their last 30min of driving. This behavior is associated with tremendous risks for the driver and the traffic environment. A recent naturalistic study on accident causation (Dingus et al., 2016) reports that observable distractions (e.g. using a handheld devices) were present in 51.93% of non-crash driving and that distraction was apparent in 68.3% of 905 recorded crashes. They also found that the risk for accidents increases through interacting with a handheld cell phone by 3.6 (odds-ratio) compared to non-distracted driving. Secondary-tasks that required glances away from the road like handheld cell dialing, texting on a handheld cell phone and reaching for a handheld cell phone exhibited the highest increases in crash risk in the context of in-car phone use (odds-ratio of 12.2/ 6.1 /4.8). Visually demanding in-car secondary-tasks have long been identified and extensively studied as a potential risk for drivers and passengers as it diverts attention away from the road. Secondary-tasks that require manual and visual attentional resources directly interfere with the driving task itself and are therefore critical for safe driving (e.g. Tijerina, Parmer, & Goodman, 1998; Engström, Johansson, & Östlund, 2005; Horberry, Anderson, Regan, Triggs, & Brown, 2006; Drews, Yazdani, Godfrey, Cooper,

& Strayer, 2009). That is why research on workload and distraction used to focus predominantly on visual-manual secondary-tasks resulting in guidelines and laws trying to minimize the risks resulting from visual-manual secondary-tasks concurrent to driving (e.g. National Highway Traffic Safety Administration, 2012). Additionally, efforts are made to integrate smartphone functionality into the on-board infotainment systems and make the use of those functions operable in a relatively safe manner. Car manufacturers either introduce their own systems (e.g. Ford - Sync) or cooperate with the consumer electronics industry to achieve that goal (e.g. Apple Car Play, Android Auto, Amazon Alexa). Additionally, third party suppliers offer their own solutions like MirrorLink by the Car Connectivity Consortium. Independent of the solution, the goal is to reduce dangerous interaction scenarios where drivers have to take their hands off the wheel and their eyes off the road to make calls, write a message or check social media applications. This can be achieved by using speech interaction technologies that reduce visual and manual demands to a minimum. Speech recognition technology has made big steps and is now available in almost every new car. Another recent safety enhancing approach is to combine safety and communication requirements in advanced driver systems. These systems try to take the driver's workload into account in modern infotainment systems and thereby match secondary-task demands to the current driver state (e.g. Advanced Integrated Driver-vehicle InterfacE (AIDE) project Consortium (2004-2008)). Whereas the impact of visual-manual interactions on driver performance and crash risks is well know, there is less evidence on how demands resulting from speech interaction impact drivers. Even though the advantages over classical visual-manual interaction are clear, as the hands stay on the wheel and the eyes on the road, there are e.g. vastly differing results regarding the associated crash risk. McEvoy (2015) for example report a 3.8 fold increase of crash risk by using a hands-free phone whereas in a study by Klauer et al. (2014) no significant increase in risk was found. Therefore, more research is needed to understand the workload consequences of speech interaction systems. Pure speech interaction does not involve visual-manual attentional resources but auditory-verbal and central cognitive resources, which makes it most challenging to find suitable measures to assess the actual workload state of a driver. Tools are needed to continuously and non-intrusively assess the actual workload state of the driver, to detect potentially dangerous changes in the demand of driver's attentional resources while engaging in cognitive secondary-tasks.

So far no system has been implemented that detects the driver's workload state to prevent fatal accidents. There are a couple of reasons for that. First of all, up to now

no reliable indicators have been identified to detect workload changes while driving. Driving performance measures are often sensitive to workload increases but fail to identify increasing demands before performance itself declines. Moreover, research on cognitive workload indicates that high workload levels manifest in poor reactions to critical events rather than showing an impact on continuous driving parameters like steering angle, lane-keeping or speed (Horrey & Wickens, 2006; Caird, Willness, Steel, & Scialfa, 2008). They are therefore not suitable to prevent critical situations before they happen. Alternatively, physiological measures could be used to detect workload changes before performance declines. However, there has been research on the use of physiological parameters for almost four decades now without conclusive results. The reasons for that are manifold. First, there is the incapability of single physiological parameters to discriminate between more than two different workload levels. Some researchers have tried to combine different parameters to increase the sensitivity, but most often the measurement methods that are used are intrusive or at least require the attachment of electrodes (EEG, skin conductance etc.). In addition to that, measurement equipment has been extremely expensive and complicated to use. Therefore the use was limited to research and not applicable to every day driving. Secondly, most research has limited its focus on the physiological workload consequences without fully incorporating all other workload related consequences. Only if performance, subjective and physiological measures are combined, it is possible to develop workload sensitive systems based on physiological measures. Thirdly, machine learning approaches trying to combine multiple physiological measures for the classification of workload often only differentiate between two very different workload states. Even though it is of crucial importance to detect e.g. whether the driver's workload is extremely high or extremely low, the workload assessment in between these two extrema is of utmost importance for workload management systems that can react before overload happens, but does not produce high numbers of false alarms. Additionally, a great share of classification approaches used individual models for each driver that do not allow for a generalisation of those models to other drivers. Both aforementioned restrictions hinder the use of physiological workload indicators in real world applications.

1.1 Aims of the thesis

The above section highlights that up to now, no cognitive driver workload assessment system based on contactless physiological measurement methods exists that is

capable of differentiating between multiple workload levels and can be applied to real world driving. Therefore the aim of this dissertation is to explore which physiological measures are most suitable to be combined in a multi-measure approach of multiple cognitive workload levels. To accomplish this I examine traditional physiological measures. I also include measures that have not yet been studied thoroughly in the context of drivers' cognitive workload but are relevant due to recent technological advances. Based on literature and past research, physiological measures will be included that can in principle be collected without contact and have been shown to be sensitive to cognitive workload. To allow for a holistic interpretation of the physiological workload consequences, the first aim of my thesis is to review the concept of cognitive driver workload with respect to the interaction between physiological, performance as well as subjective measures. My second aim of the thesis is then to study the different physiological measures regarding their sensitivity to three different cognitive workload levels induced by an auditory-verbal secondary- task parallel to basic simulated driving. The third aim is to validate the most promising physiological measures in a more realistic driving scenario for higher ecologic validity, again incorporating all aspects of workload. The result of this step is a set of physiological features that can be used for the classification of three cognitive workload levels. The final aim of the thesis is to develop a subject-independent workload classification model on the basis of the physiological features that were identified before. To this end the results of different machine learning approaches will be compared and different feature subsets will be used to find the optimal solution regarding accuracy and effort to measure the features.

The central aims of this thesis can be summarized as follows:

1. Reviewing the concept of cognitive driver workload and the consequences regarding performance, subjective workload and physiology that are associated with it.

2. Identifying suitable physiological measures that are sensitive to more than two different cognitive workload levels and that can be captured non-intrusively.

3. Validating the most promising physiological measures in a more realistic scenario to enhance ecologic validity and test reliability.

4. Combining the most promising parameters using machine learning algorithms.

1.2 Structure of the thesis

The thesis is structured in eight chapters. The first chapter provides an introduction to the topic, the aims of the thesis are defined and the structure of the thesis is described (chapter 1). This is followed by a chapter (chapter 2) that defines mental workload in general and summarizes the underlying attentional resource models, the relationship between mental workload and performance as well as factors that affect mental workload to provide the reader with a basic understanding of the concept of mental workload. Furthermore, that chapter covers performance, subjective and physiological measurement methods commonly used to examine a person's mental workload. Chapter 3 conveys the theoretical considerations from chapter 2 to the context of drivers' cognitive workload. For this, first of all the driving task and auditory-verbal secondary-tasks are described with respect to the attentional demands that arise from the combination of both. This is followed by a review of the impact of drivers' cognitive workload on performance, subjective workload ratings and physiology with a particular focus on different aspects of driving performance (continuous vs. event-related) and the identification of non-intrusive physiological measures that are sensitive to multiple levels of cognitive workload.

The conclusions of the first three chapters form the basis of the following two experiments. Chapter 4 describes the first experiment which aims at testing the sensitivity of different physiological measures to multiple levels of cognitive workload by analyzing physiological responses to different levels of secondary-task difficulty induced by an auditory-verbal task. The results will be discussed with respect to the impact of drivers' cognitive workload on performance and subjective workload levels as well as the suitability of the physiological measures for multi-measure approaches. On the basis of this, suitable physiological measures will be selected for validation. The second experiment (chapter 5) subsequently tests the sensitivity of the selected physiological measures to multiple levels of workload in a more naturalistic environment comprising hazardous events. Again the impact of drivers' cognitive workload will be thoroughly discussed with a particular focus on physiological indicators.

The insights resulting from the two experiments form the basis for the following chapter 6 that first describes the methodology of physiological computing and examples of driver state classification based on those methods. Subsequently, models of cognitive workload classification are built on the basis of the physiological measures identified in chapter 4 and 5. The results are discussed with regard to classification

accuracy and compared to other studies dealing with the classification of cognitive workload based on physiological measures.

In a more general discussion (chapter 7) the results of the two experiments and the classification chapter will be discussed with respect to the aims of the thesis. The thesis ends with chapter 8 that reviews the most important findings of the thesis and provides an outlook regarding important topics in driver cognitive workload research.

Chapter 2

Mental workload

There is a surprisingly great confusion regarding the definition of driving related cognitive concepts like mental workload and attention. Attention is a concept that has been known for a long time but besides dictionary definitions, no common scientific definition is available. For workload a plethora of different definitions exist, depending on the area of research and the personal understanding of the researcher, often resulting in operational definitions for a specific study. As it is not the aim of this thesis to solve this problem, the next sections will be limited to common understandings regarding the concepts of workload. The author therefor focuses on topics that are practically relevant to measure cognitive workload, using physiological measures. First, mental workload will be defined in the context of this thesis (section 2.1) including different perspectives on attentional resources (subsection 2.1.1, 2.1.2 and 2.1.3), the relationship between performance and mental workload (subsection 2.1.4) as well as factors influencing mental workload of an individual (subsection 2.1.5). The subsequent section 2.2 will elaborate on the measurement of mental workload and therefore cover selection criteria for workload indicators as well as performance (subsection 2.2.1), subjective (subsection 2.2.2) and physiological (subsection 2.2.3) costs associated with mental workload.

2.1 Different aspects of mental workload

Mental workload is a field of study to which an extensive amount of research has been dedicated over the years. Depending on the context in which the research is carried out, the definitions of the concept differ. In their recent review on mental workload in ergonomics M. S. Young et al. (2015) state that for almost four decades researchers have investigated the concept in the context of ergonomics but up to now it still remains "*... one of the most nebulous concepts, with numerous definitions*

and dimensions associated with it." (M. S. Young et al., 2015, p.1). Nonetheless, M. S. Young and Stanton (2004, p.39-1) proposed the following global definition of mental workload "*The mental workload of a task represents the level of attentional resources required to meet both objective and subjective performance criteria, which may be mediated by task demands, external support, and past experience.*" The following section aims at providing a basic understanding of the relevant concepts for this thesis and is far from complete. For a more comprehensive review of mental workload in ergonomics M. S. Young et al. (2015) and Cain (2007) are recommended as they cover many aspects in more detail.

Regardless of which of the numerous definitions they adhere to, all studies on mental workload share some common understandings of the concept. The core of the concept is the assumption that mental workload is the result of the interaction between task demands and limited mental resources. When task demands exceed the available mental resources, operators have to compensate for it or performance will decline. This can e.g. be accomplished by investment of compensatory mental effort but is limited by the operator's overall attentional capacity (G. Mulder, 1986; Hockey, 1997; M. S. Young et al., 2015). The understanding of the nature of these mental resources differs. However, the different assumptions complement rather than exclude each other. On the one hand mental resources can be understood as structural attentional resources (Kahneman, 1973; Wickens, 1980, 2002) and on the other hand as energetical resources influencing the efficiency of information processing stages (Pribram & McGuinness, 1975; G. Mulder, 1986; Hockey, Gaillard, & Coles, 1986). The following subsections will elaborate on the two concepts and their implications for the understanding of mental workload.

2.1.1 Structural attentional resource models

Attentional resources are limited. Whether this limit is fixed or to a certain degree adaptive is still controversial (M. S. Young & Stanton, 2002b). Moreover, attentional resources can either be of single or multiple nature. Single resource theories e.g. by Moray (1967) or Kahneman (1973) describe the attentional system as one processor with limited capacity that can allocate resources between different tasks. Multiple resource theories e.g. by Navon and Gopher (1979) or later Wickens (1980, 2002, 2008) however proposed the concept of specific resources for specific functions, i.e. modalities, to explain good performance results when multiple resources were employed (e.g. Manzey, 1988). Such findings are in conflict with a single resource assumption, as the amount of resources needed to complete all tasks should exceed the capacity of a

single resource. The perfect completion of multiple tasks would be impossible if the same task load involved only one modality, unless the concurrent tasks would be fully automated and therefore would not demand additional attentional resources. Despite assuming the existence of different, distinct resources, most authors of multiple resource theories do not reject the idea of a central processing unit shared by different tasks (Wickens, 1991, 2008). Wicken's multiple resource model (2002, 2008) is shown in Figure 2.1. It discriminates between four dimensions that can have an impact on dual-/multi-task performance: 1. *stages of processing*, 2. *code of processing*, 3. *modalities* and 4. *visual channels*.

Stages of processing differentiate resources needed for perception, cognition and response. The perception stage as well as the cognition stage (e.g. involving working memory) share the same resources whereas response selection and execution is separated from the first two stages (Wickens, 2002). The *code of processing* distinguishes between resources needed for spatial vs. verbal activity over the stages of processing. *Modalities* differentiate between resources needed for the perception of auditory vs. visual stimuli and *visual channels* distinguishes between focal and ambient perception, hence between resources within visual perception. If two or more tasks share common resources on these dimensions, they will interfere more than tasks that do not overlap. For example, an auditive task like listening to the radio will interfere less with a visual-manual task like driving a car than another visual-manual task like manipulating the radio to tune to a new radio station. Still that does not mean that the radio listening task will not interfere with the driving at all as limited higher central processing may be involved. One "bottleneck" in this context could be the central executive of working memory, thought of as a modality unspecific central managing unit controlling modality specific subunits (Baddeley & Hitch, 1974; Baddeley, 1992, 2003).

The idea of discrete multiple resources that are not completely independent from each other is supported by neuro-structural research showing that the attention system is a distinct network of multiple brain structures responsible for specific tasks like e.g. orientating to stimuli, maintaining an alert state and executive control. Although those structures are locally separated they interact with each other (Posner & Petersen, 1990; Posner & Fan, 2008; Petersen & Posner, 2012).

2.1.2 Energetical resource models

Besides the theories of structural attentional resources the idea of energetic resources is of great relevance for the understanding of mental workload. Early research by

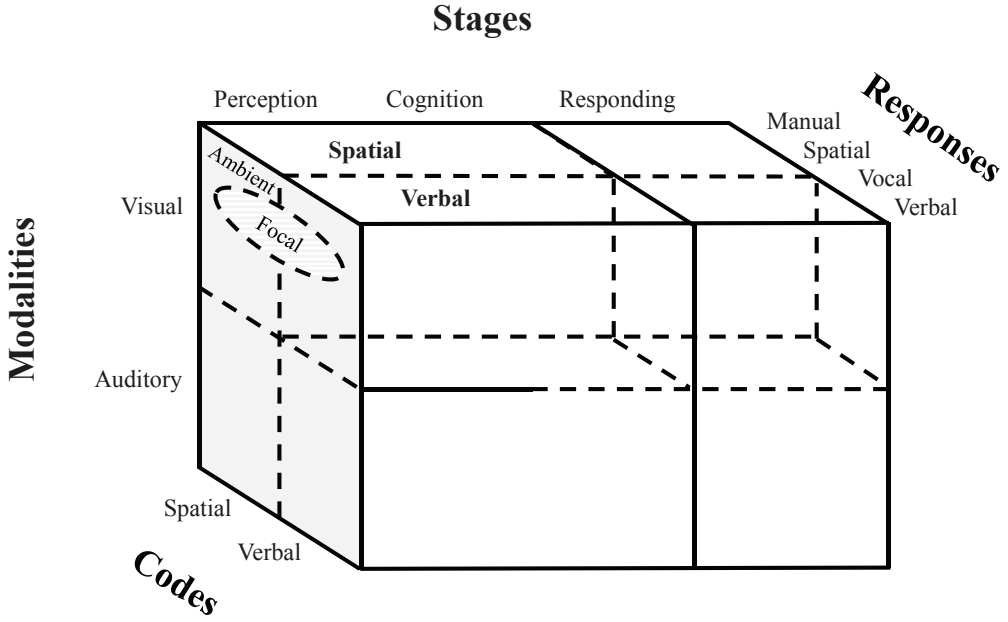


Figure 2.1: Wickens Multiple Resource Model adapted from Wickens (2008)

Yerkes and Dodson (1908) suggested an inverted U-shaped relationship between physiological arousal and performance. Optimal arousal levels in the sense of optimal performance prerequisites differ depending on the task difficulty, constituting higher arousal levels for higher task difficulties (Wickens & Hollands, 2000). Pribram and McGuinness (1975) introduced the concept of three interacting but separate neural systems responsible for the control of attention: an arousal system reactive to stimuli intensity, an activation system for the preparation of motor responses and a superior effort system coordinating the other two. Two types of effort are often distinguished. Effort can be understood as energetic foundation for central information processes (computational, passive or processing effort) and/or as a voluntary compensatory function (compensatory effort) that adjusts to task demands (Sanders, 1983; G. Mulder, 1986; Hockey, 1997). Computational effort can be described as an unspecific energetical basis for information processing structures that enhances the signal-to-noise ratio (G. Mulder, 1986). Compensatory effort, on the other hand, is thought of as an active voluntary mechanism to protect performance in highly demanding situations through the mobilization of additional resources (Hockey, 1997). Both types of effort are associated with physiological costs reflected in autonomic changes that can be used to determine the amount of effort invested (G. Mulder, 1986; L. Mulder, 1992; Fairclough & Houston, 2004; Radulescu, Nagai, & Critchley, 2015).

2.1.3 Combination of structural attentional and energetic resource models

In their attempt to combine energetic and attentional resource theories, Sanders (1983) picks up the idea of arousal, activation and effort as an energetic foundation by Pribram and McGuinness (1975) affecting the efficiency of four different (structural attentional) processing stages: stimulus preprocessing, feature extraction, response choice and motor adjustment (see figure 2.2). Arousal is thought of as being responsible for stimulus intake, passively depending on stimulus intensity or actively on attention allocation. Activation is responsible for the preparation of motor responses and effort directly affects central processes and acts as a compensatory function of the subsystems of arousal and activation. Overall the interplay between those mechanisms results in different activation patterns reflected in autonomous nervous system (ANS) responses, depending on the attentional demands put on an individual and the resources invested by the individual.

Hockey (1997) summarizes the combination of the two different views on attentional resources by assuming that mental workload arising from concurrent task performance is affected by the competition of those tasks for one or more task specific and capacity limited pools of structural attentional resources (see e.g. Wickens, 1980, 2002, 2008). This competition for limited resources from an energetical resource perspective, is on the one hand associated with computational or processing effort as well as compensatory effort to cope with increasing task demands (see e.g. G. Mulder, 1986; Hockey, 1997). Thus to understand and measure mental workload both resource models are crucial. The structural attentional models, to infer which tasks will compete for which resources and the energetical concepts to measure and interpret the physiological costs associated with different levels of mental workload.

2.1.4 Mental workload and performance

One of the main objectives of mental workload research in ergonomics is to evaluate how different levels of mental workload affect operators' performance (M. S. Young et al., 2015). The limited capacity assumption comprises the idea of a relationship between the quantity of invested resources and performance (Kahneman, 1973). According to M. S. Young et al. (2015) workload levels can roughly be divided into three areas with different implications for performance as illustrated in Figure 2.3. Similar to Yerkes and Dodson (1908) it is assumed that optimal levels of workload exist where performance is optimal. Additionally, there are areas of mental overload

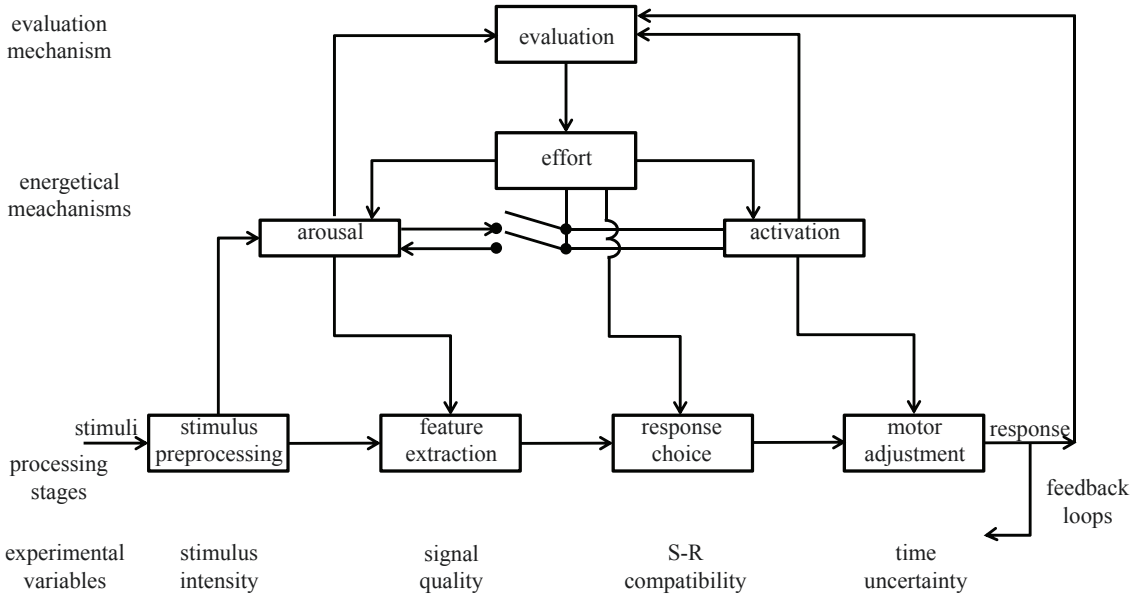


Figure 2.2: Cognitive energetic model adapted from Sanders (1983, p.79)

and underload. If workload is low, an increase of mental workload will lead to better performance whereas in areas of very high workload, a further increase in workload will lead to performance decrements (De Waard, 1996; M. S. Young et al., 2015). Especially the areas of over- and underload are of great importance, as they have been shown to result in degraded performance due to distraction and inadequate information processing respectively reduced alertness and lowered attention (De Waard & Brookhuis, 1997). Mental overload occurs when the operator is not able to compensate for increasing task demands by effort investment to maintain performance (De Waard, 1996). Underload on the other hand is not yet well defined but best described as the operator's state where task demands are below a certain level and an inappropriate amount of resources is deployed in relation to the demands. Whether this occurs because of a shrinkage of attentional resources (M. S. Young & Stanton, 2002a) or because too little voluntary effort is invested in reaction to a misperception of the demands (M. S. Young et al., 2015) remains unclear.

2.1.5 Factors affecting mental workload

Mental workload can be affected temporarily by factors like fatigue (e.g. Hancock & Verwey, 1997), exceptional emotional states (e.g. Myrtek et al., 1994), drugs (e.g. Brookhuis, de Waard, & Samyn, 2004) etc. or more constitutional factors like working memory capacity (e.g. Ross et al., 2014), age (e.g. Cantin, Lavallière, Simoneau, &

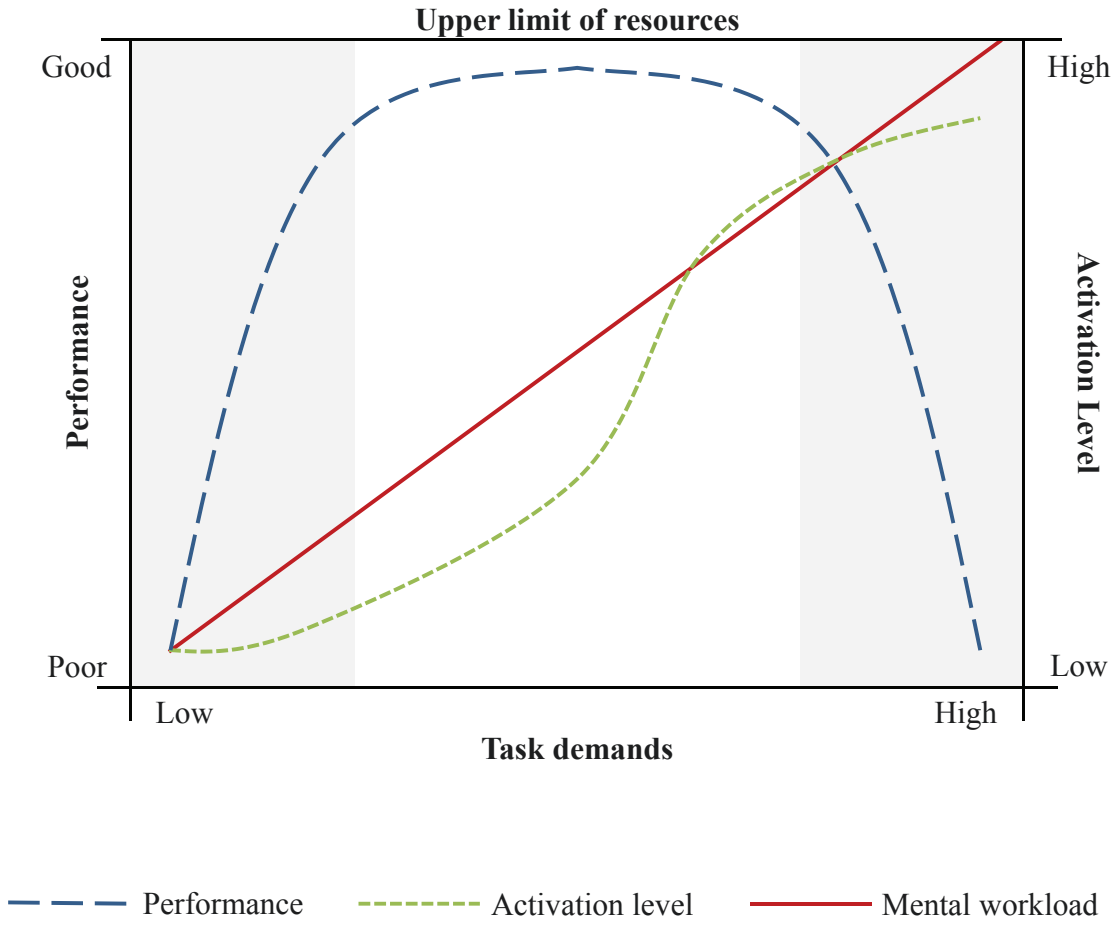


Figure 2.3: Model of performance, activation and mental workload adapted from M. S. Young et al. (2015)

Teasdale, 2009), personality (e.g. G. Matthews & Campbell, 2009) and diseases (e.g. Baddeley, Bressi, Della Sala, Logie, & Spinnler, 1991; Baddeley, Chincotta, & Adlam, 2001).

The workload definition described in section 2.1 also contains the idea that effective performance criteria may vary between individuals but may also depend on internal goals (M. S. Young & Stanton, 2004). Even though a task seems to be objectively definable by its characteristics (time constraints, input/output modalities, difficulty etc.), the same task does not necessarily put the same demand on different individuals. Individual knowledge, skills or past experience may influence the objective task demands. For a highly trained operator the same task will demand less resources compared to an untrained operator as with practice parts of the task or the complete task can be performed automatically with less attentional resource investment (Schneider & Detweiler, 1988; Xie & Salvendy, 2000a; Ayaz et al., 2012). Additionally, the definition of mental workload by M. S. Young and Stanton (2004) features external support as a mediating factor of workload. Such support can e.g. be technological or other external instances aiding the individual.

2.2 Measurement of mental workload

The measurement of mental workload aims to quantify the amount of cognitive resources needed to perform a task properly (Cain, 2007). Workload assessment techniques are generally classified in three main categories, consisting of physiological, subjective, and performance measures (O'Donnell & Eggemeier, 1986; Wierwille & Eggemeier, 1993; Cain, 2007). Each of the aforementioned categories of measures captures different aspects of workload (Cain, 2007). Whereas performance measures reflect the costs to performance efficiency induced by suboptimal workload levels, subjective measures give insights in the operators consciously perceived workload (Gopher, 1994). Physiological measures offer the opportunity to continuously asses the bodily reactions to different levels of mental workload. Most ideally a mixture of all approaches towards measuring workload is used in experimental research to explore all aspects of task demands. However, in applied contexts this is often unrealistic.

Based on the definitions by O'Donnell and Eggemeier (1986) and Eggemeier, Wilson, Kramer, and Damos (1991), G. Matthews, Reinerman-Jones, Barber, and Abich (2015, p.126) defined six criteria to evaluate mental workload measures (see table 2.1). According to G. Matthews et al. (2015) the criteria sensitivity, diagnosticity and selectivity are most important to the validity of workload measures. Nevertheless, the

Table 2.1: Workload instrument evaluation criteria according to Matthews et al. 2015 (p.126)

Criterion	**Description**
Sensitivity	Capacity of the instrument to detect changes in task difficulty or cognitive demands
Diagnosticity	Capacity of the instrument to differentiate distinct sources of workload, such as specific capacities or multiple resources
Selectivity/Validity	Sensitivity of the instrument only to differences in cognitive demands, not to changes in other variables (e.g., physical workload, stress)
Reliability	Consistent assessment of mental workload
Intrusiveness	Lack of interference with primary-task performance
Implementation requirements	Practical constraints associated with instrumentation, software and training
Operator acceptance	Operator perception of the validity and usefulness of the procedure

consideration of all listed criteria is necessary particularly in applied settings. Ideally, a measure is sensitive to variable cognitive demands but insensitive to confounding influences like physical or emotional factors. This is referred to as selectivity. Furthermore, it should differentiate between the structural resources needed to cope with the task demands. Additional factors like reliability, intrusiveness with the primary-task, practical considerations for the implementation as well as operators' acceptance of a workload measure also need to be taken into account. For further elaborations on selection criteria please see O'Donnell and Eggemeier (1986), Wierwille and Eggemeier (1993) and G. Matthews et al. (2015).

2.2.1 Performance measures

Primary-task measures reflect performance characteristics crucial for safe, i.e. errorless performance. This is often operationalized through behavioral measures like e.g. reaction times, number of errors and accuracy. These measures allow for a continuous and interference-free assessment but are mostly not diagnostic as the same perfor-

mance decrements can be the result of structurally different task demands (Vidulich & Tsang, 2012). Furthermore it is desirable to detect workload increases before primary-task performance decreases, which makes it impossible to use those measures for prediction purposes in applied settings. Additionally, the *Subsidiary Task Paradigm* is often used in mental workload assessment. A mostly artificial concurrent secondary-task is hereby added to the primary-task. Participants are then instructed to prioritize the primary-task at the cost of secondary-task performance. Performance decrements in the secondary-task should consequently reflect the spare capacity available to the operator under the assumption that the amount of available resources is fixed over different task demand levels (O'Donnell & Eggemeier, 1986). Considering the existence of multiple structural resources, it is also crucial that primary- and subsidiary-task demand the same resources. If not so, performance results do not provide insights into the primary-tasks resource demands, as the task then does not interfere and can, at least in part, be executed concurrently (M. S. Young & Stanton, 2004).

2.2.2 Subjective measures

Subjective measures of mental workload are rating scales assessing the individually perceived workload of the operator. Most prominent examples are multidimensional instruments like NASA-TLX (Hart & Staveland, 1988) and SWAT (Reid & Nygren, 1988) or unidimensional questionnaires like the RMSE scale (Zijlstra, 1993). In general, they are easy to implement and in the case of multidimensional questionnaires, the impact of different sources of workload can be inferred. Furthermore subjective ratings of the amount of effort that needs to be invested can be a sensitive indicator of workload changes in workload where performance is unaffected (M. S. Young & Stanton, 2004). However NASA-TLX e.g. requires a time consuming process of pairwise ratings to individually apply weights to the subdimensions for the calculation of the overall workload. Additionally, these instruments interfere with the task itself, which is why they are most often applied subsequent to the task. This can result in biased ratings by the operators if time between task and rating are long or the task to be rated is stretched over long time periods (Wierwille & Eggemeier, 1993; M. S. Young & Stanton, 2004). For a detailed review on subjective workload questionnaires please see Rubio, Díaz, Martín, and Puente (2004).

2.2.3 Physiological measures

Physiological measures are highly associated to computational and compensatory mental effort, reflecting attentional resources invested to cope with the task demands (M. Young & Stanton, 2004). With increasing task demands more effort is needed to maintain performance, which is e.g. reflected in autonomic nervous system (ANS) responses (Kahneman, Tursky, Shapiro, & Crider, 1969; G. Mulder, 1986; Hockey, 1997). The autonomous nervous system is part of the peripheral nervous system (PNS) and its main function is the regulation of the body's internal milieu (Gramann & Schandry, 2009). It can be subdivided into a sympathetic and a parasympathetic system. Generally, sympathetic activation mobilizes energy in demanding or arousing situations and parasympathetic activity mostly operates to conserve energy with the purpose of bodily relaxation (Pinel, 2000). ANS innervated organs are the inner organs, smooth muscles, glands, blood vessels and the skin. Almost all autonomically innervated organs are under the control of both subsystems with the exception of peripheral arterioles that are solely innervated sympathetically (Gramann & Schandry, 2009). The most common pattern of workload invoked sympathetic and parasympathetic activity is reciprocally coupled sympathetic activation accompanied by parasympathetic withdrawal. However, research has also shown that different modes of activation vs. withdrawal and coupled vs. uncoupled sympathetic and parasympathetic responses occur (Berntson, Cacioppo, & Quigley, 1991, 1993; Berntson et al., 1994), depending on the task demands (see e.g Lenneman & Backs, 2009).

To capture changes in the ANS resulting from mental workload, several methods have been proposed and successfully tested. The following section will focus on non-intrusive measures that can in principle be measured without direct contact to the body. This precondition is chosen as it is highly unlikely that e.g. drivers or pilots will attach electrodes to their bodies in everyday real world conditions. For a complete review of physiological measures of mental workload in ergonomics research please see e.g. Kramer (1991), De Waard (1996), Brookhuis and Waard (2010) and Borghini, Astolfi, Vecchiato, Mattia, and Babiloni (2014)).

Cardiovascular measures

In general the heartbeat is generated by internal structures of the cardiac conduction system, but moderated by sympathetic and parasympathetic activity (Gramann & Schandry, 2009, p.102). Sympathetic innervation leads to a higher frequency of

heartbeats whereas parasympathetic innervation results in lower heartbeat frequency. Under resting conditions with low load the heart is dominantly under parasympathetic control whereas under higher load conditions sympathetic activity dominates (Gramann & Schandry, 2009, p.102). Robinson, Epstein, Beiser, and Braunwald (1966) showed that while resting, the heart rate does not increase due to a blockade of sympathetic activity but with parasympathetic blockade heart rate increases significantly. Under heavy exercise conditions, however, sympathetic activity dominantly contributes to an increase in heart rate whereas parasympathetic withdrawal contributes to a lesser amount.

Heart rate and heart rate variability

One of the most popular methods to measure cardiac activity is the electrocardiogram (ECG). Heart rate (HR) and heart rate variability (HRV) are commonly used indicators derived from it. Figure 2.4 shows an example of a typical ECG course over time. Heart rate is calculated as the number of R-waves (indicated by a red cross) per time unit, usually as beats per minute (bpm). For the calculation of HRV, the distance between consecutive R-waves is calculated. These intervals are known as Inter-Beat-Intervals (IBI) or as the interval between two R-waves (RR). Measures of the variability of the IBIs in the time and frequency domain can then be interpreted as HRV. Figure 2.5 shows an example of a typical summary of HR and HRV indicators in time and frequency domain using Kubios Software (Tarvainen et al., 2014). Typically the power of different frequency bands of the IBI power spectrum is thought to reflect different aspects of ANS activity. The high frequency component (HF: 0.15 - 0.4Hz) is consistently associated with parasympathetic modulation whereas the low frequency component (LF: 0.04 - 0.15) is thought to represent parasympathetic as well as sympathetic modulations (Sztajzel, 2004). Some authors, however, suggest that the LF component predominantly reflects sympathetic dominance (Berntson et al., 1997; Billman, 2011). Nevertheless, recent studies showed that the use of the LF/HF ratio does not reflect the balance between sympathetic and parasympathetic activity (Billman, 2013).

In general, HR increases and HRV decreases with increased mental effort, especially when working memory demands are high. This is assumed to be a result of increased sympathetic activity (Kahneman et al., 1969; Jorna, 1992; L. Mulder, 1992; Wilson, 1992; L. B. Mulder, de Waard, & Brookhuis, 2004). Changes in heart rate related to increased mental effort often occur with short temporal latency (i.e. within seconds) to the onset of the psychological changes. HRV in the frequency domain

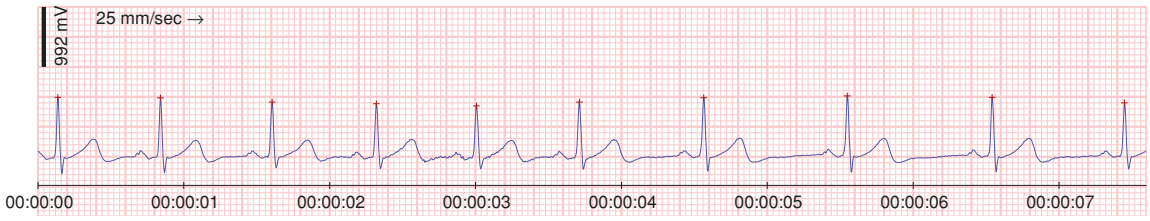
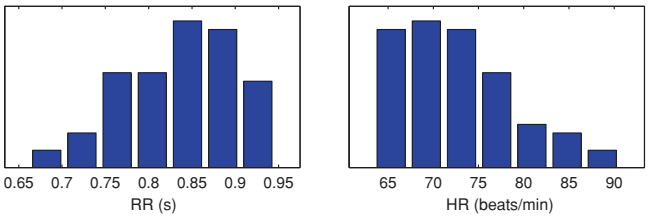


Figure 2.4: Raw ECG signal over time

Time–Domain Results

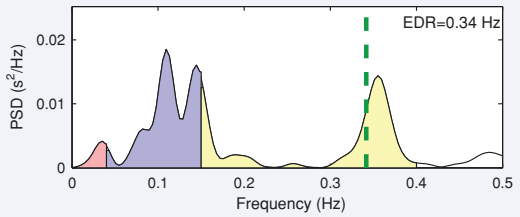
Variable	Units	Value
Mean RR*	(ms)	835.0
STD RR (SDNN)	(ms)	47.2
Mean HR*	(1/min)	72.34
STD HR	(1/min)	5.04
RMSSD	(ms)	47.8
NN50	(count)	21
pNN50	(%)	30.0
RR triangular index		10.143
TINN	(ms)	195.0

Distributions*



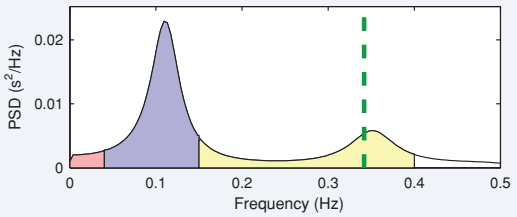
Frequency–Domain Results

FFT spectrum (Welch's periodogram: 256 s window with 50% overlap)



Frequency Band	Peak (Hz)	Power (ms²)	Power (%)	Power (n.u.)
VLF (0–0.04 Hz)	0.0352	74	4.2	
LF (0.04–0.15 Hz)	0.1094	900	50.9	53.1
HF (0.15–0.4 Hz)	0.3555	792	44.8	46.8
Total		1768		
LF/HF		1.135		

AR Spectrum (AR model order = 16, not factorized)



Frequency Band	Peak (Hz)	Power (ms²)	Power (%)	Power (n.u.)
VLF (0–0.04 Hz)	0.0391	86	4.6	
LF (0.04–0.15 Hz)	0.1094	1148	61.2	64.1
HF (0.15–0.4 Hz)	0.3516	638	34.0	35.6
Total		1877		
LF/HF		1.799		

Figure 2.5: Spektral analysis of IBI with Kubios Software by Tarvainen et al. (2014)

seems to be more sensitive to changes in mental workload compared to the variability in time domain (De Waard, 1996). The most commonly used indicator in the frequency domain, i.e. 0.1 Hz component, reliably showed to decrease with increased memory load and subjective effort ratings (see for review Kramer, 1991; Jorna, 1992). However, as the valid interpretation of spectral features of HRV requires long measurement intervals (Aasman, J., Mulder, G., & Mulder, L.J.M., 1987), the potential for online use of HRV to detect changes in mental workload is limited. Both measures are subject to several confounding influences. HR e.g. is not only influenced by mental effort but also influenced by respirational patterns as well as physical and emotional strain (L. Mulder, 1992; Nakamura, Yamamoto, & Muraoka, 1993; Brosschot & Thayer, 2003). HRV on the other hand is highly affected by even small measurement artifacts in the IBI time series (Jorna, 1992; L. B. Mulder et al., 2004). Those artifacts can either arise from simple measurement errors due to imprecision of the technical equipment or from changes in e.g. breathing patterns (Veltman & Gaillard, 1996). Hence, the use of HRV in real-time applications is highly debatable (Beda, Jandre, Phillips, Giannella-Neto, & Simpson, 2007) and is therefore not further considered in this thesis. Cardiovascular parameters are mostly measured by attaching electrodes to the body, which is not desirable in applied settings. Recent technological advancements, however, introduced new methods suitable to non-intrusively measure heart activity. Wartzek et al. (2011), for example, successfully tested a system that uses sensors integrated into the driver seat. Furthermore, systems that use video-based detection of color changes in the human face have become extraordinarily advantageous (Poh, McDuff, & Picard, 2010).

Skin temperature

In addition to the aforementioned measures of heart activity, blood flow in peripheral blood vessels of the skin can be used as an indicator of mental workload. In contrast to the heartbeat, vasodilation and vasoconstriction are almost exclusively controlled by the sympathetic nervous system (Drummond, 1994; Rimm-Kaufman & Kagan, 1996). In resting conditions the tonic sympathetic tone keeps the vessels at approximately half their maximal diameter (Iani, Gopher, & Lavie, 2004). Sympathetic activity leads to an increase in vasoconstriction accompanied by a decrease in blood flow, and sympathetic withdrawal to an increase in vasodilation accompanied by an increase in blood flow through cutaneous blood vessels. Research on mental workload by e.g. Iani et al. (2004) and Miyake et al. (2009) showed decreases in blood flow volumes with higher workload levels. The change in blood flow can e.g. be

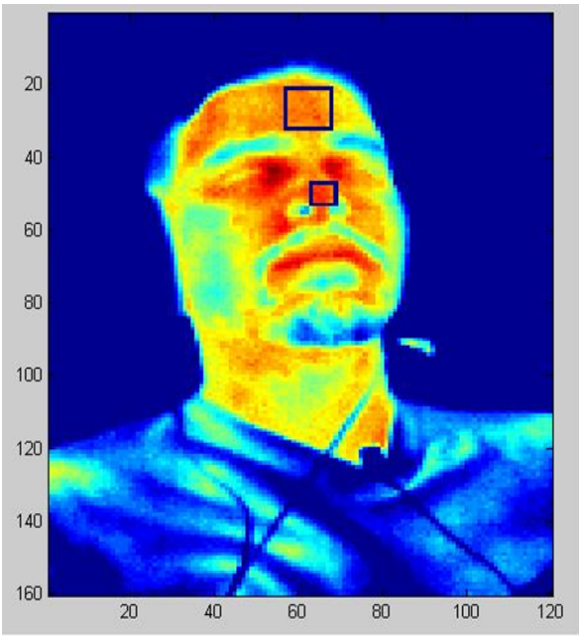


Figure 2.6: Thermal image captured with a Testo 875-1 thermography camera. The rectangles illustrate facial areas at the tip of the nose and the forehead that can be used for workload measurement

captured by laser-doppler flowmetry (Holloway & Watkins, 1977; Shepherd & Öberg, 2013). However, changing blood flow volumes are secondarily accompanied by skin temperature changes (Rimm-Kaufman & Kagan, 1996; Iani et al., 2004; Nakayama, Goto, Kuraoka, & Nakamura, 2005). Sympathetic activity decreases blood flow volumes, resulting in a drop in cutaneous temperature. Areas suitable to capture those changes are regions with a high density of blood vessels in the face like the nasal tip or periorbital areas as well as finger tips. The traditional way of measuring skin temperature is to apply sensors to the region of interest. More recently the use of thermography cameras gained a lot of attention, especially in medical diagnostics of e.g. inflammatory arthritis, osteoarthritis, soft tissue rheumatism, complex regional pain syndrome and malignant diseases like breast cancer (Ring & Ammer, 2012). Thermal imaging techniques come with the advantage of being non-intrusive and highly accurate. Thermography cameras capture the amount of infrared radiation emitted by the human skin and transform it to a thermal image by combining the radiation information with the video image (see an example of a thermography image in Figure 2.6).

Even though skin temperature as mental workload measure has not been studied extensively so far, a small number of studies showed that the temperature at the tip of the nose decreases with cognitively demanding tasks (Veltman & Vos, 2005; Duffy,

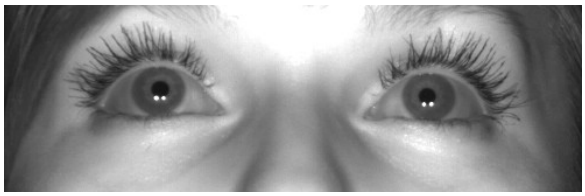


Figure 2.7: Eye-tracking eye image

2007; Kang & Babski-Reeves, 2008; Itoh, 2009). Others successfully used thermal imaging to capture perspiration in the perinasal area to assess subjects sympathetic state (Shastri, Merla, Tsiamyrtzis, & Pavlidis, 2009; Pavlidis et al., 2016). Like heart activity measures, skin temperature changes are also influenced by stress and fatigue (Genno et al., 1997) as well as emotions (Nakayama et al., 2005).

Ocular measures

Two different categories of ocular measures are used for the assessment of mental workload: characteristics of eye-movements and ANS related changes of the eye. Both can be captured contactless by remote video-based eye-tracking. Please see Holmqvist et al. (2011, p. 9-64) for a detailed review of eye-tracking hardware and properties and Hansen and Ji (2010) for a recent survey on eye and gaze models. The basic working principle of most eye-tracking systems is the use of computer vision techniques to determine the center of the pupil and the location of infrared reflections on the cornea (see Figure 2.7), induced by infrared Light Emmiting Diodes (LEDs). After calibrating the system to the human eye, the knowledge of those two locations allows for the reliable estimation of the point of gaze (Holmqvist et al., 2011, p. 24-25).

Eye-movements

Eye-movements provide insights into the distribution and allocation of visual attention during task execution and are highly task dependent. For a detailed overview of eye-movement definitions and measures please see Rötting (2001) and Holmqvist et al. (2011). Fixations are most commonly used in mental workload assessment. A fixation can be defined as a *"... state, in which the eye is in relative stasis to its object of interest"*(Rötting, 2001, p.68). According to the influential eye-mind hypothesis, information which is fixated is also processed (Just & Carpenter, 1980). Although there is evidence that visual attention precedes the actual fixation (Deubel, 2008), it can be assumed that the correlation between the location and time of a fixation and

the focus of attention is high. The most prominent fixation characteristics are fixation duration and fixation count and they are highly related. If the mean fixation duration increases from one time interval to the next equally long interval, than eventually the fixation count has to decrease relatively. It is generally implicated that longer fixation durations are related to higher cognitive task difficulties, whereas when the information necessary to perform a task is visually complex, fixation durations decrease (see Holmqvist et al. (2011, p.381-383) for a detailed review of task influences on fixation durations). Furthermore indices of fixation distribution like horizontal and vertical gaze position (e.g. Recarte & Nunes, 2000, 2003) and the randomness of fixation patterns (e.g. Di Nocera, Camilli, & Terenzi, 2007) are used as mental workload measures in applied contexts. However, the use of eye-movements in mental workload assessment is highly dependent on the task as the famous experiment by Yarbus (1967) suggests and should therefore be interpreted with care.

Pupil

The second category of ocular measures includes parameters that are not under voluntary control but are under the control of the autonomous nervous system (ANS). The most prominent of these measures is pupil diameter. Pupil diameter has long been known to increase with memory load and problem solving tasks (Hess & Polt, 1964; Kahneman & Beatty, 1966; Beatty, 1982). The regulation of the size of the pupil under no-load conditions results from opposing muscle activity of the sympathetically innervated *musculus dilator pupillae*, responsible for the dilation of the pupil and the parasympathetically innervated *musculus sphincter pupillae*, responsible for pupillary constriction (Schmidt & Schaible, 2006; Steinhauer, Siegle, Condray, & Pless, 2004). Steinhauer et al. (2004) examined the contribution of both processes to the increase of pupil diameter under sustained mental processing. Whereas sympathetic activation defines the maximum pupil size (Schmidt & Schaible, 2006), inhibitory parasympathetic activity seems to be sensitive to different levels of task difficulty. In general, pupil dilation due to mental workload is observed regardless of task modalities. Klingner, Tversky, and Hanrahan (2011) suggest that the magnitude of the pupil diameter is higher for aural presentation of different tasks than for visual presentation, suggesting higher workload levels for auditorily presented tasks. Generally, mental workload related changes in pupil size ($<= .5mm$) are much smaller than luminance induced changes (Beatty & Lucero-Wagoner, 2000; Laeng, Sirois, & Gredebäck, 2012). Although pupil diameter as a workload indicator has been shown to be sensitive to different task loads, the use in applied settings might be limited since it

is highly influenced by changing light conditions (Wyatt & Musselman, 1981; Winn, Whitaker, Elliott, & Phillips, 1994) and emotional strain (Hess & Polt, 1960; Partala & Surakka, 2003; Bradley, Miccoli, Escrig, & Lang, 2008). Nonetheless, eye-tracking technology as well as sophisticated algorithms to filter unwanted environmental effects like light changes (e.g. Marshall, 2002; Schwalm, 2009) progress, so that the use of pupil size as a workload indicator in real world environments seems feasible.

Speech signal features

In the context of speech interaction, the use of speech signal features to assess mental workload is a promising approach as characteristics like the fundamental frequency are analyzed for speech recognition anyway. The speech production system or articulatory system is a complex secondary physiological system using structures primarily responsible for respiration and digestion (Flohr, 2002). According to Flohr (2002), speech production involves three subsystems. The respirational system produces an air stream serving as a carrier. This carrier signal then passes through the larynx area and the glottis determining voiced, unvoiced and pitch properties. Finally, through modifications in the pharynx and the mouth-nose area, sounds are articulated. Motor effort related to speech production itself has been shown to elevate autonomous arousal, hence a shift to sympathetic dominance, compared to resting states (Linden, 1987; Arnold, MacPherson, & Smith, 2014; MacPherson, Abur, & Stepp, 2016).

The speech signal can be analyzed by decomposing the acoustic signal into amplitude and fundamental frequency features to obtain prosodic features of the voice (Alter, 2002). Two commonly used features in human state evaluation are fundamental frequency (F0) and voice intensity. Figure 2.8 shows an example of the speech signal and the course of F0 and voice intensity over time. F0 is defined by the number of glottis oscillations per time unit and it correlates with the perceived pitch of a voice. F0 is assumed to be mostly affected by sympathetic activation (Scherer, Grandjean, Johnstone, Klasmeyer, & Bänziger, 2002). Voice intensity can be defined as the acoustic intensity of the speech signal and correlates with the loudness of the voice. Both features have been shown to increase with increasing psychological stress and workload (Mendoza & Carballo, 1998; Scherer et al., 2002; Rothkrantz, Wiggers, Wees, & Vark, 2004; Dromey & Bates, 2005; Bořil, Omid Sadjadi, Kleinschmidt, & Hansen, 2010; Huttunen, Keränen, Väyrynen, Pääkkönen, & Leino, 2011; Giddens, Barron, Byrd-Craven, Clark, & Winter, 2013) but they are also subject to changes

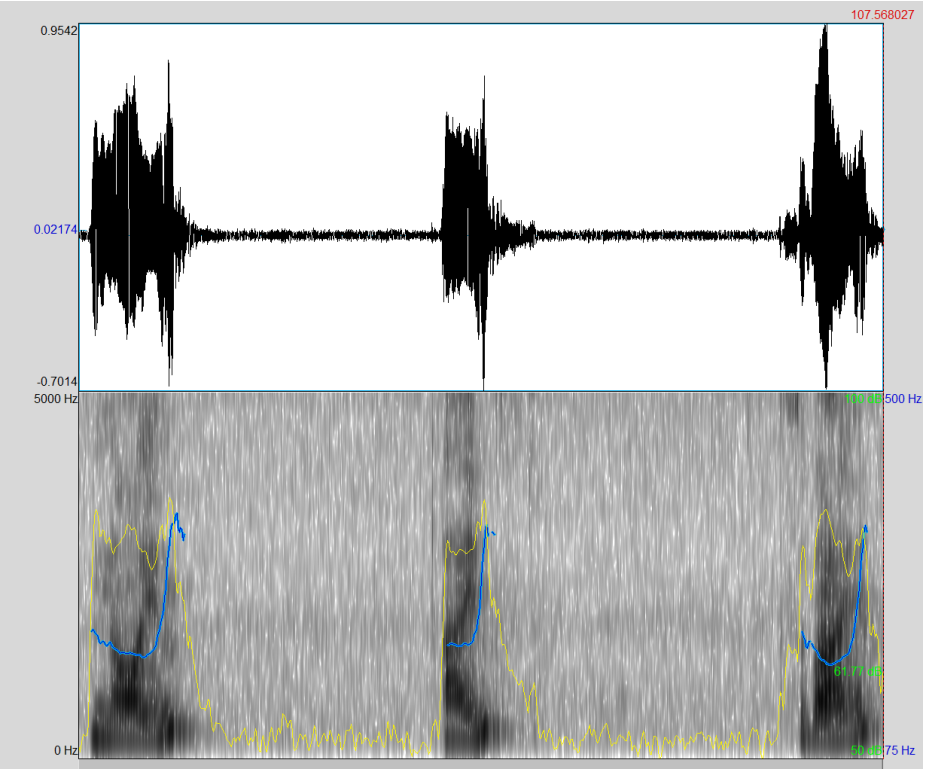


Figure 2.8: Upper window: Acoustic signal of a female speaker articulating the numbers 9, 1 and 2. Lower window: The blue lines represent the fundamental frequency contour (F0 in Hz) and the yellow lines the intensity of the speech signal in dB

due to different emotional states (Williams & Stevens, 1972; Ververidis & Kotropoulos, 2006). Additionally, features like F0-variability are reported to be sensitive to workload changes (Lively, 1993). Note that findings regarding prosodic and voice quality changes due to cognitive demands are sometimes inconsistent (see Giddens et al., 2013; MacPherson et al., 2016) as the influence of autonomic changes on speech production is complex and not entirely understood. Whereas some autonomous responses will directly affect muscle structures involved in speech production, other responses affect speech production only secondarily. Orlikoff and Baken (1989) for example showed that sympathetically innervated heart activity contributes up to 20% to F0 perturbation. Furthermore, it is assumed that increased lung pressure and bronchodilation contribute to increases in F0 and vocal intensity (Giddens et al., 2013).

2.3 Conclusions

In this chapter mental workload was defined for the context of this thesis. Structural as well as energetical attentional resources models build the foundation for the understanding of mental workload. On the one hand structural attentional models help to understand workload implications resulting from the structural characteristics of the task demands. On the other hand, energetical models allow insights into energetical processes involved in coping with the task demands. To understand consequences of workload resulting from a task, both model views have to be integrated. Additionally individual temporary and more constitutional factors affecting mental workload have to be considered. For the measurement and holistic interpretation of the actual workload state of an operator, performance, subjective as well as physiological workload consequences have to be examined thoroughly, since they reflect different aspects of mental workload. Whereas performance measures reflect behavioral consequences in response to changing attentional demands, subjective measures reflect the perceived costs of the attentional demands. Physiological measures, on the other hand primarily reflect the bodily changes in response to energetical demands. All three workload measurement methods were analyzed regarding their general sensitivity to changes in mental workload and their suitability to be used in applied scenarios. Based on the theoretical framework introduced in this chapter the following chapter will focus on cognitive workload in the context of driving.

Chapter 3

Driver cognitive workload

The nature of driving is continually changing. With the introduction of new assistance systems and the increasing availability of technology for non-driving activities like writing emails, listening to podcasts or making phone calls as well as the introduction of new interaction technologies like gestures, touch, gaze and speech, it is more than ever crucial to understand changes in demand that impact the driver to prevent critical situations. For that reason the following chapter will apply the theoretical framework presented in chapter 2 to the context of driving. To do so, it is necessary to elaborate on the nature of the driving task (Section 3.1) as well as on common concurrently performed tasks (Section 3.2) and their implications for driver workload (Section 3.3). Unfortunately, the term driver distraction is often confused or used synonymously with driver workload. For that reason subsection 3.3.2 will highlight the differences between both concepts. In the subsequent sections the emphasis will lie on drivers' cognitive workload. Insights into workload consequences with regard to performance (section 3.4), subjective perception of workload (section 3.5) and especially physiological reactions (section 3.6) that result from the combination of driving and auditory-verbal secondary-tasks will be presented.

3.1 Driving task

To understand the effects of overall task demands on driving performance, it is important to analyze the demands resulting from the driving task and additional demands by secondary-tasks. In fact, only if the type, magnitude and frequency of the present demands are understood, it is possible to grasp their impact on mental resources and with that understand potential consequences regarding performance (Hurts, Angell, & Perez, 2011). In the driver safety literature, driving is often referred to as a 'primary-task'. However, safe driving is no simple single-task but a rather complex

set of tasks involving at least three different control levels (Michon, 1979, 1985). The highest level is the strategic level that includes planning and decision making which takes mainly place prior to the trip but, if adjustment is needed, also during the trip. At this level, the driver (implicitly) sets the goal for the trip and decides which mode of transport to use and which route to take (Michon, 1979; De Waard, 1996). Accordingly, at this level the highest demands on the driver are on visual resources and working memory including central executive attentional processes (Hurts et al., 2011). At the tactical level maneuvers like overtaking, turning, as well as event monitoring and response selection are mostly situation dependent, e.g. determined by traffic density, weather conditions or behavior of other road users, but have to match the strategic goals to a certain degree (Michon, 1979; Hurts et al., 2011). This level demands visual resources, manual resources as well as working memory including central executive attentional processes (Hurts et al., 2011). The lowest control level (operational level) includes fundamental vehicle control tasks like steering, braking, lane-keeping and speed control (Michon, 1979; De Waard, 1996; Hurts et al., 2011). At this control level, demands are mainly placed on visual and manual resources (Hurts et al., 2011). At all levels task demands can exceed the resources available and therefore impact performance at the same level and/or lead to higher level functions like checking mirrors or the speedometer beeing disregarded (De Waard, 1996; Hollnagel, Nåbo, & Lau, 2003).

Primary-task demands involve perceptual and cognitive stages of information processing e.g. working memory and demands mostly affect visual spatial resources, involving focal vision for object recognition (e.g. hazards) as well as ambient vision for orientation and movement (e.g. lane-keeping) on the perceptual stage (Wickens, 2008; Briggs, Hole, & Land, 2016). Furthermore, response demands are mostly manual spatial (steering, accelerating and breaking). Figure 3.1 shows the simplified resource demands of the driving task in the multiple resource framework by Wickens (2002, 2008) described in subsection 2.1.1. However, it should be noted that structural resources needed for driving are not completely limited to the aforementioned demands. Processing auditory input (motor sounds, sound of other vehicles etc.) might be involved as well as visual-verbal resources in the processing of traffic signs, warnings and navigational information. Furthermore top-down processes, most often associated with the central executive, are not explicitly addressed in this model but have to be considered as driving is in most aspects a self-paced task, giving the driver the opportunity to modulate demands strategically e.g. by speed choice (Fuller, 2005). From an energetical resource perspective, compensatory strategies like effort

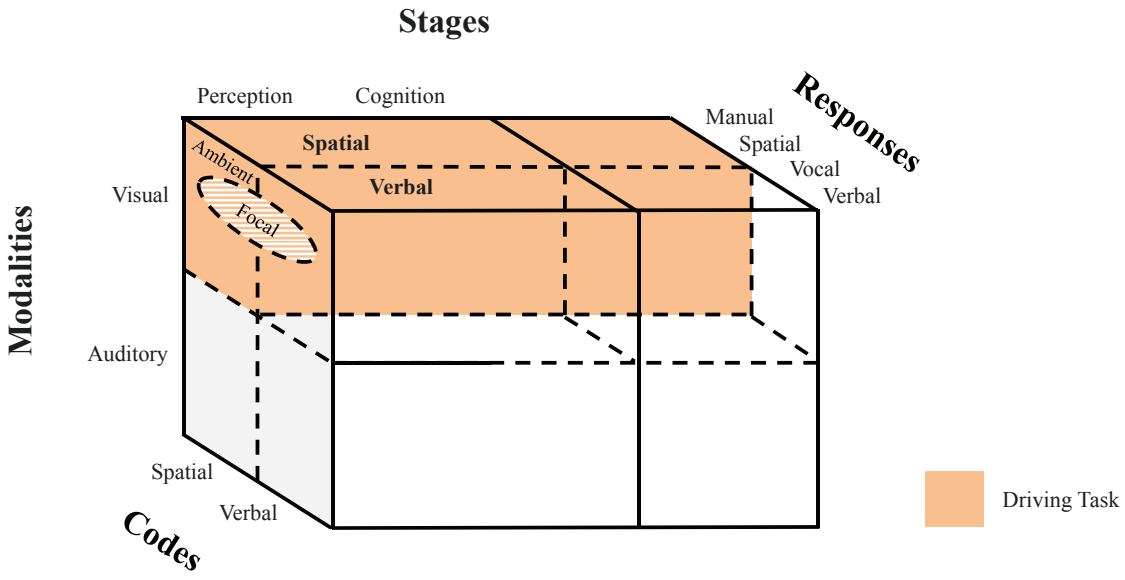


Figure 3.1: Attentional resource demands of the driving task illustrated in the multiple resource framework by Wickens (2002, 2008)

adaptions are needed to maintain performance when task demands change in order to reach the objectives set at the different levels.

3.2 Drivers' secondary-tasks

Secondary or concurrent tasks can be defined as tasks that compete for attentional resources needed for the primary-task of driving. Such tasks can either originate from inside the vehicle or outside the vehicle as e.g. advertisements signs. Tasks originating from inside the vehicle can further be separated into tasks related to in-vehicle technology as navigation systems, infotainment functions or cell phone activities or not in-vehicle technology related like conversations with the passenger, eating, drinking or grooming. Furthermore they can be classified as relevant for driving but not for the driving task itself, e.g. handling navigation system etc., or irrelevant for driving as eating or drinking, though it can of course be argued that on a broader time scale eating and drinking are relevant for safe driving. A further distinction can be made between drivers' internal or external tasks. Internal tasks include phenomena like mind-wandering and external tasks are all task demands that are imposed from outside the driver. Regardless of the source of the concurrent task, it is of great importance to identify structural attentional resources needed to perform a specific task to be able to assess any possible interference of the task with

the primary-task of driving. Hurts et al. (2011, p.13-16) conclude that most tasks involve visual-manual resources for perception and response. Classic examples are manipulating the navigation system, the infotainment system or simply the radio with traditional control elements like rotary knobs or buttons but also with newer technologies like touch screens or gesture input. Most of the listed tasks will also demand working memory and in some cases central executive attentional resources, depending on the specific task.

With the advances in speech recognition technology and regulatory efforts that aim at curbing visual distraction while driving, e.g. by the NHTSA's manual visual distraction guideline (National Highway Traffic Safety Administration, 2012), manufacturers will rely more and more on auditory-verbal interactions resulting in different demand patterns as compared to visual-manual secondary-tasks (see chapter 1). There will be a shift away from visual-manual task demands to more auditory-verbal demands at all stages of information processing. In contrast to traditional interactions, auditory-verbal interactions demand input or action of the user on moments in time that are specified by the system, i.e. *externally-paced*, adding prolonged demands. The following sections will focus on the impact of externally-paced auditory-verbal secondary-tasks on drivers' mental workload and its measurement.

3.3 Driver workload

3.3.1 Driver cognitive workload resulting from auditory-verbal secondary-tasks

When dealing with driver workload it is of utmost importance to understand all costs associated with secondary-task engagement. First of all it is crucial to examine consequences regarding driving performance, as good driving performance is the foundation of safe driving. Furthermore it is also important to analyze the psychological and physiological costs associated with it. A recent systematic review of 206 paper abstracts (1968-2012), examining the impact of secondary-task activities on driving performance, found that in 80% percent of all studies the introduction of a concurrent task led to a decrease in driving performance, 10.3% led to an increase and 9.7% showed no effect at all (Ferdinand & Menachemi, 2014). So under most circumstances, secondary-tasks seem to be detrimental to safe driving. However, as the results suggest, there is also evidence that secondary-task engagement can actually increase driving performance. Besides many other reasons, the nature of the

secondary-task and more specifically the attentional resources involved can be responsible for different results. The interference between two or more tasks is particularly high when they share the same resources on perceptual and cognition stages and involve working memory (Wickens, 2002). Consequently, visual-manual tasks can lead to excessive workload if performed concurrent or parallel to driving as overlap of attentional resources is high. Past research has consistently shown that this leads to severe performance decrements (see e.g. Tijerina et al. (1998), Engström et al. (2005), Horberry et al. (2006), Drews et al. (2009), Hosking, Young, and Regan (2009) and Caird, Johnston, Willness, Asbridge, and Steel (2014)) and contributes to a large extend to distraction related accidents (Klauer et al., 2014; Dingus et al., 2016). For auditory-verbal tasks, however, the evidence is less conclusive. In the framework of Wickens multiple resource theory (see subsection 2.1.1), driving and auditory-verbal tasks should interfere little as they demand different resource pools (see figure 3.2). Regarding the input and output modalities, it seems obvious that when the eyes stay on the road and hands at the steering wheel, this should be beneficial as compared to the performance of visual-manual tasks. But even though auditory-verbal tasks have been shown to impact driving performance less than visual-manual tasks (see e.g. Horberry et al., 2006; Maciej & Vollrath, 2009; Liang & Lee, 2010), they can still lead to detrimental effects. In fact, research on hands-free cell phone conversations has shown that it results in higher subjective workload ratings and changes in driving performance (Alm & Nilsson, 1994, 1995; R. Matthews, Legg, & Charlton, 2003; Horberry et al., 2006). These results suggest an interference on the level of a limited central resource pool, responsible for the 'imperfect' multitask performance of two structurally different tasks. This assumption is backed up by neuro-imaging studies showing that even when two tasks demand different anatomical brain regions, the overall activation in these brain areas is lower under dual-task conditions as compared to single-task performance (Just et al., 2001; Newman, Keller, & Just, 2007; Just, Keller, & Cynkar, 2008). More specifically Just et al. (2008) showed in their fMRI study that activation in driving related brain areas decreased with the introduction of a language processing task. In summary, the interference between driving and auditory-verbal tasks is neither a result of an overlap of structural attentional resources on the stages of perception and response nor of modality specific cognition stages (i.e. the phonological loop vs. visual-manual sketchpad Baddeley and Hitch (1974) and Baddeley (1992, 2003)), but seems to develop mainly through an interference on modality unspecific central stages of cognition. For that reason and for better

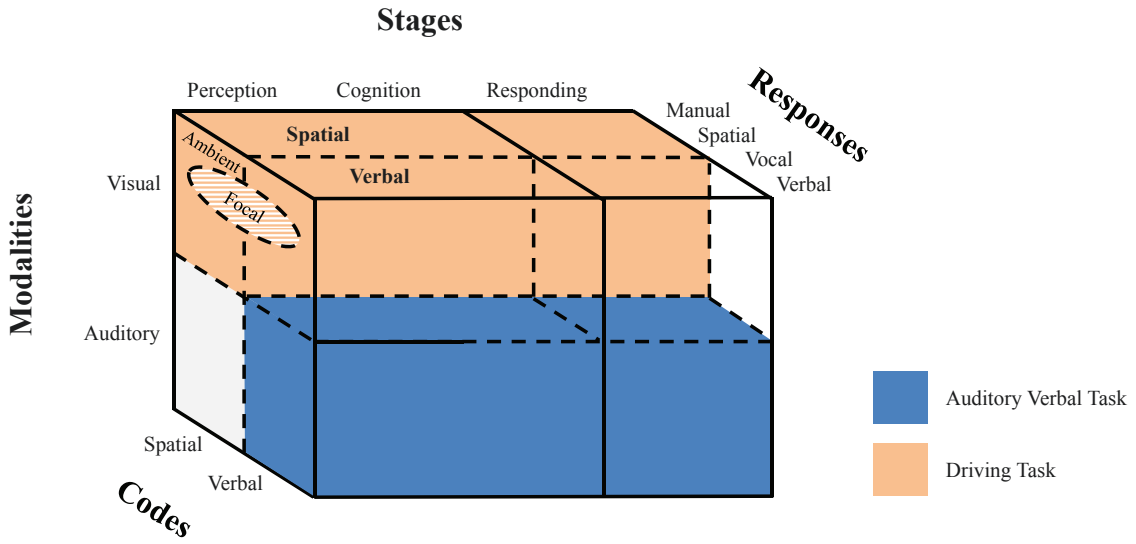


Figure 3.2: Attentional resource demands of driving and auditory-verbal secondary-task illustrated in the multiple resource framework by Wickens (2002, 2008)

readability, the workload resulting from driving and concurrent auditory tasks, will from now on be referred to as driver cognitive workload.

3.3.2 Distinction between driver distraction and driver workload

Unfortunately, the concept of distraction lacks a common understanding or definition and it is often interchangeably used with workload (Schaap, Van der Horst, Van Arem, & Brookhuis, 2013). According to R. Young (2012) there is a great variety of definitions of distraction. His critical review of the distraction literature concludes that comparability of results suffers mainly of missing definitions of attention, confusion of driver state and driver activity, varying operational definitions between studies and implicit assumptions of limited capacity of single resource models. However, driver distraction can be defined as one form of driver inattention: *"The diversion of attention away from activities critical for safe driving toward a competing activity, which may result in insufficient or no attention to activities critical for safe driving."*(Regan, Hallett, & Gordon, 2011, p.1776). As the authors state, this definition is closely related to an earlier definition by Lee, Young, and Regan (2008) and includes internal as well as external sources of distraction. One factor contributing to the confusion of distraction and workload is that distraction comprises the concept of (passive) limited attentional resource consumption as well as the active process of shifting attention between tasks. Therefore distraction can be understood as mental

overload, when the overall resource demands exceed the capacity on the one hand but on the other hand distraction also describes temporal attentional dynamics (Lee, 2014). Both points of views are with no doubt crucial to understand the consequences of secondary-task engagement while driving. However within the scope of this thesis the emphasis will be on the limited attentional resource demand aspect and how to measure the impact of a specific secondary-task demand.

3.4 Impact of driver cognitive workload on driving performance

3.4.1 Measures of driving performance

Even though an *"... exact relationship between MWL* [1] *and accident causation is not easily established ..."* (M. S. Young et al., 2015, p.3) it is clear that mental overload can result in potentially safety critical driver errors (Smiley & Brookhuis, 1987; Brookhuis, 2004). These errors can contribute to the genesis of accidents but due to the complexity of the process do not necessarily lead to accidents (Patten, Kircher, Östlund, & Nilsson, 2004). So even though high mental workload levels have consistently shown to result in performance decrements, there is conflicting evidence regarding the use of hands-free phones. Whereas e.g. McEvoy (2015) reports a 3.8 fold increase of crash risk, no significant increase in risk was found in the study by Klauer et al. (2014). Nevertheless, it is crucial to understand the specific costs of high cognitive workload levels induced by auditory-verbal tasks to assess their specific potential of contributing to traffic accidents as they have been shown to be qualitatively different to performance costs of concurrent visual-manual tasks (Engström et al., 2005; Liang & Lee, 2010; Strayer et al., 2015). In general, two types of performance measures can be differentiated. On the one hand there are lane-keeping behavior and speed maintenance measures. The former refers to continuous lateral tracking on the operational level defined by Michon (1979, 1985) and the latter to longitudinal tracking at the tactical level defined by Michon (1979, 1985). On the other hand there are event-related measures e.g. reactions to hazards in the form of brake reaction times to avoid critical situations. The first will be referred to as continuous measures and the latter as event measures. Most empirical evidence of driving performance changes induced by auditory-verbal tasks can be found in the literature dealing with cell phone conversations while driving. A meta-analysis

[1] mental workload

by Horrey and Wickens (2006) analyzed 23 studies (1994-2004) and concluded that being on the phone while driving is mainly associated with negative effects regarding the driver's reaction to external events rather than with continuous measures. They report an average increase of 130ms in reaction time when driving while engaging in cell-phone conversations as compared to driving-only conditions. Additionally, costs were the same irrespective of the type of phone use (hands-free vs. hand-held) and of whether a conversation was held on the phone or directly with a passenger. Another meta-analysis confirmed most of the results gathered by Horrey and Wickens (2006): Caird et al. (2008) analyzed 33 studies (1994-2006) and confirmed that reaction times to events increase by on average 250ms with cell phone conversations with same effect sizes for hands-free and hand-held phones. In addition the analysis revealed that drivers reduce their speed when talking on the phone compared to a baseline drive without phone conversation. Again no different effects of the type of study (simulator vs. real-world) were found regarding reaction times or speed.

3.4.2 Reactions to critical events

Studies that were published subsequent to the meta-analyses by Horrey and Wickens (2006) and Caird et al. (2008) mostly confirmed the results regarding reaction time (RT) (see e.g. Collet, Clarion, Morel, Chapon, & Petit, 2009; Reyes & Lee, 2008; Engström, Aust, & Viström, 2010; Strayer et al., 2015). Although some studies reported no detrimental effects (e.g. Liang & Lee, 2010), to the author's knowledge no study has found a positive impact of auditory-verbal tasks on RTs. Similar results are reported by research on actual speech-based interactions, e.g. speech-to-text email and texting systems or interacting with a virtual assistant like Apples Siri ™. Reaction times increase consistently with the introduction of speech-based interactions (Lee, Caven, Haake, & Brown, 2001; Jamson, Westerman, Hockey, & Carsten, 2004; Maciej & Vollrath, 2009; Strayer et al., 2015). Interestingly, reaction times to both peripherally and centrally occurring events are impaired under cognitive workload, implicating that these effects cannot solely be attributed to gaze concentration towards the center of the road or a loss of sensitivity in peripheral areas, called cognitive tunneling but are rather the result of an overall decrease in responsiveness to visual stimuli over the entire visual field (R. Young, 2012). In summary, it can be concluded that the increase in reaction times to critical events denotes an important performance decrement that can contribute to accidents.

3.4.3 Lateral vehicle control

Regarding lateral control measures the results and interpretations are more mixed than regarding reaction times. Liang and Lee (2010), He, McCarley, and Kramer (2013) and Pavlidis et al. (2016) for example showed that lane maintenance improvements are accompanied by less smooth steering behavior whereas Mehler, Reimer, Coughlin, and Dusek (2009) did not find any effect of an auditory-verbal secondary-task on lane maintenance and Drews, Pasupathi, and Strayer (2008) report decreased lane-keeping performance, when simultaneously performing an additional auditory-verbal task. Studies using more applied speech-based secondary-tasks (e.g. speech dialog systems, speech-to-text email, voice messages etc.) on lateral control mostly show less lane-keeping variability (Jamson et al., 2004) or no changes at all (Harbluk & Lalande, 2005; Maciej & Vollrath, 2009; Yager, 2013). However contrary to phone conversations, studies on interacting with a speech-based system more often report higher absolute lane deviations (Harbluk, Burns, Lochner, & Trbovich, 2007; He et al., 2014). One possible explanation could be that more realistic tasks also include a visual component and therefore demand confirmational glances at the system to check, whether the right input was made. This interrupts the hand-eye feedback loop and most likely impairs mechanisms counterbalancing erroneous motor reactions (Pavlidis et al., 2016).

Some researchers interpret reduced lane-keeping variability as a performance gain (He et al., 2013) others interpret it as a performance decrement (Östlund et al., 2005; Reimer, 2009). Both sides argue that the observed effect is the result of compensatory mechanisms, but differ in their explanations of the effect. He et al. (2013) describes it as compensational effort to circumvent the loss of lateral control whereas Reimer (2009) argues that this mechanism serves as a mean to allow attentional shifts to a secondary-task. Regardless of the interpretation, the overall effect of cognitive workload on lateral control is low and ambiguous and therefore most likely less safety critical than increased reaction times to critical events.

3.4.4 Longitudinal vehicle control

Speed reductions while performing an auditory-verbal secondary-task as reported by Caird et al. (2008) can be thought of as drivers' most effective strategy to reduce task demands, as it directly influences the number of information units to be processed at a time (Fuller, 2005). This effect, which is similar to lateral control measures, was

mostly found for high cognitive workload levels (see e.g. Reimer, 2009; Haigney, Taylor, & Westerman, 2000; Reimer, Mehler, Wang, & Coughlin, 2012). Regarding speed maintenance little to no effects of speech-based interactions were found (e.g. Yager, 2013; He et al., 2014). From a safety perspective reduced speed can be interpreted as a positive compensation mechanism to reduce task demands and risk.

3.4.5 Summary on driving performance

In general, auditory-verbal cognitive tasks consistently show decreased event-related performance, but only small changes regarding lateral and longitudinal control of the vehicle. It can be argued that e.g. lane-keeping and speed maintenance use different attentional resources than hazard responses (ambient vs. focal vision), are automated to a greater extend (operational level) (Horrey & Wickens, 2006) and are therefore less impacted by cognitively demanding tasks. However, changes in longitudinal and lateral measures might be useful to assess the mental workload of the driver and to predict hazardous event-related performance in combination with suitable physiological indicators of mental workload.

3.5 Impact of driver cognitive workload on subjective experience

Most cognitive workload studies that used subjective measures showed that drivers are aware of changes in task demands and therefore rate their perceived workload higher when performing a secondary-task while driving compared to driving-only (see e.g. Alm & Nilsson, 1994, 1995; Lee et al., 2001; R. Matthews et al., 2003; Niezgoda, Tarnowski, Kruszewski, & Kamiński, 2015; Strayer et al., 2015). However, sometimes subjective measurements are not sensitive to different task difficulties (see e.g. Rakauskas, Gugerty, & Ward, 2004; Harbluk & Lalande, 2005). The reason for this might be that drivers are not always aware of their own performance decrements as they are absorbed by increasing task demands and thus are not able to assess their workload correctly (Lesch & Hancock, 2004; Horrey, Lesch, & Garabet, 2009). Even though questionnaires are very useful as a complementary measure to get insights into the perceived workload of a driver, they are not feasible for applied settings, where workload measures are used to predict the drivers' state.

3.6 Impact of driver cognitive workload on physiology

Previous research on e.g. cardiac measures has highlighted that they are potentially more sensitive to changes in attentional demands than driving performance measures (Brookhuis, Vries, & Waard, 1991; Lenneman, Shelley, & Backs, 2005; Lenneman & Backs, 2009; Mehler et al., 2009). As described earlier in subsection 2.2.3, physiological indicators of mental workload mostly reflect energetic consequences of attentional resource investment. The mental effort needed to cope with task demands can be assessed by measuring ANS responses. However, the simplified interpretation that increased mental effort elicits solely sympathetic activity cannot be upheld against the growing empirical evidence suggesting a more complex interplay of parasympathetic and sympathetic activation patterns (Berntson et al., 1991; Backs, 1995). In their paper Lenneman and Backs (2009) present empirical evidence that different tasks and task combinations elicit different ANS activity. They summarize results by other researchers showing that e.g. central processing tasks like mental arithmetics elicit reciprocally coupled sympathetic activation and parasympathetic withdrawal, and that visual-manual tasks are more likely to result in uncoupled parasympathetic withdrawal whereas executive attentional control tasks in dual-task settings evoke uncoupled sympathetic activity. In an attempt to transfer these results to driving under workload Lenneman and Backs (2009) present evidence that a single driving task increases uncoupled parasympathetic withdrawal whereas an auditory-verbal n-back task increases reciprocally coupled sympathetic activation and parasympathetic withdrawal compared to a resting state. When adding the n-back task to driving, the results showed a significant increase in reciprocally coupled sympathetic activation and parasympathetic withdrawal for the high n-back difficulty compared to driving-only. No autonomic changes where found for driving and a simple n-back condition. These results illustrate that the use of only one physiological parameter is not sufficient to capture changes in complex autonomic activity patterns. Furthermore, integrating workload indicators that are not directly affected by the ANS, but provide information about attention allocation like eye fixations do, might be of additive value in multi-measure approaches for workload estimation.

3.6.1 Cardiac measures

Heart rate

In general heart rate has been shown to be sensitive to the demands imposed by a cognitive secondary-task compared to driving-only (Brookhuis et al., 1991; Engström et al., 2005; Lenneman et al., 2005; Liu & Lee, 2006; Collet et al., 2009; Lenneman & Backs, 2009; Mehler et al., 2009; Reimer & Mehler, 2011). However, HR seems to be limited in its sensitivity to differentiate between small differences in workload levels (Engström et al., 2005; Lenneman et al., 2005; Reimer & Mehler, 2011). Only a few studies revealed significant differences in HR between moderate and high workload levels (see e.g. Mehler et al., 2009). One reason for that might be that subtle shifts in sympathetic-parasympathetic balance are not reflected in heart rate changes. The same change in HR induced by workload can be the result of different autonomic states (Backs, 1995). Therefore, heart rate is useful as a general workload indicator in a multimodal approach but, because of its lack of differentiating smaller workload differences, not feasible as single indicator.

Skin temperature

Relatively few studies examine facial temperature features as indicators of cognitive workload in transportation research. A couple of studies have been conducted in the area of ship navigators' workload in decision making and training. They consistently showed declining nose tip temperatures in reference to forehead temperature in mentally demanding situations (Murai, Hayashi, & Inokuchi, 2004; Murai, Okazaki, Stone, & Hayashi, 2007; Murai, Hayashi, Okazaki, Stone, & Mitomo, 2008; Nishimura, Murai, & Hayashi, 2011). Similar results have been found in the context of driving under workload (Or & Duffy, 2007; Yamakoshi et al., 2008; Itoh, 2009). A driving simulator study by Or and Duffy (2007) showed that the temperature at the tip of nose drops significantly and forehead temperature remains constant, when adding a mental addition task to the driving task. Facial temperature features of different regions in the face have been shown to correlate to different degrees with longitudinal and lateral control measures (Reyes, Lee, Liang, Hoffman, & Huang, 2009). However, there seem to be big inter-individual differences in the reaction to mental workload regarding nose tip temperature (Itoh, 2009). Periorbital temperature (Reyes et al., 2009) or the supraorbital areas (Wesley, Shastri, & Pavlidis, 2010) were also used successfully to measure cognitive workload. More recently promising attempts are made to use thermal imaging to measure changes in perspiration in the perinasal region

(Shastri, Papadakis, Tsiamyrtzis, Bass, & Pavlidis, 2012; Pavlidis et al., 2016). In summary, the temperature at the tip of the nose appears to be the most promising amongst facial temperature features to be used for cognitive workload estimation. It has shown to be sensitive to increased sympathetic activity and can be assessed non-intrusively. Additionally this region seems to be superior to other facial regions as it is not obstructed by sculp hair, facial hair or glasses as other regions such as the periorbital or perinasal areas can be. Nevertheless, as the empirical data on the use of nose tip temperature as workload indicator is insufficient, more studies examining the systematic influence of cognitive workload are necessary.

3.6.2 Ocular measures

Measures of gaze dispersion

In the driving under cognitive workload literature, there are consistent results indicating that drivers tend to concentrate their gaze on the center of the road resulting in reduced horizontal and vertical gaze dispersion (see e.g. Recarte & Nunes, 2000; Engström et al., 2005; Victor, Harbluk, & Engström, 2005; Tsai, Viirre, Strychacz, Chase, & Jung, 2007; Horrey, 2011; Niezgoda et al., 2015). These results are in line with earlier research, showing smaller saccadic extend with auditory task load (May, Kennedy, Williams, Dunlap, & Brannan, 1990) and lower entropy of fixation patterns from pilot workload studies (Ephrath, Tole, Stephens, & Young, 1980; Tole, 1983). Gaze concentration is often accompanied by decreased visual scanning of the periphery (Reimer, 2009; Reimer et al., 2012), less checking of mirrors and speedometer in cognitive workload conditions (Recarte & Nunes, 2000, 2003; Harbluk et al., 2007) and impaired decision making (Recarte & Nunes, 2003). In a series of studies, Recarte and colleagues consistently found changed fixation dispersion patterns with auditory-verbal concurrent tasks as compared to driving-only tasks. In an early study, Recarte and Nunes (2000) report a significant decrease of up to 60% in horizontal and vertical fixation position variability with a verbal spatial imagery task as compared to driving-only. In a follow-up study they could extend their findings to a series of other cognitive tasks, including a phone task, but also showed that the extend of the gaze concentration depends on task difficulty (Recarte & Nunes, 2002). Later they reproduced the results of their first study and additionally found impaired detection performance and decision making with higher cognitive demands as a result of attentional interference (Recarte & Nunes, 2003). Particularly horizontal gaze dispersion seems to linearly decrease with cognitive task demands. Reimer et al. (2012) showed

that incrementally increasing the difficulty of a concurrent auditory-verbal n-back task (0-back, 1-back, 2-back) resulted in a respective linear decline in the standard deviation of fixation positions for the 0-back and 1-back condition compared to a reference with no concurrent activity. The standard deviation of fixation positions also decreased for the 2-back condition, but no further decrease was found for this condition compared to the 1-back condition. A similar pattern was found by Niezgoda et al. (2015) indicating a possible floor effect, when cognitive workload increases. Apart from the study conducted by Reimer et al. (2012), there are only few studies explicitly researching the sensitivity of the phenomenon to graded levels of workload in the context of driving.

Recarte and Nunes (2003) interpret gaze concentration as a compensatory mechanism to keep attention at the most critical area of driving, the road ahead. This appears reasonable as the most relevant information for lateral and longitudinal control, as well as hazard perception, are ahead of the driver. Crash statistics support this assumption, as almost 70% of crash impact angles in the United States are reported to be in areas visible to the driver through the windshield (R. Young, 2012).

Gaze concentration can be explained in the framework of the SEEV - model by Wickens and colleagues (Wickens, Goh, Helleberg, Horrey, & Talleur, 2003). According to the SEEV model, visual scanning is influenced by saliency, expected location and the value of a stimulus for a particular task. These parameters drive focal vision by influencing eye-movements (Horrey, Wickens, & Consalus, 2006). Especially the value of a stimulus for the execution of a task has been shown to be highly related to visual attention allocation (Horrey et al., 2006). In driving the scene in front of a driver is of the highest value and the expected location of relevant information for the driving task in front. Complementary, effort also plays a role when switching visual attention from one location to another. Longer eye-movement distances are related to higher effort and therefore avoided in mentally demanding situations (Wickens, 2014). The influence of effort on visual scanning, however, seems to be small and only to appear in highly demanding situations (Horrey et al., 2006). Another line of argument for a compensatory mechanism can be derived from laboratory studies examining the effect of auditory secondary-tasks on visual span and visual search. The visual span or perceptual span is the *"region of the visual field from which we can extract information during an eye fixation"*(Pomplun, Reingold, & Shen, 2001, p. B57) and is often identified by using the gaze contingent moving display technique (see for a review Duchowski, Cournia, & Murphy, 2004). Pomplun et al. (2001) showed that reaction times in visual search tasks increased and the visual span decreased significantly with

the introduction of an auditory secondary-task compared to a visual single-task condition. In another experiment they additionally found that the decrease of visual span size is dependent on the difficulty of the auditory secondary-task. So if the visual span size decreases through a concurrent auditory task while driving, then it seems beneficial to focus visual attention on areas that are most important for safe driving. Moreover, the results regarding reaction times are also comparable to the drivers' reaction time results, indicating that even if foveal attention is on the visual task, a concurrent auditory task may inhibit effective vision through interference of a shared limited attentional resource.

To infer, how gaze concentration might contribute to performance changes with cognitive workload, two visual channels are distinguished, *focal vision* and *ambient vision* (Leibowitz & Dichgans, 1980; Leibowitz & Post, 1982). Focal vision is closely linked to eye-movements and mainly associated with tasks involving high visual acuity like visual search and object recognition (Leibowitz & Post, 1982; Horrey et al., 2006). Ambient vision is concerned with spatial orientation and localization (Leibowitz & Dichgans, 1980) and is not directly linked to the execution of eye-movements (Horrey et al., 2006). Focal vision is assumed to demand more attentional resources and to more strongly degrade towards the periphery compared to ambient vision (Leibowitz & Post, 1982; McKee & Nakayama, 1984; Horrey et al., 2006). Transferred to the context of driving, focal vision plays a major role in the detection of hazardous events whereas ambient vision is primarily involved in lateral control (Summala, Nieminen, & Punto, 1996; Horrey et al., 2006; Wickens, 2008). Engström et al. (2005) observed that cognitive workload increases gaze concentration on the road and decreases lane-keeping variability. They hypothesize that both effects are connected and present two possible explanations for it. The first one states that visual scanning is impaired due to increased attentional resource demands, which lead to more focused gaze patterns on the road and consequently to enhanced perception and tracking. A second possible explanation according to the authors could be that drivers perceive the need to protect lateral control performance and therefore put more effort in it. However, there seems to be no causal relationship between eye-movements and lane-keeping variability. Although a decrease in lane-keeping variability is often accompanied by more focused gaze concentration, Cooper, Medeiros-Ward, and Strayer (2013) showed that the lateral performance effect is independent of the eye-movement distribution. These results are in line with the results of Summala et al. (1996) that showed that lane-keeping can be accomplished to a great extent with ambient vision without involving focal vision and hence is largely uninfluenced by fixation distribution. Nevertheless,

it might be possible that ambient vision benefits from the more stationary fixation patterns under high workload. The reaction to critical events is clearly dependent on focal vision. A gaze concentration on the road ahead might lead to prolonged reaction times with regard to events appearing from the periphery, simply explained by the "don't look so don't see" phenomena. However, as described in section 3.4, reaction times also increase for objects that are in the central field of vision. In situations of high cognitive workload, the competition for attentional resources is most likely to impair focal vision leading to a general decrease in visual sensitivity over the whole field of view rather than specific sensitivity loss in periphery (R. Young, 2012).

Fixations duration and fixation count

Fixation duration and fixation count are closely coupled. An increase respective decrease in fixation duration is usually associated with a decrease respective increase of the number of fixations within a fixed time interval. Laboratory findings on visual search and concurrent auditory tasks indicate longer fixations with higher auditory task difficulties (Pomplun et al., 2001). However, there is little consistent evidence in applied settings like driving with a concurrent auditory-verbal task. Recarte and Nunes (2000) for example report task dependent differences for fixation duration. Whereas a simple verbal secondary-task while driving decreased fixation durations slightly, a mental spatial imagery task led to longer fixations compared to driving-only and to the dual-task condition with the simple verbal task. Tsai et al. (2007) reported decreased fixation duration to predict errors in the auditory secondary-task. Contrary to this, on a descriptive level Niezgoda et al. (2015) found longer fixation durations with increasing task difficulty in an auditory-verbal n-back task compared to baseline driving but no significant differences between task difficulties. Other studies found equally contradictory results with longer or shorter fixation durations on an individual level for the same task (Itoh & Inagaki, 2008), emphasizing that there are inter-individual differences as well as task related characteristics that influence fixation durations. Additionally, several studies report no changes in fixation durations at all (Strayer, Drews, & Johnston, 2003; Tsai et al., 2007). As described in section 2.2.3, fixation durations are highly dependent on the task. In fact, it can be argued that both longer as well as shorter fixations might occur during the same task depending on situational and structural parameters of the setting as well as individual differences in coping strategies. Based on the intake-rejection hypothesis by Lacey (1959), Rötting (2001) proposes that fixation durations decrease during environmental intake and increase when environmental rejection takes place due to lower,

respectively higher, suppression of saccadic execution. Whereas the driving task itself is more likely to provoke environmental-intake, mental tasks involving internal manipulations like arithmetics have been shown to provoke physiological patterns reflecting environmental-rejection (Lacey, Kagan, Lacey, & Moss, 1963; Obrist, 1963). Depending on the strategy of the driver both processes might occur in different ratios. This could explain the inconsistent results as well as the finding of decreased fixation times before errors in the secondary-task reported by Tsai et al. (2007).

Pupil

Pupil size has consistently increased with increasing cognitive demands in the literature regarding driving under cognitive workload (see e.g. Recarte & Nunes, 2000, 2003; Palinko, Kun, Shyrokov, & Heeman, 2010; Niezgoda et al., 2015). In a recent study Niezgoda et al. (2015) report that the increase in pupil size is the most sensitive measure to differentiate between three difficulty levels of an auditory-verbal n-back task concurrently performed while driving. Moreover, an increase in pupil diameter is often accompanied by better performance in the auditory-verbal task (Tsai et al., 2007). Both, sympathetic activation as well as parasympathetic inhibition play an important role in the pupillary response as described earlier in section 2.2.3. However, parasympathetic withdrawal seems to play a special role as a modulator, to differentiate between different task difficulties, when sustained cognitive processing of tasks is needed (Steinhauer et al., 2004). Overall changes in pupil size in reaction to cognitive workload appear to be the most reliable and sensitive indicator. Pupil size is therefore valuable in multi-measure approaches, despite its susceptibility to confounding influences like emotions and changes in luminance (see section 2.2.3).

3.6.3 Speech signal features

To the author's knowledge there are hardly any studies that systematically examine the influence of cognitive workload on speech signal features like fundamental frequency or voice intensity. This is surprising, as speech signals can be easily recorded and speech-based interactions become more common. So far there is some evidence from aircraft pilot studies replicating the laboratory results described in section 2.2.3. For example Huttunen et al. (2011) demonstrated workload induced increases in F0 and voice intensity of military pilots with highest increases in intensive flight sections. Bořil, Kleinschmidt, Boyraz, and Hansen (2010) and Bořil, Omid Sadjadi, et al. (2010) compared two different cognitive secondary-tasks: talking to a co-driver

and conducting phone calls to a speech dialog system. F0 increased significantly in the speech dialog system condition compared to the co-driver conversation. They argue that co-driver conversations impose less workload as they are less demanding than talking to a speech dialog system. Moreover, passengers might be more aware of the traffic situation and the current demands of the driver and therefore pace the conversation accordingly (Drews et al., 2008). Even though aforementioned speech signal features are promising, more research is needed to examine the impact of cognitive workload on F0 and voice intensity.

3.7 Conclusion

This chapter elaborated on the causes and consequences of cognitive workload in the driving context. Structural aspects of the driving and common cognitive secondary-tasks were thoroughly examined to get insights into the genesis of driver cognitive workload and its consequences on performance, subjective ratings and physiology. It was shown that cognitive workload impacts driving performance in manifold ways but seems to affect reactions to critical events more severely than continuous driving. Furthermore, past research on non-intrusive physiological measures in the driving context was analyzed regarding the potential of multiple measures to be used to differentiate between different cognitive workload levels. Besides traditional measures like pupil size, fixation characteristics and heart rate, fundamental frequency of the voice and voice intensity of the speech signal as well as nose tip temperature were identified as promising. The following chapter will therefore examine the sensitivity of the aforementioned measures to workload differences induced by an auditory-verbal task with differing task difficulty while driving. To allow for a holistic understanding of the impact of the induced workload, performance as well as subjective ratings will additionally be taken into consideration.

Chapter 4

Experiment 1 - Sensitivity of physiological measures

As past research has predominantly focused on differences in workload between single driving task conditions and dual-task conditions with additional cognitive secondary-tasks, the following experiment will explicitly focus on the comparison between different cognitive workload levels induced by different dual-task conditions. In section 4.1 hypotheses regarding performance, subjective and physiological measures of cognitive workload will be generated on the basis of the previous chapters. Following that the experimental setup (4.2.1) and design (4.2.3), procedure (4.2.4) as well as data pre-processing (4.2.5) and data analysis procedures (4.2.6) will be described in the methods section 4.2. Subsequently the results regarding performance, subjective and physiological measures will be presented in section 4.3, followed by a thorough discussion of the results (4.4) and the experimental limitations (4.5). The chapter closes with the conclusions regarding the experiment (4.6).

The conduction and the analysis of this experiment was supported by two Bachelor students. Oliver Hammerschmidt analyzed the facial temperature data and Oliver Knoll the ECG data. Both helped with the conduction of the experiment. For that reason the study presented here can in part be found in the their respective theses (Hammerschmidt, 2013; Knoll, 2013). Moreover, parts of this study were published in a conference proceedings paper (Ruff & Rötting, 2013) [1] and the technical implementation and analysis was supported by employees Mario Lasch and Otto Lutz of the Chair of Human-Machine Systems at Technische Universität Berlin.

[1] Text in part similar to publication

4.1 Introduction

Chapter 3 highlighted that there is more research needed regarding the measurement of cognitive driver workload. There is a particular need for robust and non-intrusive measures for the assessment of workload in real world settings. The aim of the first experiment of this thesis is to examine the sensitivity of physiological parameters described in section 3.6 regarding different levels of cognitive workload imposed by an externally paced auditory-verbal task performed concurrently to driving. The experiment focuses on facial temperature features and speech signal features, as there has been little research regarding their sensitivity to different levels of driver cognitive workload and their potential to be integrated into multi-measure approaches to workload estimation. The author chose a secondary-task that mimics the attentional demands of a speech interaction system. This was done to increase ecological validity. Primary- and secondary-task performance as well as subjective workload ratings were assessed to gain insight into coping strategies of the participants. On the basis of section 3.4, section 3.5 and section 3.6 the following hypotheses regarding performance, subjective workload and physiology were generated:

Regarding the impact of task difficulty on performance the following is hypothesized:

- Although there are divergent results in the literature regarding lane-keeping behavior (see e.g. Liang & Lee, 2010; He et al., 2013; Pavlidis et al., 2016), the majority of studies reports no changes in lane-keeping variability (Harbluk & Lalande, 2005; Maciej & Vollrath, 2009; Yager, 2013). This was also the conclusion of two meta-studies (Horrey & Wickens, 2006; Caird et al., 2008). Lane-keeping variability is hence expected to be unaffected by increasing task difficulty.

- Increasing task difficulty is expected to result in degraded performance in the secondary-task in order to protect driving performance as driving is instructed as primary-task.

Regarding the impact of task difficulty on subjective workload ratings the following is hypothesized:

- Subjective workload ratings are expected to increase with increasing task demands induced by the secondary-task as has been shown repeatedly (see section 3.5)

Regarding the impact of task difficulty on physiological measures the following is hypothesized:

- Heart rate has long been known to be a robust indicator of mental workload. It is therefore expected that heart rate increases with increasing task demands.
- Evidence hints to sensitivity of the nose tip temperature (see Or & Duffy, 2007; Yamakoshi et al., 2008; Itoh, 2009). It is therefore hypothesized that nose tip temperature drops with increasing task demands due to heightened sympathetic activity.
- When visual demands remain constant, increasing secondary-task difficulty is accompanied by an increase in the proportion of environmental-rejection compared to environmental-intake and therefore leads to longer and fewer fixations (see subsection 3.6.2), as demonstrated by Recarte and Nunes (2000), Pomplun et al. (2001) and Niezgoda et al. (2015). It is therefore expected that fixation count drops and fixation duration increases with increasing mental workload.
- There is consistent evidence for heightened gaze concentration under cognitive workload (see subsection 3.6.2). For that reason horizontal fixation dispersion is expected to decrease with higher secondary-task demands.
- Pupil size is known to be a sensitive indicator of different cognitive workload levels (see e.g. Recarte & Nunes, 2000, 2003; Palinko et al., 2010; Niezgoda et al., 2015). Therefore pupil size is expected to increase with increasing task demands.
- Research from non-driving studies suggests that the voice's fundamental frequency and voice intensity increases with increasing task difficulties (Mendoza & Carballo, 1998; Scherer et al., 2002; Rothkrantz et al., 2004; Dromey & Bates, 2005; Bořil, Omid Sadjadi, et al., 2010; Huttunen et al., 2011; Giddens et al., 2013). The same is expected for the driving scenario in this experiment.

Figure 4.1: Experiment 1 - Driving simulator with the projected LCT at the wall in the background.

4.2 Methods

4.2.1 Experimental setup

For the study a custom-built fixed base driving simulator with a Volkswagen Touran cockpit was used (Figure 4.1). The simulator was located in a temperature regulated room (M=27.8 °C, SD=0.8 °C across all participants) that was illuminated by a single ceiling-mounted Osram LUMILUX T8 58W tube. The driving environment was projected on a 1.13m x 0.80m white wall approximately 1.50 m in front of the driver's seat with a resolution of 1024 x 768 pixel (see figure 4.1).

Experimental tasks

As primary-task, the participants performed the Lane Change Task (LCT) developed by Mattes (2003) at a fixed speed of 80 km/h. The driving task consisted of four straight two-minute tracks ('pre'/'p1'/'p2'/'post') separated by a 180° turn (see figure 4.2) and was repeated for each experimental condition. Usually the LCT task is performed at 60km/h. Pre-tests, however, revealed that four consecutive tracks at 60 km/h with a very low demanding secondary task led to undesired vigilance problems and sleepiness of the participants. Therefore it was decided to increase the speed to

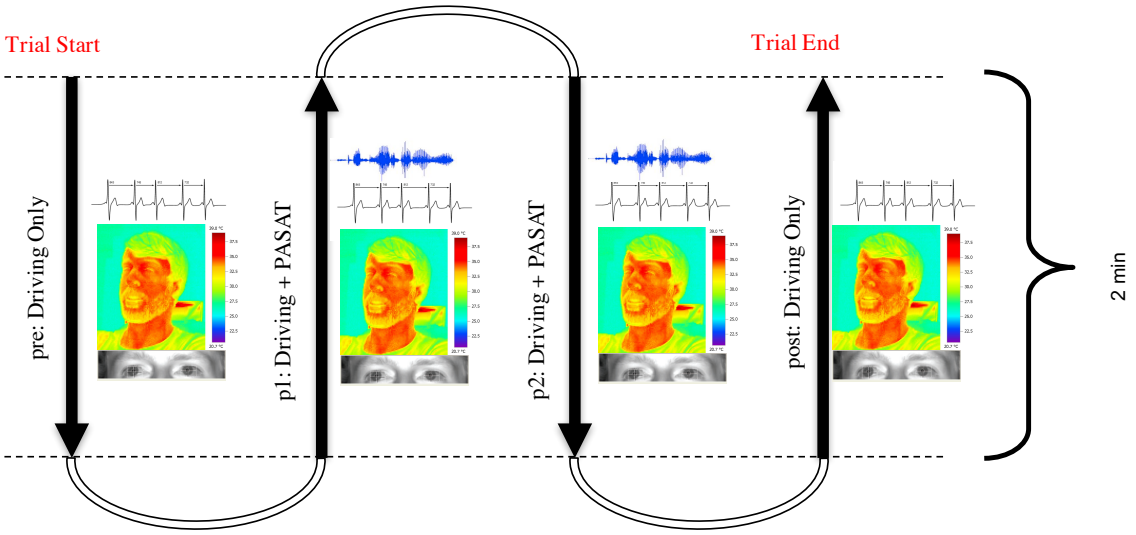


Figure 4.2: Experiment 1 - Driving track consisting of four 2 min consecutive tracks (pre, p1, p2, post), divided by three 180° turns. Additionally the different physiological measurements methods are depicted at the respective tracks.

80km/h. Lane changes on all four tracks were indicated by white traffic signs to the left and right of the driveway (see figure 4.3). Black arrows on the signs provided the information which lane to switch to. Whereas the pre and the post experimental tracks served as driving-only baselines, during track p1 and p2 participants additionally performed a modified Paced Auditory Serial Addition Task (PASAT) (Gronwall, 1977). The PASAT was chosen because of its high demands on working memory as well as sustained attention. Depending on the task condition, participants were either asked to continuously sum up the last two auditorily presented digits ranging from 1 to 9 and to respond verbally or to repeat the digits presented. For each track, i.e. for p1 and p2 in every experimental condition, a random sequence of digits was created to avoid training effects.

Setup of the physiological measurement

ECG and room temperature were recorded with a Becker Meditech Varioport biosignal recorder sampling at 512Hz and saving at a rate of 256Hz. Eye gaze behavior and pupil diameter was recorded using the SensoMotoric Instruments OEM version of the RED-m remote eye-tracker sampling at 60Hz with 0.1° spatial resolution and a gaze position accuracy of 0.5°. The eye-tracking system was placed above the steering wheel in approximately 0.70 m distance to the drivers' eyes. Facial temperature was measured using a Testo 875-1 thermography camera (optical resolution: 160*120 pixel, emission coefficient: 0.98, thermal resolution: $< 0.08K$)

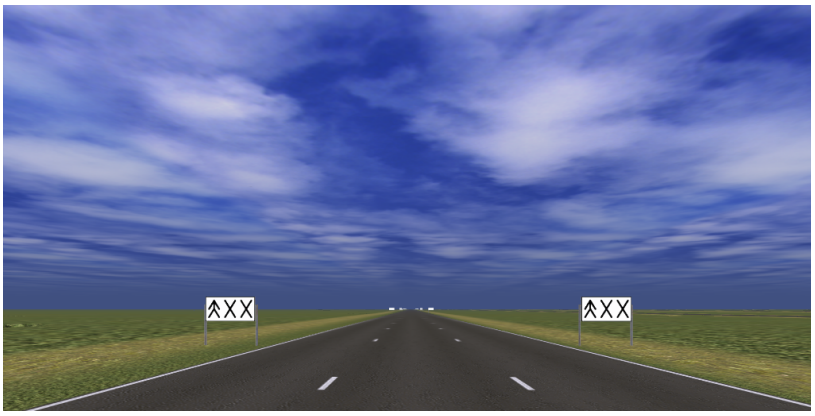


Figure 4.3: Experiment 1 - LCT signs indicating lane changes to the left and the right of the road

placed in 0.75m distance to the participant on the dashboard of the car base. A Sennheiser ew 100 g2 wireless microphone sampling at 44100Hz stereo was attached to the participants collar to record the speech signal. As central coordinating unit a Labview program was developed for synchronization of all data channels via UDP (User Datagram Protocol) messages (Figure 4.4).

4.2.2 Participants

A sample of 30 participants was targeted to ensure a balanced design (see 4.2.3). As the data sets of six of the initial 30 participants were incomplete, a further six participants were recruited. In the following, the demographics of the participants that yielded the complete 30 data sets will be presented. All data that stemmed from the participants whose recordings were incomplete, was discarded. The gender of the participants was balanced and their age ranged from 20 to 31 years ($M = 24.73$, $SD = 2.68$). Prerequisite for participation was the possession of a valid driver's licence. To assess driving experience, the participants' driving habits and their yearly mileage was inquired. The results are presented in figure 4.5 and 4.6. Most of the participants were in an educational program either at a university ($N = 23$), at a public school ($N = 2$) or in vocational training ($N = 2$). The remaining participants pursued a profession ($N = 3$). After the experiment participants were rewarded with 10 Euros per hour or alternatively a certification of student experimental hours.

4.2.3 Experimental design

This study featured a one-factorial within-subjects design. The factor *'task difficulty'* was defined as the difficulty of the PASAT and consisted of three levels. In the

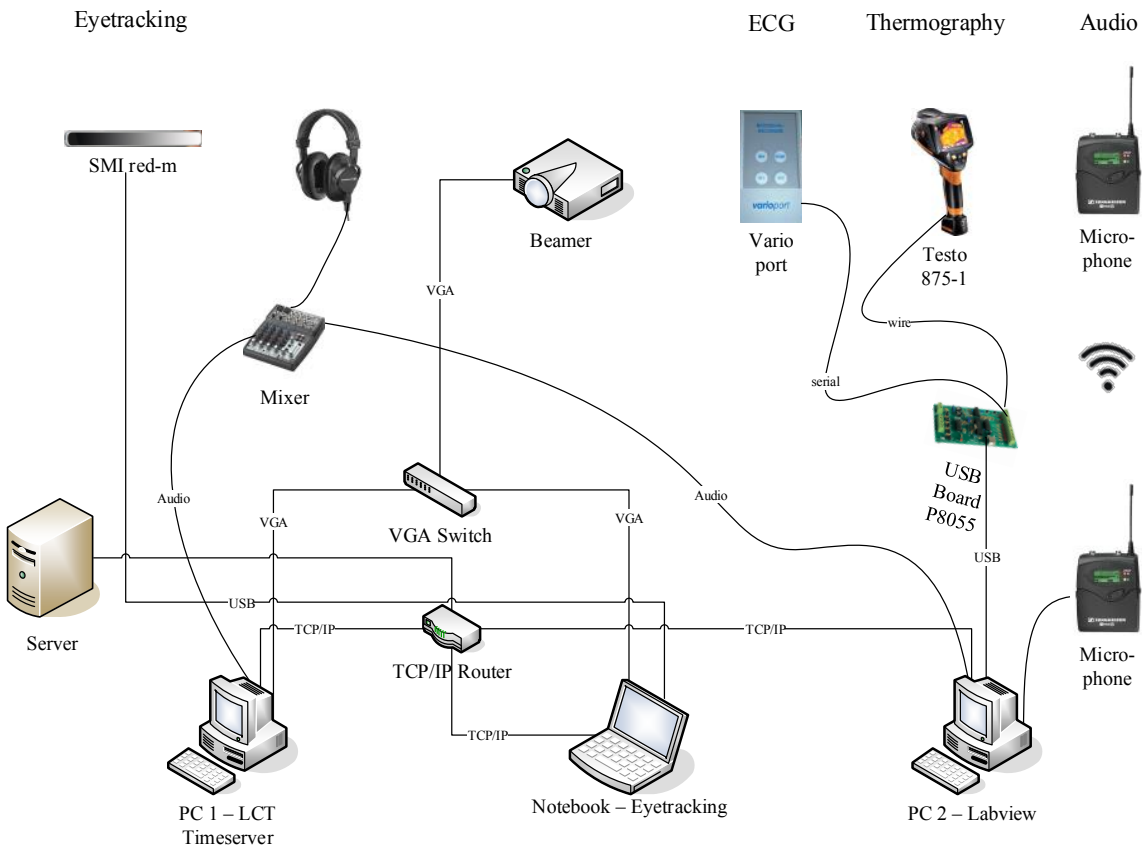


Figure 4.4: Experiment 1 - Technical setup of the experiment. Automation and synchronization of the experiment was accomplished by a custom Labview program (PC 2). The figure specifies the hardware used and the physical connections between the different components.

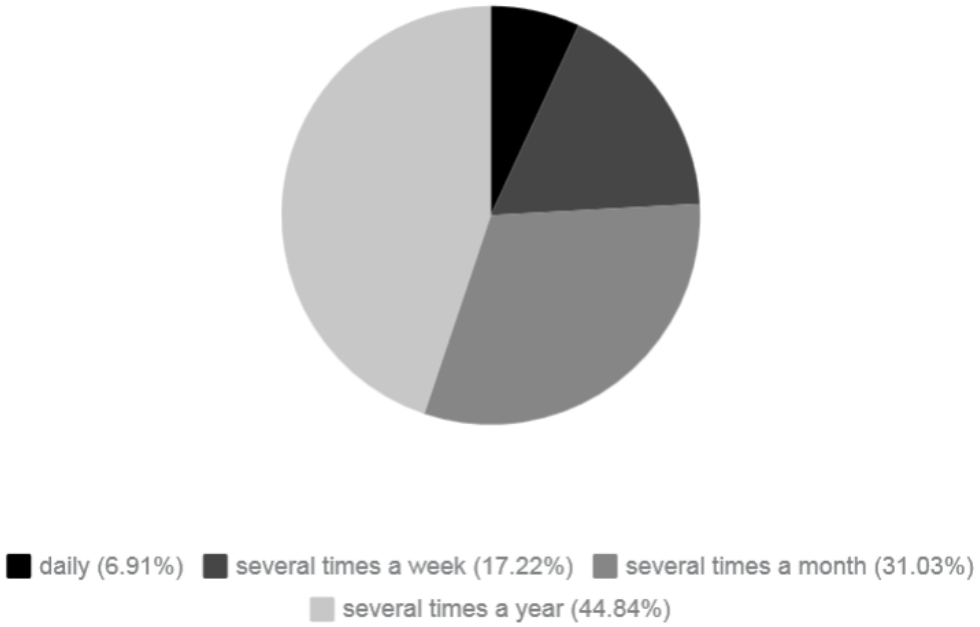


Figure 4.5: Experiment 1 - Participants' frequency of driving

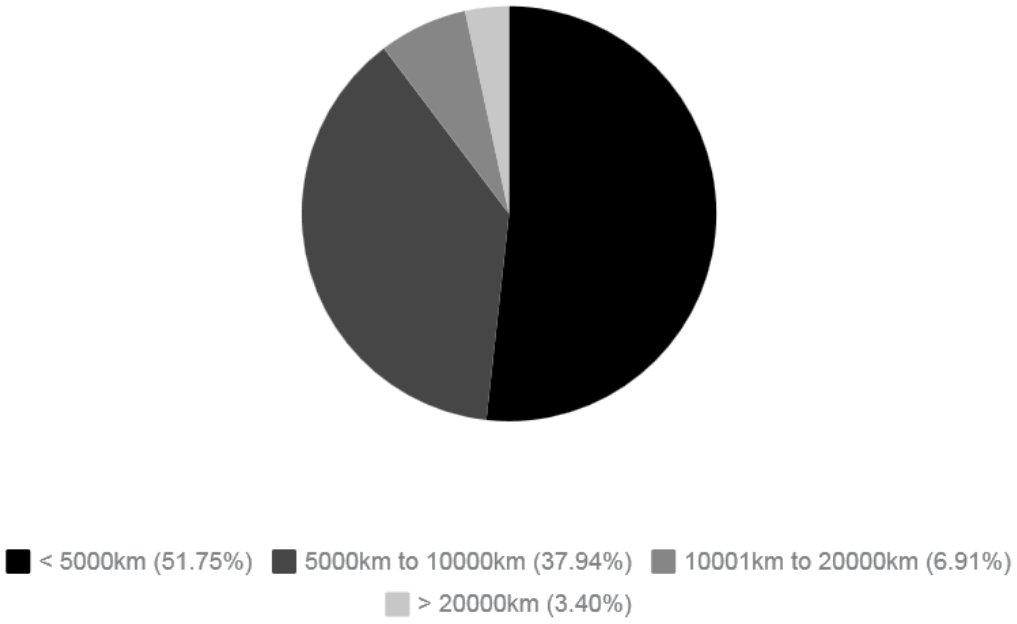


Figure 4.6: Experiment 1 - Participants' mileage per year

'low' task difficulty condition participants just simply repeated the digit that was last presented to them. In between digits, there was an inter-stimulus interval of five seconds. In the 'medium' task difficulty condition they were asked to continuously sum up the two digits that were presented last. Between digits there was an inter-stimulus interval of five seconds. The 'high' task difficulty condition differed from the medium condition with respect to the demand regarding time, as the inter-stimulus interval was reduced to three seconds. Strictly speaking this manipulation does not increase the cognitive task difficulty compared to the medium task difficulty when looking at a single sub-task. However, considering that participants in the high task difficulty condition had to perform more calculations in same time intervals as in the medium task difficulty condition, this introduced overall higher cumulative cognitive task demands. The PASAT digits were presented to the participants over headphones to avoid distracting external noises. The sequence of the three task difficulty conditions was counterbalanced, resulting in six sequences that were each completed by five of the 30 participants. The driving task described in subsection 4.2.1 was the same for all sequences. In the following, task difficulty levels are abbreviated as 'low', 'medium' and 'high' condition for the purpose of easier differentiation. It is, however, important that these labels do not describe absolute task difficulties, but should be interpreted as relative differences between the three levels.

Dependent variables

Dependent variables are grouped in three different main categories commonly agreed upon in mental workload research (O'Donnell & Eggemeier, 1986) and operationalized as follows:

Performance measures

This category includes primary-task measures of driving as well as secondary-task measures from the PASAT.

1. Primary-task performance:

   Primary-task performance of all three task difficulty conditions in the LCT was collected for all four driving tracks of the driving task introduced in subsection 4.2.1. Lane deviation (Mdev) was calculated as the mean difference between the actual driving course (red line) and a normative model (green line) in m as illustrated in figure 4.7.

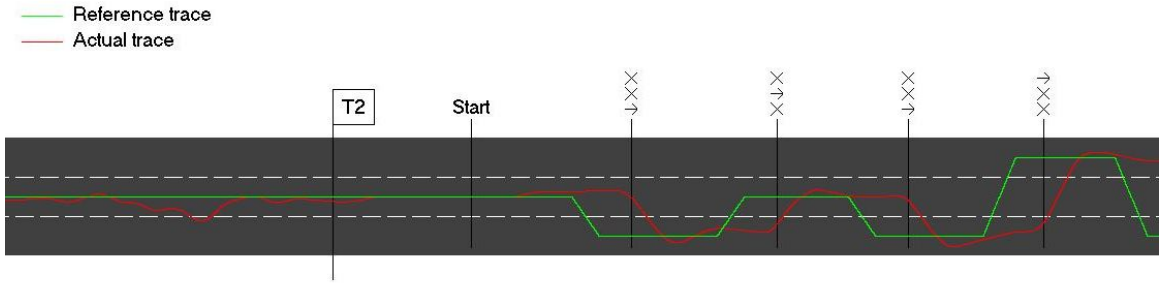


Figure 4.7: Experiment 1 - Example of normative (green) vs. actual (red) lane change behavior from LCT analysis tool by Mattes (2003)

2. Secondary-task performance:

   Secondary-task performance was manually logged by the experimenter and later analyzed as the percentage of errors in the PASAT compared to optimal performance (PaErrPerc) for the p1 and p2 driving track for each task difficulty condition. The logging was accomplished by writing the participants' responses next to the correct answers in a prepared table. To ensure the correct assignment of the participants' responses to the correct answers the experimenter listened to the PASAT digits over headphones while they were presented to the participants.

Subjective measures

1. To assess subjective workload, a German version of the NASA-TLX, originally developed by Hart and Staveland (1988), was adopted from Seifert (2002) (see appendix A.1.1):

   (a) Raw values for each NASA-TLX dimension on a scale from 0-100

   (b) Individually weighted overall score on a scale from 0-100

Physiological measures

Physiological measures, except for the speech signal features, were recorded for all driving tracks. For the speech signal features, only the tracks 'p1' and 'p2' of each experimental condition were recorded, as no speech was required in driving tracks 'pre' and 'post'.

1. Cardiovascular Measures:
   (a) Heart rate (HR) in *bpm*
   (b) Temperature at the tip of the nose (NTT) in °C
2. Ocular Measures:
   (a) Fixation duration (FD) in *ms*
   (b) Fixation count (FC) as number of fixations
   (c) Horizontal fixation dispersion (SDHor) as the standard deviation of fixation positions in x direction
   (d) Pupil size as pupil diameter in *mm*
3. Speech signal measures:
   (a) Fundamental frequency of the voice (F0) in *Hz*
   (b) Voice intensity as the root mean square (RMS) energy over five pitch pulses (RMSInt) in dB (see Shue, 2010b)

4.2.4 Procedure

Following a general introduction to the experimental procedure (see appendix A.1.3), participants were asked to fill out a demographic questionnaire (see appendix A.1.2) and signed an informed consent form (see appendix A.1.4). After that, the participants were instructed on the placement of the ECG electrodes and then attached them to their chest as illustrated in figure 4.8. Afterwards they sat down in the driver's seat. They were asked to bring the seat into a comfortable position, so that they could easily reach the pedals and the steering wheel. Subsequently, all technical devices were arranged to the individual's position followed by an introduction to the driving and the secondary-task (see appendix A.1.5). Then they were asked to practice the tasks individually to ensure both tasks were fully understood. They were randomly assigned to one of the six task difficulty condition sequences to balance training effects. The subjects were then instructed that their primary-task was driving and that they should perform the PASAT as good as possible without disregarding the driving task. The focus on driving was chosen, as it is believed that this most closely mimics attentional and motivational strategies in real world driving. Subsequently, they rested for three minutes before being told which task difficulty

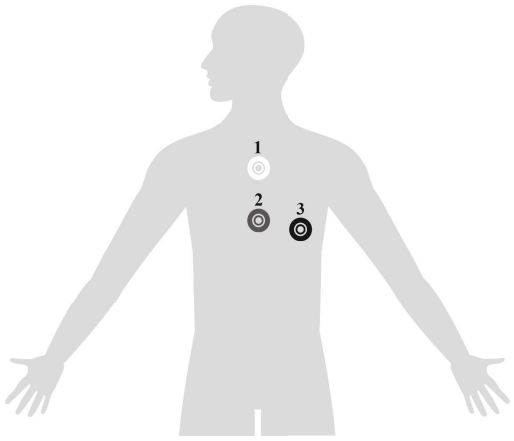


Figure 4.8: Experiment 1 - Placement of ECG electrodes as implemented in the experiment adapted from Varioport manufacturer's recommendation

condition they were going to perform next. The eye-tracking system was then calibrated to the participant. The calibration criteria was defined as an accuracy better than 0.5° although in some cases this threshold needed to be adapted to $< 1°$. Every participant performed all three task difficulty conditions (low/medium/high). After each trial the participants were asked to fill out a paper version of the NASA-TLX and to rest for a period of three minutes before starting the next trial. Each trial lasted for approximately 10 minutes. After completing the entire experiment, the ECG electrodes were detached, the participants were thanked and rewarded.

4.2.5 Data pre-processing

Pre-processing and data analysis was mostly done with the free statistics software R statistics version 3.0.1 (R Core Team, 2015a). All R packages that were used are documented in appendix A.2.1. Before the actual data analysis, specific pre-processing had to be performed for all physiological measures. Pre-processing was done as follows:

Varioport ECG data

At first each Varioport raw data file was converted to a .txt file using the Variograf software (revision 4.78). To generate HR time series for these files, Kubios Software (version 2.2) was used (Tarvainen et al., 2014). Initially, the ECG signal was visually checked for correct R-peak detection and artifacts and if necessary corrected according to a procedure proposed by L. Mulder (1992). Based on this, the data of two subjects had to be excluded from data analysis due to noisy recordings.

Eye-tracking data

During eye blinks the eyelids partly or fully cover the pupil leading to invalid pupil size values. Therefore the raw eye-tracking data was corrected by removing these episodes using the blink detection algorithm developed by Pedrotti, Lei, Dzaack, and Rötting (2011). Fixations were calculated with the open source software OGAMA (version 5.0) (Voßkühler, 2015). The maximum dispersion of gaze points considered to belong to a fixation was set to 20px and a minimum fixation duration of 80ms was defined.

Thermography data

The thermography camera was set up to take an image every four seconds, resulting in approximately 30 images for each driving track. The images were semi-automatically analyzed with a custom-built tool that allowed the calculation of mean values over predefined regions of interest (ROI). For the calculation of the temperature at the tip of the nose a fixed size ROI of 5px*5px was defined (see figure 2.6 in chapter two). For further details please see Hammerschmidt (2013).

Speech signal data

For the analysis of the speech signal features the software Voicesauce (version 1.15) (Shue, 2010a, 2010b) was used. Internally, Voicesauce resorted on the Praat algorithms (Boersma, 2002) for the fundamental frequency estimation. The settings used for pitch estimations were the same for all participants and can be found in Appendix A.2.2. Excel .csv files with F0 and voice intensity time series resulted from this procedure.

4.2.6 Data analysis

To compensate for individual differences, the physiological and driving performance data was averaged over the two driving-only tracks (pre/post) and the dual-task tracks (p1/p2) for each task difficulty condition. Then the percentage of the difference between the driving-only tracks and the dual-task tracks was calculated so that the parameters could be interpreted and compared as changes from driving-only phases to driving with secondary-task phases. As no speech recordings were made when participants had no secondary-task, the absolute values of fundamental frequency and voice intensity were statistically analyzed. Within-subjects ANOVAs were used for

statistical testing. Distributions of the variables were visually assessed by checking the frequency histograms for normality. In cases of moderate deviations from normality for all task difficulty conditions, an ANOVA was performed, as it has been shown to be robust against such violations of the assumptions (Glass, Peckham, & Sanders, 1972; Harwell, Rubinstein, Hayes, & Olds, 1992; Lix, Keselman, & Keselman, 1996). In case of a violation of the assumption of sphericity, a Greenhouse-Geisser correction was applied. Post-hoc analyses of the main effects were performed, using pairwise t-tests with Bonferroni correction. All statistical analyses were conducted using R statistics version 3.0.1, ANOVAs were calculated using the 'ez' package (Lawrence, 2015) and post-hoc comparisons using the 'stats' package (R Core Team, 2015b).

4.3 Results

4.3.1 Performance measures

For statistical testing of the main effect of task difficulty the α-level was adjusted to $p = .2$ as no differences were expected. This is a conservative approach to reduce the probability of committing a β -error (Bortz, 2005). The lane change performance in the driving task showed a slight decline over the three task difficulty conditions with best driving performance in the 'low' condition (see table 4.1 and figure 4.9a), resulting in a significant main effect of task difficulty, $F(2, 58) = 3.45, p < .05, \eta_p^2 = .11$. However, the changes compared to driving-only were small and differences between the three conditions did only reach statistical significance between the 'low' and the 'high' condition ($p_{low-high} = .051$) in post-hoc analysis.

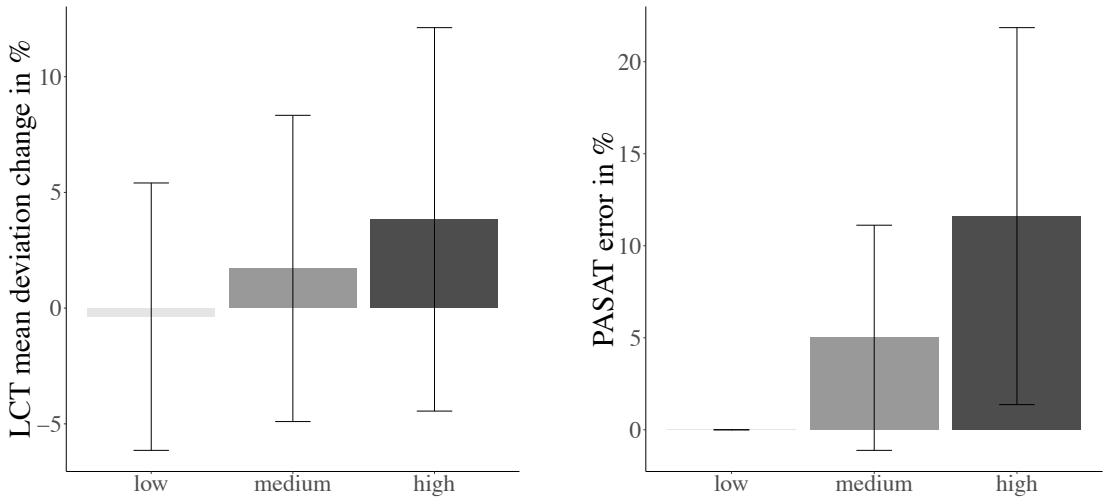
The performance in the PASAT can be found in table 4.1 and figure 4.9b. Whereas in the 'low' condition no errors were made, there was an incremental increase of errors in the cognitively more demanding conditions ('medium' and 'high'). A non-parametric Friedman test of differences rendered a χ^2 value of 41.54, which was significant ($p < .001$). Post-hoc analysis was performed with a Friedman-Nemenyi test. The results revealed that secondary-task performance discriminated between all three task difficulty conditions ($p_{low-medium} < 0.01$, $p_{low-high} < 0.001$ and $p_{medium-high} < 0.05$).

4.3.2 Subjective measures

The weighted overall workload ratings increased as a function of the task difficulty, showing lowest ratings for the easiest condition (table 4.2). The results of the ANOVA revealed a highly significant effect of task difficulty on workload, $F(2, 58) = 72.56, p <$

Table 4.1: Experiment 1: Means (M) and standard deviations (SD) of the driving and PASAT performance for the three task difficulty conditions

	low		**medium**		**high**	
	M	*SD*	*M*	*SD*	*M*	*SD*
Mdev %	-0.37	5.78	1.72	6.62	3.84	8.29
PaErrPerc	0	0	5.00	6.12	11.61	10.24



(a) Mean LCT deviation (Mdev) compared to driving-only baseline in %

(b) Percentage of errors in the PASAT compared to optimal performance in % (PaErrPerc)

Figure 4.9: Experiment 1 - Mean values of performance measures for each task difficulty. Error bars represent the standard deviations

Table 4.2: Experiment 1 - NASA-TLX mean (M) overall workload ratings along with the standard deviations (SD)

	low		**medium**		**high**	
	M	*SD*	*M*	*SD*	*M*	*SD*
Overall workload	27.9	14.2	52.3	15.8	64.9	17.5

Table 4.3: Experiment 1 - ANOVA and post-hoc results for all NASA-TLX dimensions

	Main effect of task difficulty	low - medium	low-high	medium-high
Mental Demand	$F(2, 58) = 111.24, p < .001, \eta_p^2 = .79$	$p < .001$	$p < .001$	$p < .05$
Physical Demand	$F(2, 58) = 2.30, p = .11, \eta_p^2 = .07$	–	–	–
Temporal Demand	$F(2, 58) = 46.31, p < .001, \eta_p^2 = .61$	$p < .001$	$p < .001$	$p < .001$
Performance	$F(2, 58) = 15.13, p < .001, \eta_p^2 = .34$	$p = .05$	$p < .001$	$p < .05$
Effort	$F(1.55, 44.95) = 40.02, p < .001, \eta_p^2 = .58$	$p < .001$	$p < .001$	$p = .24$
Frustration	$F(2, 58) = 25.07, p < .001, \eta_p^2 = .46$	$p < .001$	$p < .001$	$p = .06$

$.001, \eta_p^2 = .71$. A post-hoc analysis of this effect showed significant differences between all three task difficulties (all pairwise comparisons: $p < .001$). The individual dimensions of the NASA-TLX mostly exhibited the same pattern as the overall workload, i.e. displaying increased ratings with increasing task difficulty except for physical demand, which showed no significant changes regarding the three task difficulties (figure 4.10). Detailed ANOVA results for all dimensions can be found in table 4.3. Post-hoc analyses revealed that the dimensions mental demand and temporal demand differentiate between all three task difficulties whereas the effort and frustration ratings only indicate a difference between the 'low' and the two higher conditions. The performance dimension differentiates between the two higher conditions but only reaches marginal significance between the 'low' and 'medium' condition.

4.3.3 Physiological measures

Cardiovascular measures

Heart rate

Heart rate increased as a function of task difficulty compared to the baseline driving sections resulting in a significant main effect of task difficulty, $F(2, 52) = 38.64, p < .001, \eta_p^2 = .60$. The smallest differences between the driving baselines and the respective driving with secondary-task tracks were found with regard to the 'low' condition and highest for the 'high' condition (see table 4.4 and figure 4.11a). Pairwise comparisons revealed significant differences between the 'low' and the

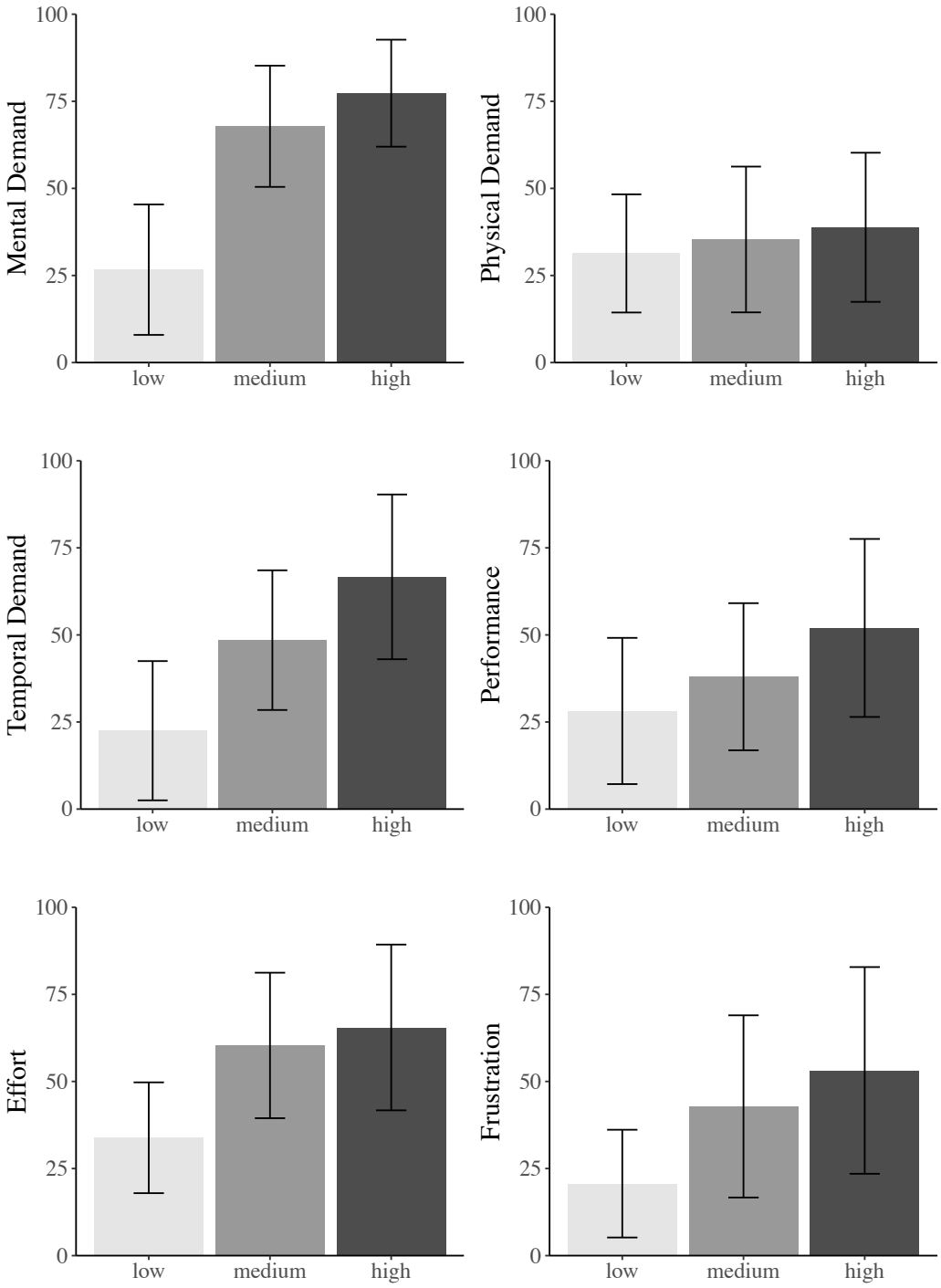
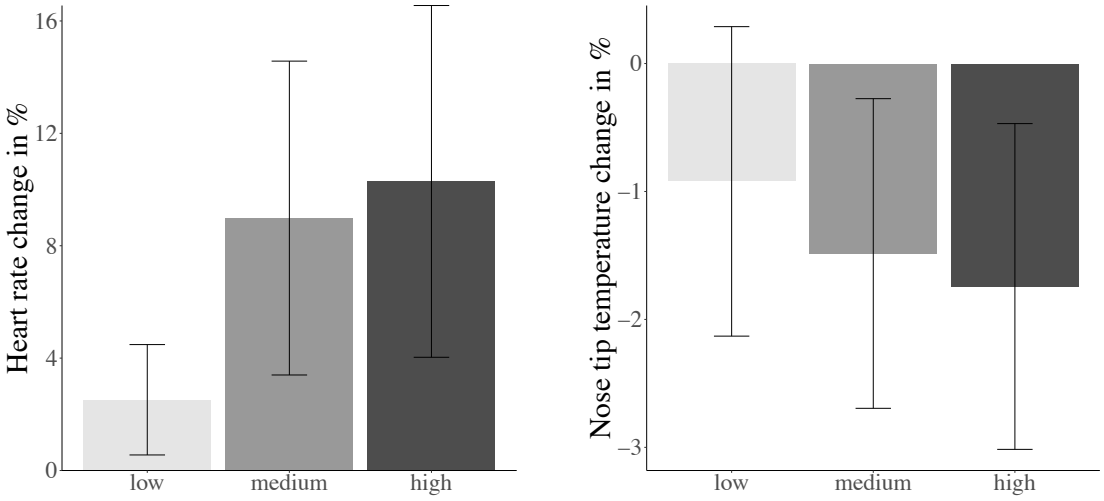


Figure 4.10: Experiment 1 - Mean values and standard deviations of the six NASA-TLX dimensions

'medium' respectively 'high' condition, but failed to differentiate between 'medium' and 'high' condition ($p_{low-medium} < 0.001$, $p_{low-high} < 0.001$ and $p_{medium-high} = .44$).

Nose tip temperature

Temperature at the tip of the nose decreased as a function of task difficulty compared to the baseline driving sections (see table 4.4 and figure 4.11b), resulting in a significant main effect for task difficulty, $F(2, 58) = 11.95, p < .001, \eta_p^2 = .29$. Post-hoc analyses showed that the drop in temperature only differentiated between the 'low' and 'medium' respectively 'high' ($p_{low-medium} < .05$ and $p_{low-high} < .001$) conditions but not between 'medium' and 'high' condition ($p_{medium-high} = .21$).



(a) Change in heart rate compared to driving-only baseline in %

(b) Change in nose tip temperature compared to driving-only baseline in %

Figure 4.11: Experiment 1 - Mean values of cardiovascular measures for each task difficulty. Error bars represent the standard deviations

Ocular measures

Fixation duration and fixation count

Results for fixation parameters showed an increase in the number of fixations as a function of task difficulty compared to baseline driving sections whereas fixations duration decreased (see table 4.4 as well as figure 4.12a and figure 4.12b). Both variations resulted in a significant main effect of task difficulty (fixation count: $F(2, 58) = 3.40, p < .05, \eta_p^2 = .10$ and fixation duration: $F(2, 58) = 8.62, p < .01, \eta_p^2 = .23$).

Table 4.4: Experiment 1 - Mean values (M) and standard deviations (SD) of the percentage of difference between driving with secondary-task and driving-only for each task difficulty condition

	low		**medium**		**high**	
	M	*SD*	*M*	*SD*	*M*	*SD*
Heart rate	2.52	1.96	8.99	5.59	10.29	6.26
Nose tip temperature	-0.92	1.21	-1.49	1.21	-1.74	1.27
Fixation duration	-3.24	7.64	-7.36	10.51	-10.34	11.98
Fixation count	3.14	8.55	5.49	13.32	8.45	14.48
Horizontal fixation dispersion	-0.20	7.92	8.88	12.08	14.17	18.59
Pupil size	1.07	2.41	7.84	3.41	10.44	3.51

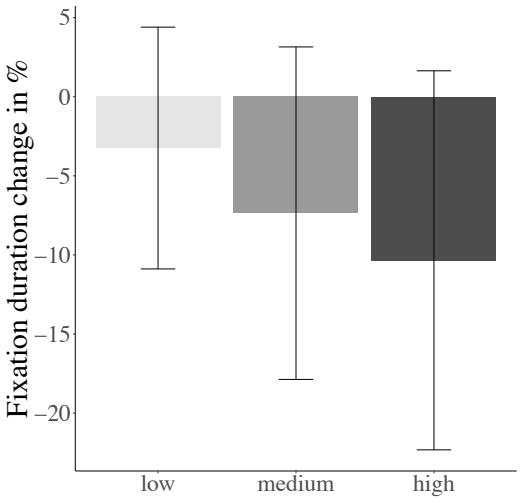
However both parameters showed little sensitivity in post-hoc comparisons. Only the difference between the 'low' and the 'high' condition was significant with regard to fixation duration ($p_{low-high} < 0.01$).

Horizontal fixation dispersion

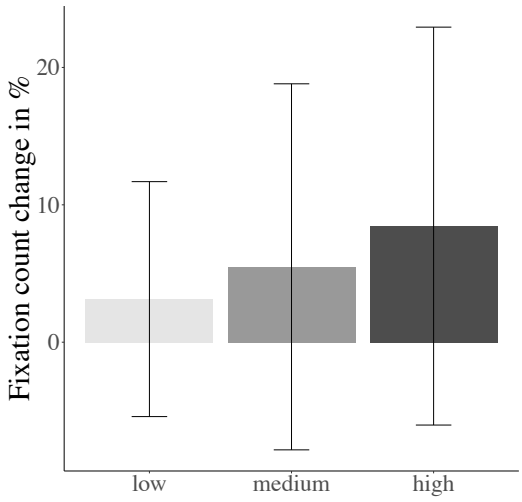
Horizontal dispersion of fixations increased as a function of task difficulty compared to baseline driving (see table 4.4 and figure 4.12c). Whereas there was almost no difference in the 'low' condition compared to baseline driving, horizontal variability leaps in the 'medium' and 'high' condition compared to baseline driving resulting in a significant main effect, $F(1.619, 46.951) = 10.11, p < .001, \eta_p^2 = .26$. Pairwise comparisons exhibited the same pattern as nose tip temperature and heart rate by not differentiating between the 'medium' and the 'high' condition ($p_{low-medium} < 0.05$, $p_{low-high} < 0.01$ and $p_{medium-high} = 0.19$).

Pupil size

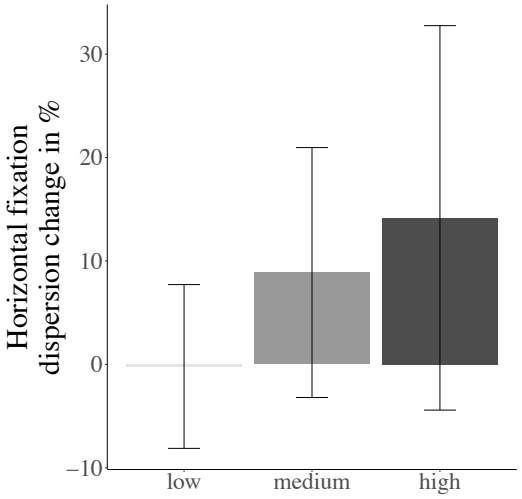
Task difficulty showed a highly significant impact on pupil size, $F(1.66, 48.04) = 141.71, p < .001, \eta_p^2 = .83$. Whereas in the 'low' condition the increase compared to the baseline driving section of the pupil was small, the more demanding secondary-task difficulties resulted in increasingly higher pupil dilations (see table 4.4 and figure 4.12d). All pairwise comparisons were significant, indicating that pupil size can differentiate between all three conditions ($p_{low-medium} < 0.001$, $p_{low-high} < 0.001$ and $p_{medium-high} < 0.001$).



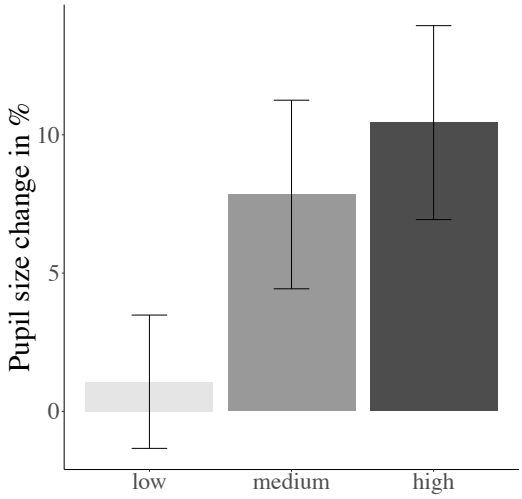
(a) Change in fixation duration compared to driving-only baseline in %



(b) Change in fixation count compared to driving-only baseline in %



(c) Change in horizontal fixation dispersion compared to driving-only baseline in %



(d) Change in pupil size compared to driving-only baseline in %

Figure 4.12: Experiment 1 - Mean values of ocular measures for each task difficulty. Error bars represent the standard deviations

Speech signal features

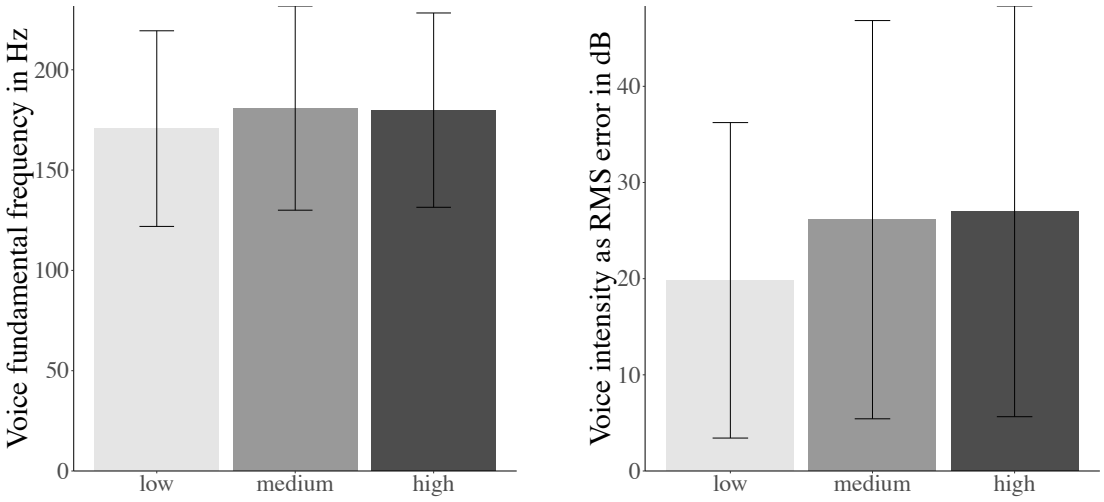
As there was no speech present on the driving-only baseline tracks the following results are reported as absolute values of driving with the PASAT present. Additionally, the factor gender as between-subjects factor was added to the ANOVA, as i.e. fundamental frequency is higher for female speakers compared to male speakers (Baken & Orlikoff, 2000). Overall, fundamental frequency and mean voice intensity increased as a function of task difficulty compared to the baseline driving

Table 4.5: Experiment 1 - Mean absolute values (M) and standard deviations (SD) of the speech signal features for each task difficulty condition

	low		***medium***		***high***	
	M	*SD*	*M*	*SD*	*M*	*SD*
Voice fundamental frequency in Hz	170.70	48.76	180.90	50.89	179.9	48.44
Voice intensity as RMS energy in dB	19.83	16.40	27.00	21.35	26.00	20.71

sections (see table 4.5 and figures 4.13a and 4.13b) resulting in significant main effects of the factor task difficulty (F0: $F(2, 56) = 6.21, p = .004, \eta_p^2 = .18$ and RMSInt: $F(2, 56) = 10.84, p < .001, \eta_p^2 = .28$). However, both measures did not differentiate between the 'medium' and the 'high' condition (F0: $p_{low-medium} < 0.001$, $p_{low-high} < 0.001$, $p_{medium-high} = 1$; RMSInt: $p_{low-medium} < 0.001$, $p_{low-high} < 0.001$, $p_{medium-high} = 1$).

Additionally, the factor gender was highly significant for both measures (F0: $F(1, 28) = 113.09, p < .001, \eta_p^2 = .80$ and RMSInt: $F(1, 28) = 36.12, p < .001, \eta_p^2 = .56$). Overall female participants exhibited higher fundamental frequencies compared to the male participants (female: $M = 219.62, SD = 25.38$, male: $M = 134.79, SD = 24.51$). Inversely the voice intensity for females was lower compared to males (female: $M = 10.37, SD = 7.28$, male: $M = 38.27, SD = 18.41$). No significant interaction of gender and task difficulty was found for either of the two measures, indicating that the effect of task difficulty was the same for both genders.



(a) Voice fundamental frequency in Hz (b) Voice intensity as RMS energy in dB

Figure 4.13: Experiment 1 - Mean values of speech signal features for each task difficulty. Error bars represent the standard deviations

4.4 Discussion

The aim of this experiment was to evaluate different physiological measures regarding their capability to differentiate between differences in cognitive workload evoked by an externally paced auditory-verbal cognitive secondary-task. For this reason hypotheses regarding performance, subjective and physiological measures were formulated. First of all the impact of task difficulty on performance and subjective measures will be discussed to infer to what extend the variation of the task demands invoked different cognitive workload levels. Afterwards the physiological consequences will be interpreted with regard to this.

4.4.1 Performance measures

Primary-task performance

It was expected that lane-keeping performance is not affected by the secondary-task. However, the mean deviation from a normative line increased with increasing task difficulty. It should be noted that, even though the difference was significant, the absolute deviations are low (low: $\sim -0.6cm$, medium: $\sim 2.8cm$ and high: $\sim 6.4cm$ condition) and no pairwise comparison showed significant differences. A reason for the change could lie in the parameter that is used to measure performance. It is assumed that the LCT parameter 'mean deviation from a normative line in m' is not perfectly

suitable to measure continuous lane-keeping as it comprises straight driving sequences as well as event-related lane changes and therefore reflects lane-keeping and lane changing performance in this one parameter (Mattes & Hallén, 2008). Lane change performance is expected to be impacted by auditory-verbal secondary-tasks, whereas lane-keeping is expected to be more robust against such influences (see section 3.4). Furthermore it should be noted that the LCT is validated at a speed of 60km/h. In the experiment, however, a speed of 80km/h was chosen. Due to the counterbalanced design, it can be assumed that the speed choice did not introduce a systematic effect on driving performance for the three task difficulties. Nevertheless, particularly at this speed delayed reactions to the LCT signs are very costly with respect to the deviation from the normative line (see figure 4.7). Therefore it can be argued that the lane changes contribute substantially more to the overall deviation from the green normative line as compared to the deviations resulting from lane-keeping sections. In summary, the results are likely to predominantly represent event-related lane changing behavior. The results however, are in line with studies that used the same LCT metrics to assess the impact of auditory-verbal tasks on overall lane-keeping, reporting higher mean deviations compared to driving-only conditions (Maciej & Vollrath, 2009) as well as increasing task difficulties (Ross et al., 2014). Nevertheless in following experiments, dependent variables that clearly differentiate between lane-keeping and event-related performance should be used, as both aspects of driving rely on different attentional resources (Horrey & Wickens, 2006). Changes in those measures induced by a secondary-task therefore imply qualitatively different workload related changes.

Secondary-task performance

Changes in mental workload were hypothesized to be mainly reflected in secondary-task performance as participants were instructed to perform the secondary-task as good as possible without disregarding the driving task. According to the results presented in subsection 4.3.1, this hypothesis can be confirmed. The results show that the low task difficulty condition was handled perfectly by all participants which can be ascribed to the very simple nature of the task. The immediate repetition of the presented digits did not involve complicated working memory processes or higher executive functions and can therefore be assumed to be a highly automatic task with very little attentional demands. In contrast, the two higher demanding task difficulty conditions involved a mental arithmetic task. Mental arithmetic has been shown to put high demands on central executive resources (see e.g Hubber, Gilmore, & Cragg, 2014). These resources are most likely shared with the driving task. Additionally, the

PASAT requires continuously keeping the last digit that was presented in memory and recalling it after the next digit is presented to sum them up, thus it continuously loads working memory. This is reflected in significantly higher error rates with increased task difficulty. With increasing time demands between medium and high task difficulty, the error rate rises subsequently. Overall, the secondary-task is sensitive to the cognitive demands as well as the temporal demands imposed by the PASAT. These changes in performance can be interpreted as a compensational strategy to cope with increasing task demands aimed at maintaining primary-task performance at an acceptable level.

In summary, the results from primary- and secondary-task suggest that three different cognitive workload levels were induced by the task combination. It is assumed that optimal workload levels constitute in optimal performance (see 2.1.4). To validate this assumption, the subjectively perceived workload of the participants can be used as an additional indicator. With respect to the primary-and secondary-task results, it can be assumed that compared to the other two induced workload levels, optimal workload was present for the easiest task difficulty condition. With increasing task difficulty both performance indicators increasingly declined, suggesting a transition to overload workload regions.

4.4.2 Subjective measures

The NASA-TLX ratings back up the aforementioned assumption that indeed three different cognitive workload levels were induced. As hypothesized, the overall workload ratings increased with increasing task demands and differentiated between all three task difficulties. Interestingly, the single dimensions of mental demand, temporal demands and performance of the NASA-TLX seem to be very sensitive to the inter-task differences (task difficulty and inter-stimulus interval) whereas the frustration and effort scales do not differentiate between the medium and the high task difficulty. These results indicate higher differential sensitivity of the mental, temporal and performance NASA-TLX dimensions to the induced workload differences compared to the effort and frustration ratings. Effort and frustration on the other hand cannot easily be rated quantitatively and are more likely to reflect overall subjective energetic demands. Similar ratings for the medium and high condition indicate that participants either limited their effort at an acceptable level or a ceiling level was reached, where further effort investment would not have resulted in better performance. Additionally, similar ratings on the physical demand dimension for all

task difficulty conditions suggest no confounding physical influences on physiological measures.

As it was laid out in chapter 2 and 3, mental workload is mainly influenced by the structure and difficulty of the tasks, but can also be affected by factors like fatigue, emotional states, age, personality etc. (see 2.1.5). Therefore, different task difficulties do not necessarily result in different workload levels. However, with respect to the presented results on performance and subjective workload, it can be assumed that three different cognitive workload levels were imposed by the three task difficulty conditions. The following subsection will therefore discuss the sensitivity of the physiological measures to the three different workload levels.

4.4.3 Physiological measures

Changes in physiological measures reflect complex changes of sympathetic and parasympathetic activity induced by mental workload, particularly as a result of effort that was invested in a task (see subsection 2.2.3 and section 3.6 of this thesis). As described in section 3.6, the sensitivity of different physiological measures to different cognitive workload levels varies. The following discussion will therefore critically examine the results with regard to the factors that might limit their sensitivity. Furthermore, the section will focus on their potential to be integrated in a multi-measure approach to differentiate between different cognitive workload levels.

Cardiovascular measures

Heart rate

Heart rate is subject to parasympathetic and sympathetic changes in activity (see subsection 2.2.3 and 3.6.1). However, in highly demanding situations the influence of sympathetic activity is predominantly responsible for increases in heart rate (Kahneman et al., 1969; Jorna, 1992; L. Mulder, 1992; Wilson, 1992; L. B. Mulder et al., 2004). Heart rate was expected to increase with increasing task demands, especially with high working memory demands. The results of the experiment are in line with this hypothesis, exhibiting increased heart rates, especially with the task difficulties involving mental arithmetic tasks. However, heart rate did not differentiate between medium and high task difficulty conditions. This pattern reflects what has been reported in other studies, showing little sensitivity in higher cognitive workload areas (Engström et al., 2005; Lenneman et al., 2005; Reimer & Mehler, 2011). In the present case, the results indicate that heart rate is not sensitive to additional temporal

demands imposed by the high task difficulty condition. It is possible that demands in the highest task difficulty condition exceeded the capabilities of the participants, hence that additional effort investment did not result in the maintenance of performance in both tasks. Participants therefore might have compensated the increasing task demands by lower secondary-task performance at roughly similar effort expenditure (see figure 4.10). Alternatively, a ceiling effect of the physiological reaction might also be a valid explanation, indicating that even if more effort had been invested by the participants in the high task difficulty condition, the physiological reaction would have been limited. However, this explanation is improbable, as the absolute values of heart rate in the high task difficulty condition ($M = 84.33bpm, SD = 13.80bpm$) did not reach physiological limits. Heart rate can easily reach 200-220 bpm during physical activity (AmericanHeartAssociation, 2015). In summary, heart rate did not differentiate between all three cognitive workload levels that were induced but clearly discriminated between relatively low workload conditions and medium respectively high cognitive workload levels. Interestingly, this pattern almost exactly depicts the NASA-TLX effort ratings, suggesting that heart rate functions more as an indicator of mental effort than as an indicator of task difficulty or overall workload. Nevertheless, it is necessary to examine whether a further increase in cognitive task difficulty rather than an increase in temporal demands would increase heart rate additionally. Mehler et al. (2009) e.g. report heart rate discriminating between three n-back task difficulties.

Nose tip temperature

The results on facial temperature features confirmed the hypothesis that the temperature at the tip of the nose drops significantly with increasing task difficulty. It is assumed that changes in nose tip temperature reflect solely sympathetic activity leading to less blood supply to the blood vessels under the skin (Drummond, 1994; Rimm-Kaufman & Kagan, 1996; Iani et al., 2004; Miyake et al., 2009). The results clearly differentiate between the low vs. medium and low vs. high task difficulty conditions but fail to discriminate between the medium and high condition and therefore exhibit the same pattern as heart rate. One explanation for the insensitivity in higher workload regions is that the effort invested in the two different tasks is the same, as the performance and subjective measures suggest. Again, a physiological ceiling effect is unlikely, as the absolute temperature values at the nose tip did not reach physiological limits ($M = 32.63°\text{C}, SD = 1.82°\text{C}$). Studies researching the influence of cold temperature on nose tip temperatures have shown that skin temperatures

can decrease to $2 - 10°C$ (Gavhed, Mäkinen, Holmér, & Rintamäki, 2000; Brajkovic & Ducharme, 2006). This study reveals the potential of nose tip temperature as a cognitive workload indicator. To the author's knowledge, this has been the first study that examined the impact of three different cognitive task difficulties on nose tip temperature. However, it has to be noted that the change in skin temperature is only a secondary-effect of sympathetic changes and is therefore a slow indicator and might not be suitable in applications where workload assessment is time critical. Nevertheless, it is a promising indicator of effort that can be assessed contactless.

Summary on cardiovascular measures

Heart rate and temperature at the nasal tip have consistently shown to be promising indicators of cognitive workload. Both measures are predominately influenced by sympathetic activity and therefore reflect the effort component of cognitive workload. Heart rate as well as nose tip temperature can easily be acquired non-intrusively and even contactless and are relatively robust to measurement artifacts, as e.g. compared to heart rate variability (see subsection 2.2.3). The results suggest that in multi-measure approaches of cognitive workload both, heart rate and nose tip temperature, might probably be redundant. Whether this is desirable or not depends on system requirements. The use of heart rate in time critical situations is preferable as the timely latency of changes in nose tip temperature is greater compared to changes in heart rate because it is a secondary-effect of cardiovascular changes.

Ocular measures

As described in subsection 2.2.3 and section 3.6 of this thesis ocular measures reflect attention allocation processes through eye-movement parameters and ANS related changes through e.g. variations in pupil size. Therefore the different measures are assumed to reflect different aspects of cognitive workload. Whereas eye-movement parameters are more likely to correlate with visual demands of the task, ANS related changes are expected to reflect the overall computational and compensatory effort invested by the participant to cope with task demands.

Fixation duration and fixation count

Inconsistent with the hypothesis, fixation durations decreased whereas fixation counts increased with increasing task difficulty. Additionally the results indicate high

inter-individual differences in the response to task demands, as the high standard deviations suggest. Firstly, these results implicate that even without a change in visual demands, a change in visual attention allocation occurs due to a central interference of attentional resources between the visual-manual driving task and the auditory-verbal secondary-task. The direction of change is not in line with most of the previous results using highly demanding auditory-verbal secondary-tasks (e.g. Recarte & Nunes, 2000; Pomplun et al., 2001; Niezgoda et al., 2015). They are, however, in line with results reported by Recarte and Nunes (2000) using a simple verbal secondary-task and to some extend with the results by Tsai et al. (2007), which indicate that decreasing fixation durations precede errors in the secondary-task. As the results of performance and subjective measures suggest (see subsections 4.4.1 and 4.4.2), the secondary-task is clearly demanding, at least with regard to the medium and high task difficulty condition. That means that the results cannot be explained by the simplicity of the task demands. Decreasing fixation durations have previously been linked to errors in the secondary-task (Tsai et al., 2007). This would be a valid explanation for the results in the medium and high task difficulty condition but would not explain why even in the low condition the fixation durations substantially decrease compared to the driving-only baseline sections. It may be more likely that the ratio of environmental-intake to environmental-rejection increasingly shifts to a dominance of intake to protect the driving performance against increasing secondary-task demands. However, this seems contradictory as the LCT is predictable with regard of the 'when and where' of a new lane change, so that extensive scanning of the visual field does not seem to be necessary.

Another line of argument for the fixation results could also be an interference of manual resources rather than an interference of central resources. Both involve manual motor resources to some extend. It can be argued that the variations in fixation duration and count are a result of an overlap of manual resources responsible for speech production and eye-movement generation. This, however can be ruled out as single explanation for the fixation count and duration variations. If the interference of manual resources would be the only contributor to the variations, one would expect comparable results between the low and the medium task difficulty condition and a clear difference between low respectively medium condition with an inter-stimulus interval of five seconds and the high condition with an inter-stimulus interval of three seconds. In contrast, the variations appear to linearly increase with increasing task demands (see table 4.4).

In summary, the benefit of fixation duration and count as cognitive workload indicators is limited. Due to their conjunction with the dynamic task demands (see subsection 2.2.3 and 3.6.2), they can hardly be used as generic workload measures but always have to be interpreted with regard to the specific task demands. So particularly in applied scenarios like real-world driving, fixation measures might contribute in cases where the reaction to specific events is of interest, e.g. assuming a potential hazard can be detected by sensors, whether and how drivers allocate their attention to critical events. However, as measure of a 'tonic' workload state over a longer period of time, fixation measures may be of lesser value.

Horizontal fixation dispersion

According to the hypothesis, horizontal fixation dispersion was expected to decrease with increasing task difficulty. In contrast to the rich body of evidence supporting this assumption (see subsection 3.6.2), the fixation dispersion increased with increasing task demands in this experiment. According to Recarte and Nunes (2003), the concentration of gaze dispersion under cognitively demanding situations can be interpreted as a compensational mechanism to focus visual attention on the most critical aspects of driving. In the present case lane changes can be assumed to be the most demanding sub-tasks of driving. Therefore, the signs indicating lane-changes were probably prioritized higher by the participants than the road ahead regarding visual attention. Thus, the most significant aspects of the driving task were not ahead on the road, but on the left and right side of the road. This effect increases with the proximity of the signs in relation to the driver (see figure 4.14). Additionally, the center of the road changes with every lane change. While this explains a general broad spread of fixations over the visual scene, it does not explain the task difficulty differences with respect to horizontal fixation dispersion. It is possible that participants concentrated their gaze towards the center of the road between lane changes. However, in order to identify the upcoming lane change, participants had to fixate on the signs to the left and the right of the road. Assuming that this shift of visual attention from the road towards the signs occurred increasingly delayed with increasing workload levels, participants in higher workload regions had to direct their gaze farther to the left and the right to fixate the oncoming signs as compared to participants in lower workload regions. This very likely led to the increase in horizontal fixation dispersion with increasing workload levels.

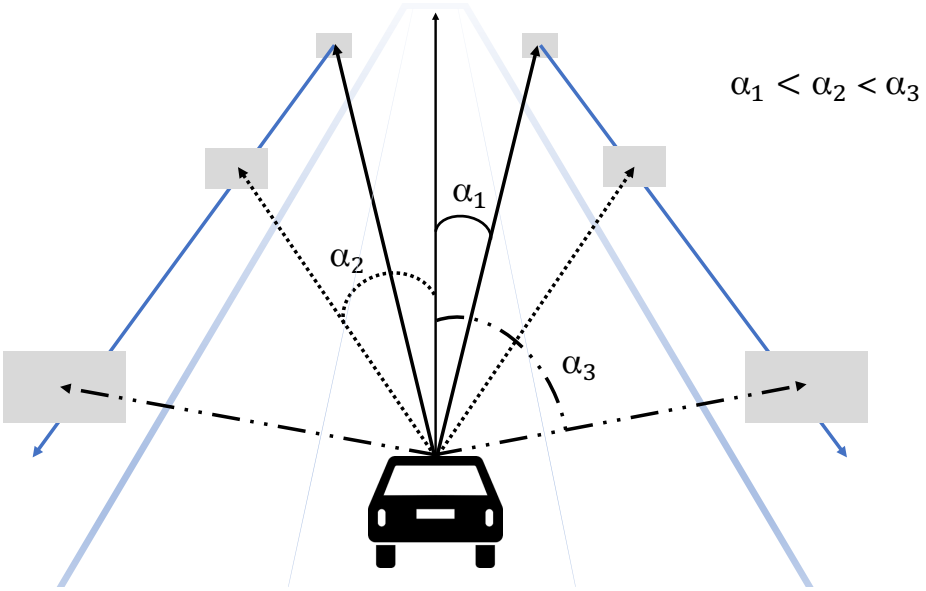


Figure 4.14: Experiment 1 - This schematically illustrates the LCT task from the drivers' perspective. The closer the LCT signs (rectangles) approach the further the drivers have to direct their gaze to the left and the right

In summary, besides the consistent evidence of spatial gaze concentration under cognitive workload in the literature (see 3.6.2), the results obtained in this experiment provide contradictory evidence. Although these results are probably highly influenced by the specific nature of the driving task, they emphasize that similar to fixation duration and count, gaze dispersion measures also have to be interpreted in relation to task dynamics. However, final conclusions, whether to integrate fixation dispersion into a multi-measure approach of cognitive driver workload cannot yet be drawn based only on the results of this experiment. Firstly, horizontal gaze dispersion exhibited sensitivity to different cognitive workload levels, even though not as expected regarding the direction of change in this experiment. Secondly, results from past research allow for the assumption that horizontal gaze dispersion might be valuable in more naturalistic driving scenarios. This, however, has to be tested in a follow-up experiment.

Pupil size

The results on changes in pupil size fully confirm the hypothesis. Pupil sizes increased incrementally with increasing task difficulty and discriminated between all three task difficulty conditions. This is in line with past research corroborating that pupil size can be used as a reliable and sensitive measure of cognitive workload (see

subsection 3.6.2). In contrast to indicators like heart rate and nose tip temperature, the changes in pupil size seem to fully depict sympathetic and parasympathetic changes related to cognitive workload. Whereas heart rate and nose tip temperature are predominantly subject to sympathetic changes, the modulation of pupil size is to a greater degree influenced by parasympathetic changes (Steinhauer et al., 2004). For this reason changes in pupil size also appear to be more sensitive to the different task demands and respectively cognitive workload levels in the present experiment than cardiovascular measures. From a technical perspective, continuous and reliable pupil detection can be achieved through contactless eye-tracking or even simple video-based techniques. Additionally, advances in signal processing allow the distinction of luminance-induced and workload-induced changes (see e.g. ICA - Index of Cognitive Activity Marshall (2002)). Taking the sensitivity to changes of cognitive workload and the technological possibilities into account, pupil size represents a promising indicator of real world multi-measure approaches when confounding influences like e.g. luminance changes can be controlled.

Speech signal features

So far there is little research on fundamental frequency of the voice and voice intensity in the context of drivers' cognitive workload. It was expected that both measures increase with increasing task difficulty conditions. This assumption can be confirmed by the results of this experiment. F0 as well as voice intensity exhibited higher absolute values in the higher task difficulty conditions with larger effect sizes for the voice intensity measure. The sensitivity, however, seems to be limited similar to that of heart rate and nose tip temperature. In contrast to all other physiological measures, no driving-only baseline is available for the speech signal features, so that inter-individual differences in initial frequencies are not accounted for. This might limit the sensitivity of measures of the speech signal. Even though it is not easily possible to acquire "unloaded" reference values from the driving-only baseline sections, it might be of value to obtain speech-only baselines as individual references for future studies to further increase the sensitivity of the measure. In summary, the presented results regarding speech signal features emphasize the potential to be included in multi-measure approaches of cognitive workload. A further advantage of the measures from a technological point of view is that modern speech-based interaction systems already extract speech signal features like fundamental frequency for speech recognition. Therefore, these features can easily be integrated with already existing in-car technology.

Summary on performance, subjective and physiological measures

Table 4.6 summarizes the performance, subjective and physiological results of the first experiment. Overall, all dependent variables were sensitive to differences in cognitive task difficulty. Primary- as well as secondary-task measures revealed that the experimental manipulation led to three different levels of cognitive driver workload. Performance in the PASAT as well as overall workload and the NASA-TLX dimensions of mental and temporal demand clearly differentiated between all three task difficulties.

All physiological measures were sensitive to differences in cognitive driver workload. With respect to effect sizes of the main effects of task difficulty, the results indicate that heart rate and pupil size were particularly sensitive to the differences in cognitive driver workload. Although effect sizes for nose tip temperature as well as the speech signal features were lower, these features also proved to be promising indicators of cognitive driver workload. Measures of eye fixations also changed as a function of cognitive driver workload but have to be interpreted carefully with respect to their task dependency described earlier. Only pupil size changes exhibited differential sensitivity to all three levels of cognitive driver workload whereas the other physiological measures failed to differentiate between the medium and high workload condition. The results therefore indicate that the differential sensitivity of physiological measures is limited at higher levels of cognitive driver workload.

4.5 Limitations of the experiment

The presented experiment has several methodological limitations. First of all the choice and implementation of the secondary-task PASAT might have led to undesired effects on the physiological measures. Whereas in the low and medium task difficulty condition the inter-stimulus interval was set to five seconds, the inter-stimulus interval was three seconds in the highest task difficulty condition. For this reason, in the highest task difficulty condition, the number of responses invoked by the secondary-task was higher as compared to the other conditions, thus resulting in a higher number of mental calculations. This might have caused more task-evoked pupillary responses (Beatty, 1982), resulting in an overall higher average of the pupil size in the high difficulty condition. For this reason, it is possible that the sensitivity of pupil size to discriminate between the medium and the high task difficulty condition is the result of the difference in the number of reactions to the secondary-tasks and not the result of two different cognitive demands. This problem potentially expands to cardiovascular

Table 4.6: Experiment 1 - Statistical significance for main effect of task difficulty, effect sizes of the main effect and pairwise comparisons for performance, subjective and physiological measures

	Main effect of task difficulty	**Effect size** η_p^2	**low vs. medium**	**low vs. high**	**medium vs. high**
Primary task performance					
Driving					
Mean lane deviation*	sig.	.11	n.s.	sig.	n.s.
Secondary task performance					
PASAT					
Proportion of Errors	sig.	-	sig.	sig.	sig.
Subjective measures					
NASA-TLX					
Mental Demand	sig.	.79	sig.	sig.	sig.
Physical Demand	n.s.	.07	-	-	-
Temporal Demand	sig.	.61	sig.	sig.	sig.
Performance	sig.	.34	n.s.	sig.	sig.
Effort	sig.	.58	sig.	sig.	n.s.
Frustration	sig.	.46	sig.	sig.	n.s.
Overall workload	sig.	.71	sig.	sig.	sig.
Physiological measures					
Heart rate	sig.	.60	sig.	sig.	n.s.
Nose tip temperature	sig.	.29	sig.	sig.	n.s.
Fixation duration	sig.	.23	n.s.	sig.	n.s.
Fixation count	sig.	.10	n.s.	n.s.	n.s.
Horizontal fixation dispersion	sig.	.26	sig.	sig.	n.s.
Pupil size	sig.	.83	sig.	sig.	sig.
Voice fundamental frequency	sig.	.18	sig.	sig.	n.s.
Voice intensity	sig.	.28	sig.	sig.	n.s.

'sig.': significant (α-level of $p = .05$)
'n.s.': not significant
'*': α-level adjusted to $p = .2$
'-': not calculated

measures and speech signal features. Due to more reactions that involve speech in the high difficulty condition, a change in breathing patterns is likely, thus influences on heart rate as well as speech signal features cannot be ruled out. Future studies should address this problem by keeping the inter-stimulus interval constant between different task difficulty levels. This could either be accomplished by using a more difficult mental arithmetic task (e.g. multiplication) instead of an addition for the high task difficulty condition in the PASAT or by a different secondary-task with constant ISIs like the n-back (see e.g. Kirchner, 1958; Mehler, Reimer, & Dusek, 2011). Regarding the secondary-task it would also be useful to assess single-task performance in the PASAT to gain more detailed insights in the compensatory mechanisms involved in dual-task performance and to more specifically infer the attentional interference between the two tasks.

Secondly, the LCT might be a suboptimal choice for a driving task with regard to ecological validity. Even though the LCT demands the same attentional resources as general driving, the gaze patterns that it induces are highly unlikely to appear in real world driving. Whereas this most probably does not influence heart rate, nose tip temperature and speech signal features, it very likely impacts eye-movement related measures as discussed in subsection 4.4.3. Therefore, future studies should use more naturalistic driving tasks and environments.

Thirdly it is not precluded that simulator sickness did occur for some participants. Although they were instructed to report any symptoms of simulator sickness, it is possible that physiological reactions to simulator sickness as e.g. reported by Min, Chung, Min, and Sakamoto (2004) influenced the physiological reaction of those participants.

Fourthly, the comparatively high room temperatures (M=27.8 °C, SD=0.8 °C) might have caused increases in heart rate by regulatory sympathetic activation to counter higher blood volume levels. These may have been induced by a decrease in resistance of the vaines due to high temperature (Klinke & Bauer, 1996). However, no systematic effect of room temperature on the heart rate results is expected as the task difficulty conditions were counterbalanced.

Furthermore, measurement related issues have to be addressed. Regarding the eye-movement measures it is worth noting that under the assumption that the participants used smooth pursuit eye-movements to fixate the moving signs that indicate lane changes, it is possible that fixation duration, count and hence the fixation dispersion, are flawed. The fixation detection algorithm used in this experiment is based on local proximity which is the common approach with eye-tracking data sampled at

a frequency of 60Hz. However, if smooth pursuit eye-movements by the participants occurred and exceeded the maximum dispersion of 80px, then the algorithm would erroneously split up one fixation into several fixations. Further, low resolution of the thermography equipment should be mentioned as a limitation. Even though the thermal resolution of $< .08$ Kelvin is reasonably good, the optical resolution of the camera is low (160px x 120px), leading to different skin tissue coverage with ROIs of fixed size (5px x 5px) between subjects depending on the size of their nose and the distance to the thermography camera. Using a within-subjects design and baseline corrected temperature values should reduce these effects. However, the results of this experiment with regard to nose tip temperature should be validated with measurement systems of higher accuracy in further studies.

4.6 Conclusions

The secondary-task successfully induced three different levels of cognitive workload. All physiological measures that were used, were able to discriminate between different cognitive workload levels but mostly lacked sensitivity in higher workload regions. Pupil size proved to be most sensitive amongst the traditional measures. Speech signal features and nose tip temperature, which are relatively new measures of cognitive workload, revealed their potential to be used in systems assessing cognitive workload. Eye-movement measures revealed changes in visual attention allocation strategies in response to auditory-verbal secondary-tasks. Several limitations of the experiment were recognized and are addressed in the following experiment which is described in chapter 5.

Chapter 5

Experiment 2 - Validation of physiological measures

The previous experiment examined the sensitivity of multiple non-intrusive physiological measures to different levels of cognitive driver workload in a basic driving simulation. Suitable measures were identified and the results are now to be validated in a more realistic driving environment that addresses several limitations from the first experiment. Hypothesis regarding performance, subjective and physiological workload measures will be generated in section 5.1. The following methods section (5.2) will cover the experimental setup (subsection 5.2.1), sample (subsection 5.2.2), design (subsection 5.2.3), procedure (subsection 5.2.4) as well as data pre-processing and data analysis steps (subsections 5.2.5 and 5.2.6) of the second experiment. Subsequently the results on performance (subsection 5.3.1), subjective (subsection 5.3.2) and physiological measures (subsection 5.3.3) will be presented and discussed in section 5.4. This is followed by the description of the experimental limitations (section 5.5) and the conclusions that can be drawn from this experiment (section 5.6).

This experiment was supported by two psychology student interns at Technische Universität Braunschweig. Lena Lüneburg and Ronja Gerdes helped with the preparation and conduction of the experiment. The technical implementation and analysis was supported by Mario Lasch, employee of the Chair of Human-Machine Systems at Technische Universität Berlin and Doris Sonntag of the Institut für Psychologie Ingenieur- und Verkehrspsychologie at Technische Universität Braunschweig. The experiment presented in this chapter was approved by the ethics board of the Department of Psychology and Ergonomics at Technische Universität Berlin (see appendix B.1.1).

5.1 Introduction

The aim of the second experiment was to validate the results obtained in the first study in a more applied scenario and to evaluate which measures are most suitable for the discrimination between multiple cognitive workload levels. To overcome several methodological limitations from the first experiment (see section 4.5) and to increase ecologic validity of the results, the primary- and secondary-task as well as the driving environment were adapted according to the concerns discussed in section 4.4. Additional focus of the second experiment was to explicitly study the impact of different cognitive task demands on event-related driving as well as continuous driving performance measures. To differentiate between continuous and event-related driving performance, additional safety critical events were introduced. On the basis of section 3.4, section 3.5 and section 3.6, the following hypotheses regarding performance, subjective and physiological workload measures were generated:

Regarding the impact of task difficulty on performance the following is hypothesised:

- Continuous lateral vehicle control is expected to be unaffected by increasing task difficulty. Although results in the literature regarding changes of lateral vehicle control in response to increasing cognitive demands are mixed (see subsection 3.4.3), two meta-studies (Horrey & Wickens, 2006; Caird et al., 2008) as well as studies by e.g. Harbluk and Lalande (2005), Maciej and Vollrath (2009) and Yager (2013) support this hypothesis.

- In their meta-analysis Caird et al. (2008) report that drivers tend to reduce their speed under high cognitive demands. For speech-based interaction tasks, however, little to no changes in continuous longitudinal vehicle control were reported. Therefore only a small decrease in speed is expected with increasing task difficulty.

- As described in subsection 3.4.1, the literature consistently shows that the reactions to critical events are more severely affected by cognitive workload compared to lateral and longitudinal control. Reactions to safety critical events are therefore expected to diminish with increasing task difficulty.

- Increasing task difficulty is expected to lead to degraded performance in the secondary-task in order to protect performance in the driving task, as driving is instructed as primary-task.

Regarding the impact of task difficulty on subjective workload the following is hypothesised:

- Subjective workload ratings are expected to increase with increasing task difficulties induced by the secondary-task, as has been shown repeatedly (see section 3.5)

Regarding the impact of task difficulty on physiological measures the following is hypothesised:

- Consistent with the results of experiment 1 and the literature (e.g. Brookhuis et al., 1991; Engström et al., 2005; Lenneman et al., 2005; Liu & Lee, 2006; Collet et al., 2009; Lenneman & Backs, 2009; Mehler et al., 2009; Reimer & Mehler, 2011) heart rate is expected to increases with increasing task difficulty.
- Results of experiment 1 as well as from past research (see Or & Duffy, 2007; Yamakoshi et al., 2008; Itoh, 2009) indicate that heightened sympathetic activation leads to a drop in nose tip temperature. Therefore it is expected that nose tip temperature decreases with increasing cognitive task difficulty.
- Based on extensive evidence of decreased gaze dispersion from past research (e.g. Recarte & Nunes, 2000; Engström et al., 2005; Victor et al., 2005; Tsai et al., 2007; Horrey, 2011; Niezgoda et al., 2015), horizontal fixation dispersion is expected to decrease with higher cognitive secondary-task demands. Experiment 1 exhibited opposite results. However, it is believed that the observed increase in horizontal fixation dispersion resulted from the specific task characteristics of the LCT.
- Consistent with the results from experiment 1 and the literature (see e.g. Recarte & Nunes, 2000, 2003; Palinko et al., 2010; Niezgoda et al., 2015) pupil size is expected to increase with increasing task difficulty.
- Experiment 1 has provided evidence that speech signal features can be promising indicators of cognitive workload in the context of driving. Therefore, the fundamental frequency of the voice and voice intensity are expected to increase with increasing task difficulty.

5.2 Methods

5.2.1 Experimental setup

The study was conducted at Technische Universität Braunschweig - Institut für Psychologie Ingenieur- und Verkehrspsychologie under the direction of Prof. Dr. Mark Vollrath. For the experiment two rooms were available. One room accommodated the driving simulator. The second room was located next to the first one and served as monitoring station. From this room the examiner was able to monitor the participant, control the experimental procedure as well as monitor and control all experiment related measurement systems (figure 5.2). To monitor the simulator room, three cameras were installed (figure 5.3). The first was pointed at the front of the participant, the second one was installed in the foot well of the driving booth and the third one in the back of the subject. The communication between participant and experimenter was bilaterally possible at all times during the experiment via microphones and speakers respectively headphones.

Driving simulator setup

The driving simulator is shown in figure 5.1. Driver seat, steering wheel and pedals were mounted on a custom made construction 10cm above floor level. Seat as well as the steering wheel were adaptable within limits to the anthropometric characteristics of the participants to allow for a comfortable seating position. The simulation was displayed on three flatscreen monitors mounted on mono pods and covered almost 180 degrees of the subjects visual field in horizontal direction. The distance between the subjects and the middle screen, depended on the seat position and amounted to approximately 1m - 1.5m. The simulation software was SILAB 5.0, developed by Würzburger Institut für Verkehrswissenschaften (WIVW). The software is a highly realistic and scenario-based simulation that offers the possibility to integrate external hard- and software via several interfaces (WIVW, 2016). The synchronization with all other measurement systems was accomplished by using UDP (User Datagram Protocol) messages. The software constantly generated markers with information about the status of the simulation (i.e. participant number, experimental condition, driving section and sequence, event status) at 100Hz and broadcasted them to the receiving measurement systems. The receiving systems were programmed to detect changes in the UDP messages and attached the respective marker to the recording. This allowed for a simple and efficient synchronization of the data streams of the simulator and the physiological measurements with sufficient accuracy.

Figure 5.1: Experiment 2 - Driving simulator setup. The picture shows the three screens that displayed the driving environment and the position of the seat, driving wheel and pedals in relation to the screens.

Figure 5.2: Experiment 2 - Monitoring room. The monitoring room was located next to the driving simulator room and contained the equipment necessary to control the experiment as well as for the communication with the participant.

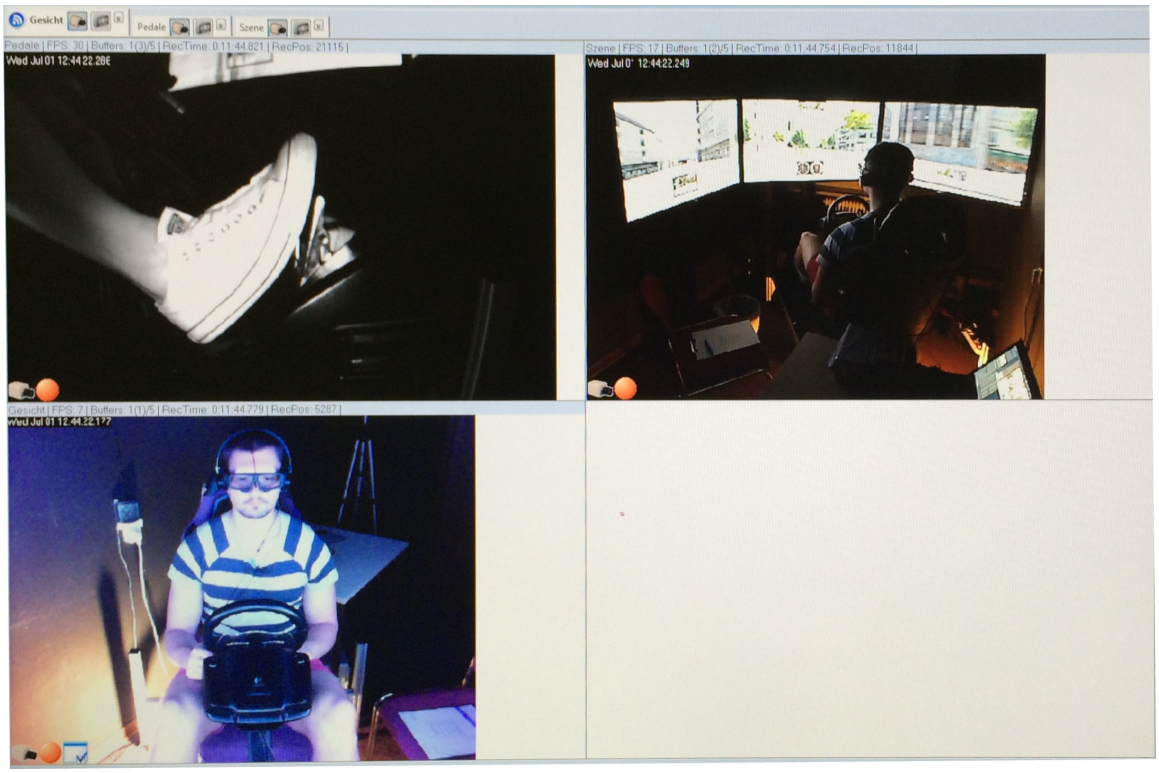
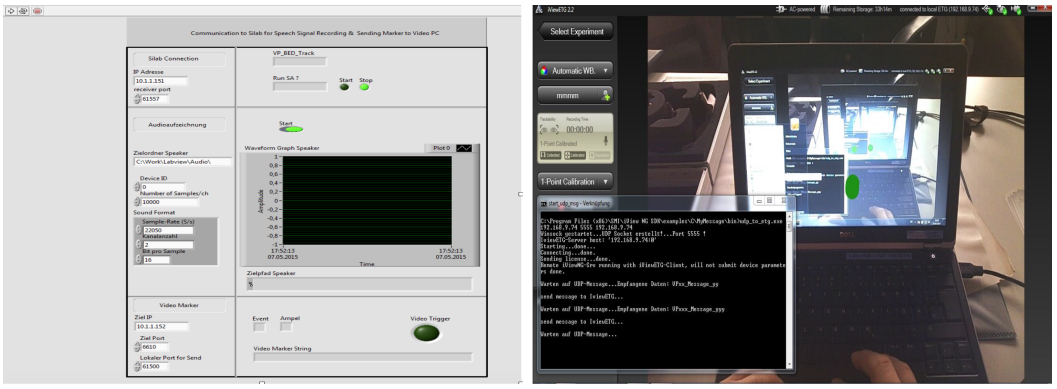


Figure 5.3: Experiment 2 - Camera positions. Upper left: foot well camera. Upper right: back camera. Lower left: front camera



(a) Graphical user-interface for audio, video and Varioport control

(b) Eye-tracking synchronization setup. The small window in the front establishes connection between the simulation software and the eye-tracking software in the back

Figure 5.4: Experiment 2 - Technical setup for synchronization

ECG and temperature measurement: For the acquisition of ECG and temperature data a Becker Meditec Varioport portable biosignal recorder was used. ECG and temperature were sampled at 512 Hz and saved at 256Hz. As the resolution of the thermography camera used in the first experiment was limited, this time temperature at the tip of the nose was collected with an attachable temperature sensor with an accuracy of 0.1°C (in the range of 20-37°C) and a temperature resolution of 0.003°C. Furthermore, the Varioport provided an external marker input channel to synchronize with the experimental setup.

Eye-movement measurement: Eye-movements were recorded binocularily using the SMI Eye-Tracking Glasses 2 (ETG 2) sampling at a frequency of 60Hz. The glasses were connected to a laptop that allowed for experimental manipulations, synchronization via the SDK (Software Development Kit) as well as the local storage of the eye-tracking data files.

Speech signal measurement: The speech signal was recorded by wireless Sennheiser ew 100 g2 microphones at 22050Hz mono. The resulting files were stored on a PC. A Labview program (see figure 5.4a) was developed to receive the UDP packages from SILAB and to trigger and mark audio and video recordings as well as send triggers to the Varioport. The markers in the eye-tracking data were generated by a C++ program that interpreted the SILAB UDP messages (see figure 5.4).

Driving scenario

The driving task was embedded in a lively urban environment including oncoming and preceding traffic as well pedestrians and cyclists. With the exception of intended critical events, the scenario was programmed so that aforementioned road users did not interfere with the participants driving.

Driving task

The driving course consisted of four blocks (see figure 5.5). Each block was divided into two different sections. The first section of each block represented the driving-only baseline section where no secondary-task was present. The second section represented the driving section where, in addition to the driving, an auditory-verbal secondary-task was presented. Even though both sections were slightly different, the demand originating from the driving task were comparable in both sections as it mostly consisted of straight driving without turns or overtake scenarios. At a speed of 50 km/h the completion of each section lasted approximately four minutes.

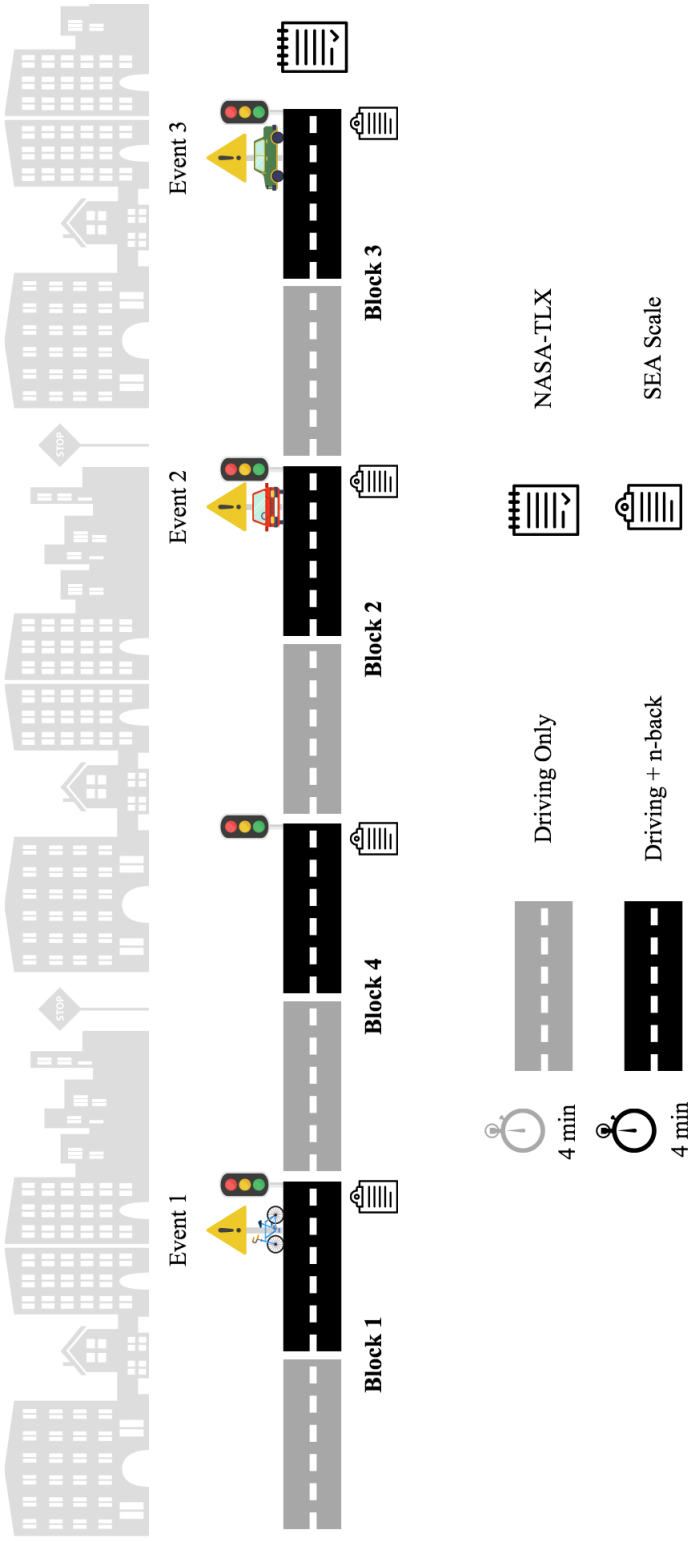


Figure 5.5: Experiment 2 - An example of the driving task sequence. Note that the order in which critical events occured was randomized between participants

Critical events

In three out of four blocks (block 1-3), a critical event occurred in the last minute of the driving with secondary-task section but not at the exact same time. For the purpose of reducing predictability, block 4 comprised no event. The events were positioned towards the end of the subsection to ensure that the initial demand related adaptation phase of the slowest physiological indicator had passed. In this case the slowest parameter was the temperature at the tip of the nose. In accordance with Itoh (2009) and the first experiment it is assumed that this phase lasts approximately 2 minutes on average.

The order of the blocks 1-3 was fully counterbalanced between participants, while the fourth block always came second. Each participant completed all four blocks, was instructed to follow the road and to comply with the rules of the German road traffic act (StVO) for inner-city driving.

1. **Event - Bicycle:**
   The bicycle event is visualized in figure 5.6. The occurance of the bicycle is triggered by the ego vehicle, 30 m away from the theoretical intersection of both vehicles. From this point on, the bicycle moves at a speed of 6.5 m/s and the distance to the intersection is 8.4 m.

Figure 5.6: Experiment 2 - Bicycle event: Bicycle appearing from the right side

2. **Event - Oncoming car:**
   The oncoming-car event is visualized in figure 5.7. The speed and velocity of the oncoming vehicle is coupled to the speed of the ego vehicle. The left turn

is triggered by the ego car at a distance of 50 m to the theoretical intersection. The oncoming car then reduces its speed to 7 m/s and turns left.

Figure 5.7: Experiment 2 - Oncoming-car event: Oncoming car turning left

3. **Event - Parking car:**
   The parking-car event is visualized in figure 5.7. It is triggered at a distance of 47 m between the position of the ego vehicle and the theoretical point of intersection between both vehicle trajectories. At this point the car that is coming out of the parking lot is 6 m away from the intersection point and accelerates at a rate of 3 m/s^2 up to 14 km/h

Figure 5.8: Experiment 2 - Parking-car event: Parking car turns on the street from the right

Secondary-Task

For the second experiment the author decided to replace the PASAT as secondary-task due to the limitations described in subsection 4.5. Instead, a n-back task was used which has broadly been applied as working memory task and as a measure of individual differences (Jaeggi, Buschkuehl, Perrig, & Meier, 2010). *"In this test, subjects are presented with a series of stimuli (e.g. spatial locations, visual objects, auditory objects) and required to decide whether the current stimulus is the same as the stimulus seen n trials back."*(Cools, 2010, p.822). The n-back test comes with the advantage that the difference between two task difficulties is not confounded by the different number of tasks within the same time interval as it is with the PASAT. In its original form *"[t]he task involves multiple processes, such as the encoding of the incoming stimuli, the monitoring, maintenance, and updating of the material, as well as matching the current stimulus to the one that occurred n positions back in the sequence. Decision, selection, inhibition, and interference resolution processes are also involved."* (Jaeggi et al., 2010, p.395). Initially it was used as a visio-spatial matching/non-matching task by Kirchner (1958) to compare age related differences in short-term retention. The n-back task is also often used in neuro-imaging studies and has shown to consistently activate the same neural areas in the brain as well as stimulus specific areas related to the modality of presentation and response (for meta-analysis see Owen, McMillan, Laird, & Bullmore, 2005). In recent years it has also proven useful as auditory-verbal secondary-task to induce different cognitive workload levels in driving studies (e.g. Mehler et al., 2009; Lenneman & Backs, 2009; Mehler, Reimer, & Coughlin, 2012).

For this experiment a slightly modified version of MIT's auditory-verbal delayed response task (n-back) was implemented (Mehler et al., 2011). Participants were presented a sequence of 90 digits and their task was to repeat the digit n steps before the digit that was presented last. Digits 1-9 were generated by a natural voice synthesizing software (Linguatech, 2016) and then randomly combined with an inter-stimulus interval of 2.25s. Within the digit sequences 'lures' were avoided, as they have shown to result in cognitive interference which add additional demands to the task (e.g. Gray, Chabris, & Braver, 2003; Kane, Conway, Miura, & Colflesh, 2007; Szmalec, Verbruggen, Vandierendonck, & Kemps, 2011). Lure trials occur, when a digit that is no longer or not yet relevant matches the currently presented digit, e.g. **1**-3-2-1 or 3-2-**1**-1 in a 2-back task (Szmalec et al., 2011). That is why digits that matched the n+1 or n-1 positions were excluded from the sequence. For

each driving with secondary-task section a different continuous sequence of digits was generated, hence totalling to four sequences. At the beginning and the end of the number sequence, a message stating the beginning and end of the task was included.

5.2.2 Participants

Overall 100 subjects participated in the study. As only complete datasets were used, 10 participants had to be excluded due to a loss of data or discontinuations of the experiment due to simulator sickness. This resulted in 90 complete data sets. Within this group the mean age was $M = 22.71$ years (range 19-30 years) with a standard deviation of $SD = 2.62$ years. Gender distribution was almost balanced (female: 44, male: 46). The majority of the participants were students enrolled at Technische Universität Braunschweig with diverse study backgrounds. Additionally one research associate and one person who was unemployed at the time of the experiment took part. All participants spoke fluent German. They came from five different nationalities (German: 86, Bosnian: 1, Chinese: 1, Romanian: 1, Czech: 1). Pre-condition for participation was the possession of a valid driver's licence for at least two years.

5.2.3 Experimental design

The study featured a one-factorial between-subjects design. The factor *'task difficulty'* was defined as the task difficulty of the n-back and consisted of three levels. In the easiest condition, subsequently called 'low' condition, participants had to verbally repeat the last presented digit. In the 'medium' and 'high' condition they were asked to continuously verbalize the digit that was presented $n = 1$, respectively $n = 2$ positions back in the sequence. Task difficulty levels are here described as low, medium and high for the purpose of easier differentiation for the reader in the following sections. It is however important, that these labels do not describe absolute task difficulties but only relative differences between the three levels.

Dependent variables

Dependent variables are grouped in three different main categories commonly agreed upon in mental workload research (O'Donnell & Eggemeier, 1986) and operationalized as follows:

Performance measures

This category includes primary-task measures of driving as well as secondary-task measures from the n-back task.

1. Continuous longitudinal and lateral driving performance:
   (a) Mean and standard deviation of the participants lateral lane deviation in m calculated as the distance to the middle of the drivers' driving track
   (b) Mean and standard deviation of steering acceleration in rad/s
   (c) Mean and standard deviation of speed in km/h
2. Event-related driving performance:
   (a) Number of crashes and near crashes
   (b) Number of omitted brake reactions as reaction to the critical events
   (c) Brake reaction times in ms. Reaction time was calculated from the moment the critical incident could theoretically be identified by the participant.
3. Secondary-task performance:
   (a) n-back performance is measured as the proportion of correct responses to the overall number of possible responses

Subjective Measures

This category includes two questionnaires used to measure the participants' perceived workload.

1. SEA - Scale (Eilers, Nachreiner, & Hänecke, 1986):
   (a) Perceived effort on a scale from 0 to 220
2. NASA-TLX (Hart & Staveland, 1988; Seifert, 2002):
   (a) Raw values for each NASA-TLX dimension on a scale from 0-100
   (b) Individually weighted overall score on a scale from 0-100

Physiological measures

1. Cardiovascular Measures:
   (a) Heart rate in *bpm*
   (b) Temperature at the tip of the nose in °C
2. Ocular Measures:
   (a) Horizontal fixation dispersion as the standard deviation of fixation positions in x direction
   (b) Pupil size as pupil diameter in *mm*
3. Speech signal features:
   (a) Fundamental frequency F0 in Hz
   (b) Voice intensity in voiced speech segments in dB

5.2.4 Procedure

The experimental procedure was the following. Overall the experiment was conducted by three different experimenters. Beside the author of the thesis two psychology student interns from TU Braunschweig, Lena Lüneburg and Ronja Gerdes, helped carrying out the experiment. All experimenters had the same training on the experimental procedure. First of all, the participants were greeted and thanked by the experimenter and asked to switch off their mobile phones. Subsequently, the participants were informed orally and in writing about the course and terms of the experiment (see appendix B.1.2). After that they signed a general consent form (see appendix B.1.6) as well as a consent form on audio and video recordings (see appendix B.1.7). Baseline speech was recorded by asking the participants to count slowly from one to nine. As a next step, a demographic questionnaire including questions on their well-being was handed out (see appendix B.1.8). The participants were instructed and trained on their respective n-back task condition (see example for 2-back in appendix B.1.4) until at least 50 percent of the answers were correct and the participants felt comfortable with the task. As long as one of the two conditions was not fulfilled participants performed another set of ten digits until both conditions were met (see training document in appendix B.1.5). Following this, the driving simulator was introduced to the participants. They were asked to find a comfortable position and instructed to

explore steering and braking characteristics while driving a section which was comparable to the later experimental task. They were also informed that after a period of approximately 5 minutes of solely driving, the n-back task would be started and should be performed as good as possible without disregarding the driving task. After that, the physiological sensors were applied to the participant. The participants were instructed on how to attach the ECG electrodes (see figure 4.8 for placement information) and then did it themselves in a separate room. After their return the quality of the ECG signal was checked by the experimenter. The participants were seated in the simulator and the temperature sensor was applied to the tip of the nose. They put on the eye-tracking glasses and the microphone was attached to their collar. Then the eye-tracker was calibrated to the participant with a 3-point calibration on the center screen. The experimenter handed out the instructions on the experimental tasks (see in appendix and B.1.3). Participants were instructed to focus on the driving task and perform the n-back task as good as possible without disregarding the driving task. They were also informed that an additional amount of 4 Euro could be earned if their overall performance reaches a given threshold. No information was provided on the critical events occurring during driving with secondary-task sections. After that the experiment started. Participants started driving on their assigned first driving-only section. Each driving-only section was followed by a driving with n-back section. At the end of each block, the participants stopped at a red traffic light and filled out the SEA Scale (Eilers et al., 1986). After completing all blocks they were asked to rate their subjective mental workload on the NASA-TLX scales (Hart & Staveland, 1988). Both questionnaires can be found in appendix B.1.9, respectively appendix A.1.1. All measurement sensors were detached by either the experimenter or the participants themselves. In the end participants were thanked by the experimenter and rewarded with 20 Euro, independent of their performance.

5.2.5 Data pre-processing

Data analysis as well as pre-processing of the data was mostly done with the free statistics software R (R Core Team, 2015a). All R packages that were used are documented in appendix B.2.3. Before the actual data analysis pre-processing had to be performend for all physiological measures and the driving data. Data was pre-processed as follows:

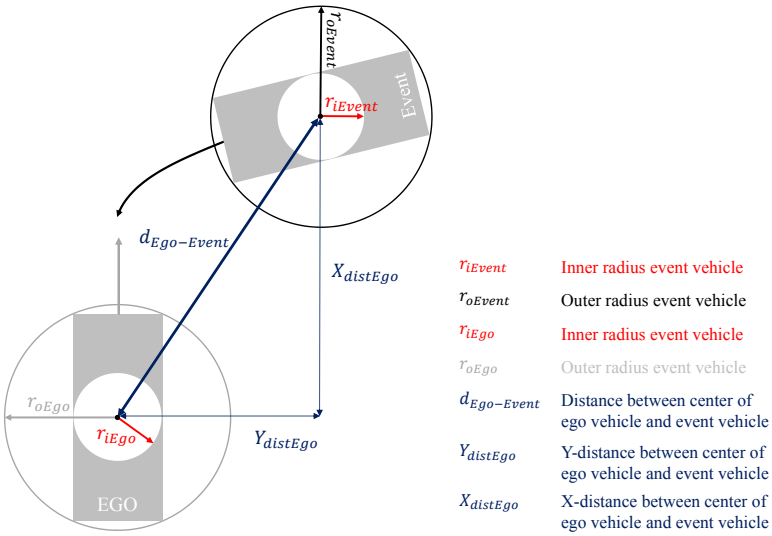
Driving Data

Continuous lateral and longitudinal measures were averaged over the driving sections for each block. For the driving-only baseline sections of each block, all data from start until the end of the section was used. For driving with additional secondary-task sections, the average was calculated from the beginning of the section until the start of the critical event. Event-related performance was calculated for each block using the driving data from the sections of driving with additional secondary-task, starting when the ego vehicle triggered the event vehicle and ending 10 m behind the theoretical point of intersection of both vehicles (see subsection 5.2.1 for details).

Collisions

Unfortunately, the driving simulation software did not calculate whether collisions occurred in the critical event sections or not. For this reason a two-fold approach was used to determine crashes post-hoc. Firstly, all eye-tracking videos were visually inspected and categorized into three different event outcomes. Events were labeled as 'no crash', when no contact between the ego- and the event vehicle was detected. The label 'crash' was applied when a contact was detected with certainty. All other cases were labeled as 'near crash'. In situations where the decision for one group was ambiguous, an additional comment describing the observation was added to the label. These comments later served as additional factor in the final decision on the outcome of the critical events which will be described later in this subsection. Secondly, the classification was determined by a mathematical approximation. Figure 5.9 visualizes the calculation of the events. The inner and outer radii of the ego and the event vehicle were determined for all three events. The inner radius constitutes a circle around the center of the vehicle to the closest side of the vehicle (left or right side). The second circle was drawn around the center of the vehicle with a radius that depicts the largest distance between the center of the vehicle and the farthermost corner of the car. In principal, the label 'no crash' was applied, when at no time during an event the two outer circles overlapped. The label 'crash' was applied, when at any point in time during the event, the inner circles overlapped. All other cases were labeled as 'near crash'. For the parking car event all data points, where both vehicles drove side-by-side were excluded. The calculation for the bicycle event differed slightly from the approach for the other two events because the trajectory of ego and event vehicle (see figure 5.6) allowed for a more accurate calculation of possible accidents. Here it was calculated whether the x-distance between the centers of the two vehicles was smaller

Figure 5.9: Experiment 2 - Geometrical relation between the ego vehicle and the event vehicle for the parking car event. However, the variables describing the relations can be transferred to all critical events.



than the sum of the distance between the center and the front of the ego vehicle and the distance between the left side and the center of the event vehicle when at the same time the y-distance between the vehicle centers was smaller than the distance of ego vehicle center and left/right side plus the distance between the ego vehicle center and the front of the event vehicle. When the x-distance was smaller, the event was labeled as 'crash'. When the outer circles of the vehicles overlapped, it was labeled as 'near crash'. All other cases were labeled as 'no crash'.

Finally, the results based on the visual inspection and those based in the calculation were compared and final decisions were made according to the following rules.

No Crash

1. when 'visually = near crash' and 'mathematical = no crash'
2. when 'visually = no crash' and 'mathematical = near crash' and no additional comment

Near Crash

1. when 'visually = no crash' and 'mathematical = near crash' when additional comment indicated an uncertain visual decision

Crash

1. when 'visually = near crash' and 'mathematical = crash'
2. when 'visually = crash' and 'mathematical = near crash'

ECG and nose tip temperature

The Varioport raw data files, containing ECG as well as facial temperature data, were converted to .txt files and afterwards segmented into the different sections of each block. Further steps regarding the ECG data followed the procedure described in subsection 4.2.5. No further pre-processing was necessary for the temperature data.

Eye-tracking

The eye-tracking and marker files were loaded to the gaze analyzing software SMI BeGaze version 3.5.101 and converted to text files containing all fixations including start and end time, duration, x and y position in the video file as well as the mean pupil size during the fixation. The default fixation detection algorithm used by the software is location based, summarizing all gaze samples within an area of two degrees of visual angle to one fixation. The minimal fixation duration was set to 80ms. Subsequently the resulting files were loaded into R. By averaging the pupil sizes over the length of each fixation, the procedure of excluding the eye blinks like in experiment 1 (see subsection 4.2.5) was obsolete. Data points containing invalid pupil sizes bigger 9 mm and smaller 1 mm were excluded (Loewenfeld & Lowenstein, 1993; Laeng et al., 2012; Sirois & Brisson, 2014). Fixation positions smaller than zero were filtered out, to avoid negative values for the variability calculations. Ultimately, the files were segmented into the relevant driving blocks and sections according to the UDP markers (see 5.2.1 for driving block and section definition).

Speech signal

Time series of the speech signal features were calculated using a customized Praat (version 6.018) script (see appendix B.2.2 for the full script). The parameters used for the estimation of F0 and intensity can be found in appendix B.2.1. The files were then segmented into the different sections for each block.

5.2.6 Data analysis

To compensate for individual differences, the physiological and driving performance data was averaged over the driving-only baseline sections and the corresponding driving with secondary-task sections for each block as described for continuous driving performance in the pre-processing section 5.2.5. Then the difference between those two was calculated so that the parameters can be interpreted and compared as differences between driving-only to driving with secondary-task. Speech signal features were analyzed by calculating the difference values of the dual-task segments with the recorded resting baseline. For final data analysis, block 4 was excluded as its only purpose was to reduce predictability of the critical events. Blocks 1-3 were averaged for each condition. Therefore between-subjects ANOVAs were calculated for all dependent variables. Shapiro-Wilks tests as well as QQ-plots were used to test for normality and the Levene test to check for homogeneity of variance. Post-hoc analysis of the main effects was performed, using the Tukey-Kramer multiple comparisons of means test which is robust against mildly unbalanced data (Hayter, 1984). All statistical analyses were conducted using R statistics version 3.0.1, ANOVAs were calculated using the 'ez' package (Lawrence, 2015) and post-hoc comparisons using the 'stats' package (R Core Team, 2015b). Exceptions of this procedure are explicitly mentioned. Additionally, time series analyses of secondary-task performance and all physiological measure were performed to gain further insights into other attributes of workload than average workload (e.g. instantaneous workload, peak workload and accumulated workload) as proposed by Xie and Salvendy (2000b)

5.3 Results

5.3.1 Performance measures

Driving performance - Continuous driving measures

As no changes were expected for lateral control measures, the significance level was adjusted to $p = .2$ to reduce the probability of committing a β -error (Bortz, 2005). The results of continuous driving measures for each task difficulty can be found in table 5.1. Mean values and standard deviations of the lateral deviation showed a trend to improved lane-keeping with increasing task difficulty, resulting in a significant main effect of task difficulty for the standard deviation of lateral deviation ($F(2, 87) = 13.86, p < .001, \eta_p^2 = .24$) and for the mean lateral deviation

Table 5.1: Experiment 2 - Means (M) and standard deviations (SD) of driving performance measures presented separately for the three task difficulties levels

	low		**medium**		**high**	
	M	*SD*	*M*	*SD*	*M*	*SD*
M lateral deviation in m	0.0215	0.0477	0.0049	0.0452	0.0033	0.0468
SD lateral deviation in m	0.0094	0.0259	-0.007	0.0223	-0.0098	0.0189
M steering acceleration in rad/s	0.31	0.18	0.46	0.18	0.55	0.47
M speed in km/h	-0.55	1.46	-0.62	1.74	-0.51	1.98

($F(2, 87) = 2.83, p = .064, \eta_p^2 = .06$). Even though the hypothesis could not be confirmed, post-hoc tests were calculated to gain further insights into the relationship between task difficulty and lane-keeping. Pairwise comparisons of the lateral standard deviation revealed that this measure discriminated between all conditions except between the 'medium' and the 'high' condition ($p_{low-medium} < 0.001$, $p_{low-high} < 0.001$ and $p_{medium-high} = .59$). The same pattern was found for the mean lateral deviation ($p_{low-medium} < .05$, $p_{low-high} < .05$ and $p_{medium-high} = .97$). The differences between the task difficulty conditions for both measures, however, varied only in the range of centimeters. Steering acceleration increased with increasing task difficultiy resulting in a main effect of task difficulty ($F(2, 87) = 5.09, p < .01, \eta_p^2 = .10$). Post-hoc comparisons showed a similar pattern for the standard deviation of the lateral deviation measure ($p_{low-medium} < 0.01$, $p_{low-high} < 0.001$ and $p_{medium-high} = .12$). No significant changes with respect to the driving speed were found ($F(2, 87) = 0.04, p = .959, \eta_p^2 < .01$) but overall, there was a general trend towards a slight reduction in speed in the sections with secondary-task compared to the driving-only baseline sections. However, the reduction was small for all three task difficulty conditions ($\sim 0.5 km/h$).

Driving performance - Event-related driving measures

Brake reaction times

All reaction times smaller than 150ms were removed to exclude brake reactions not associated with the critical events. Together with incidences where participants did not break at all, this resulted in 22 missing brake reaction times in the data set (Bicycle: 14 missing, Oncoming car: 5 missing, Parking car : 3 missing). All participants with at least one missing value were excluded from the data set resulting in 75 samples in the 'low', 81 in the 'medium' and 63 in the 'high' condition. The detailed distribution of the number of complete cases for each event and task difficulty condition can be found in table 5.2. Because of large differences in mean brake reaction

Table 5.2: Experiment 2 - Number of valid reaction times per event for each task difficulty

	Bicycle	Oncoming-car	Parking-car
low	25	25	25
medium	27	27	27
high	21	21	21

Table 5.3: Experiment 2: Dunnet-Tukey-Kramer test results for mean adjusted and log-transformed RT pairwise comparisons. Confidence Interval (CI) of 95%. Note: if the range from lower CI to upper CI of the mean difference between the conditions is above or below zero, the difference between the conditions is significant on a $p = .05$ level (Lau, 2009)

	Mean Difference	Lower CI	Upper CI
medium vs. low	-0.088	-0.153	-0.023
high vs. low	-0.007	-0.078	0.063
high vs. medium	0.082	0.020	0.143

times between the three events, the mean values of the reaction times were adjusted by calculating the difference between the mean RT of the event with the longest RTs (Oncoming-car: $M = 1273ms$) to the other events (Bicycle: $M = 445ms$, Parking-car event: $M = 651ms$) and applying it as offset to the other events' values. Subsequently, all brake reaction times were log-transformed to reduce violations of normality and homogeneity of variance assumptions (see appendix B.3.1 for details of the distribution). Other transformations, e.g. box-cox, 1/sqrt, etc., were tested but led to inferior results compared to the log-transformation.

Figure 5.10 shows that reaction times in the 'medium' condition were smallest compared to the 'low' and the 'high' condition. These differences resulted in a significant main effect of task difficulty ($F(2, 70) = 4.83, p < .05, \eta_p^2 = .12$). Post-hoc comparisons of the three task difficulty levels were calculated using an adjusted Dunnett-Tukey-Kramer test which is robust to unequal sample sizes and unequal variances between groups (Lau, 2009). The test revealed that the brake reaction time in the 'medium' condition was significantly shorter compared to the other two conditions (see table 5.3).

Collisions

Table 5.4 shows the number of crashes, near crashes and no crashes for all three task difficulty conditions. Overall the number of crashes was low in all three task

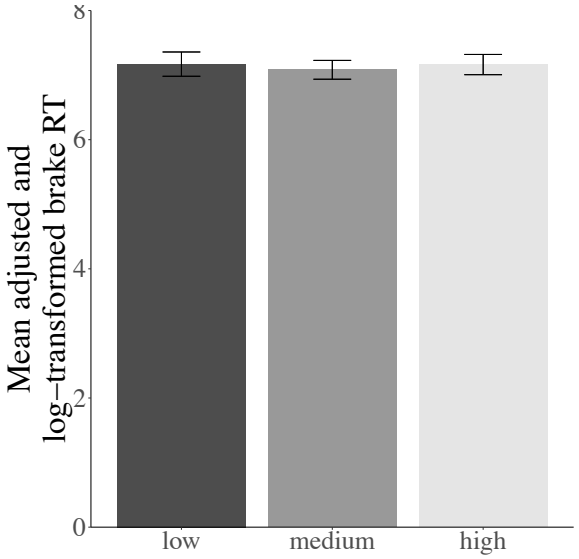


Figure 5.10: Experiment 2 - Mean adjusted and log transformed brake reaction times for all task difficulties. Error bars represent the standard deviations.

Table 5.4: Experiment 2: Number of no crashs, near crashs and crashs for each task difficulty

	low	medium	high
no crash	71	75	72
near crash	12	12	12
crash	7	3	6

difficulty conditions with the smallest number of crashes in the medium condition. The number of near crashes was equal amongst all task difficulties. However, a Friedmann test showed no significant difference in the distribution of crashes, near crashes and no crashes between the three task difficulties ($\chi^2(4, N = 270) = 1.74, p = 0.78$).

Omitted brake reactions

As a third measure of event-related driving performance, the number of omitted brake reactions was compared between the three task difficulties. Table 5.5 reveals that the number of omitted brake reactions was highest for the 'high' condition and lowest for the 'medium' condition. This resulted in a significant difference between the three task difficulties calculated by a χ^2-test, $\chi^2(2, N = 270) = 7.5, p = 0.02$. However, a post-hoc comparison using a Fisher test with Bonferroni adjusted p-values showed no significant pairwise differences, even though the difference between

Table 5.5: Experiment 2: Number of omitted brake reactions for each task difficulty

	low	medium	high
no brake	5	2	11
brake	85	88	79

the 'medium' and the 'high' condition was close to significant ($p_{low-medium} = 1.00$, $p_{low-high} = .57$ and $p_{medium-high} = .05$).

N-back performance

N-back performance decreased as a function of task difficulty ($M_{low} = 0.0003, SD_{low} = 0.002, M_{medium} = 0.03, SD_{medium} = 0.05$ and $M_{high} = 0.19, SD_{high} = 0.16$). A Kruskal-Wallis rank sum test revealed a significant main effect of task difficulty on n-back performance, $H(2) = 70.952, p < .001$. Pairwise comparisons, using the Dunn-test with Bonferroni adjusted p-values, resulted in significant differences between all three task difficulty conditions ($p_{low-medium} : z = -4.26, p < .001$, $p_{low-high} : z = -8.42, p < .001$ and $p_{medium-high} : z = -4.17, p < .001$). Additionally, a visual time series analysis of the n-back performance for all three task difficulty conditions showed that participants in the 'low' and the 'medium' condition were able to keep up performance over time, whereas performance in the 'high' condition declined rapidly within the first 50 seconds of the task and stabilized for the remaining time (see N-back section in appendix B.3.2).

5.3.2 Subjective measures

SEA

Participants scored their subjective effort on the SEA scale directly after experiencing the critical events in each block. To assess whether the different driving blocks differed regarding the subjective effort, the factor 'block' was included in the ANOVA for statistical testing. In general, the SEA ratings increased as a function of task difficulty, showing lowest ratings for the 'low' condition and highest for the 'high' condition (see figure 5.11). This resulted in a significant main effect of task difficulty, $F(2, 87) = 22.28, p < .001, \eta_p^2 = .34$. Post-hoc analyses of this effect showed significant differences between the 'low' and the 'medium' as well as between the 'low' and the 'high' condition but failed to differentiate between the 'medium' and the 'high' condition ($p_{low-medium} < .001$, $p_{low-high} < .001$ and $p_{medium-high} = .20$). No

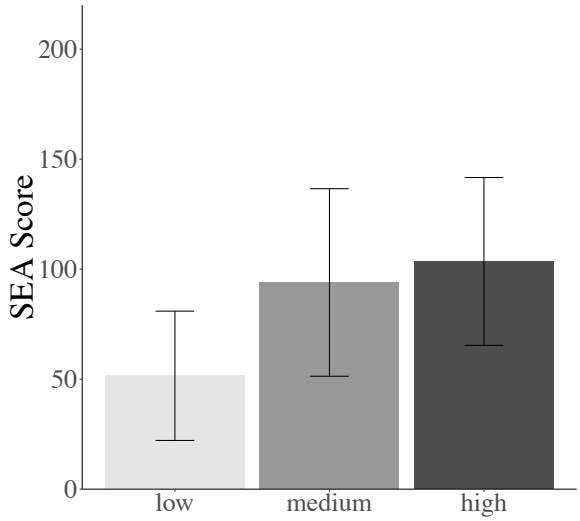


Figure 5.11: Experiment 2 - Mean values of the SEA ratings for each task difficulty. Error bars represent the standard deviations.

significant effects were found for the factor 'block' and the interaction between 'block' and 'task difficulty'.

NASA-TLX

In contrast to the SEA ratings, participants completed the NASA-TLX only once after their respective last block. All participants with missing ratings or missing weights for the pairwise ratings of the single NASA-TLX dimensions were excluded from statistical analysis ($N_{low} = 27, N_{medium} = 26, N_{high} = 30$). The weighted overall workload ratings increased as a function of task difficulty, showing lowest ratings for the 'low' condition and highest for the 'high' condition ($M_{low} = 30.26, SD_{low} = 10.98$, $M_{medium} = 51.16, SD_{medium} = 13.92$ and $M_{high} = 57.63, SD_{high} = 13.82$). The results of the ANOVA revealed a highly significant effect of task difficulty on overall workload, $F(2, 80) = 33.75, p < .001, \eta_p^2 = .46$. A post-hoc analysis of this effect showed significant differences between the 'low' and the 'medium' as well as the 'low' and the 'high' task difficulty condition but failed to differentiate between the 'medium' and the 'high' condition ($p_{low-medium} < .001$, $p_{low-high} < .001$ and $p_{medium-high} = .15$).

The individual dimensions of the NASA-TLX exhibited similar patterns to the overall workload ratings acquired by the NASA-TLX, displaying higher ratings regarding higher task difficulty except for physical demand which showed no significant changes regarding the three task difficulty conditions (figure 5.12). Detailed ANOVA results for all dimensions can be found in table 5.6. Post-hoc analyses revealed that

Table 5.6: Experiment 2 - ANOVA and post-hoc results for all NASA-TLX dimensions

	Main effect of task difficulty	low - medium	low-high	medium-high
Mental Demand	$F(2,80) = 31.01, p < .001, \eta_p^2 = .44$	$p < .001$	$p < .001$	$p < .05$
Physical Demand	$F(2,80) = 1.42, p = .25, \eta_p^2 = .03$	–	–	–
Temporal Demand	$F(2,80) = 7.77, p < .001, \eta_p^2 = .16$	$p < .001$	$p < .05$	$p = .23$
Performance	$F(2,80) = 4.74, p < .05, \eta_p^2 = .11$	$p = .1$	$p < .01$	$p = .68$
Effort	$F(2,80) = 50.96, p < .001, \eta_p^2 = .56$	$p < .001$	$p < .001$	$p = .91$
Frustration	$F(2,80) = 6.51, p < .01, \eta_p^2 = .14$	$p = .07$	$p < .01$	$p = .45$

Table 5.7: Experiment 2 - Mean values (M) and standard deviations (SD) of the physiological measures, represented as the percentage of difference between driving with secondary-task and driving-only baseline sections for each task difficulty

	low		**medium**		**high**	
	M	*SD*	*M*	*SD*	*M*	*SD*
Heart rate	4.07	3.03	10.72	6.53	9.37	6.60
Nose tip temperature	0.26	0.76	-0.13	0.70	-0.20	0.90
Horizontal fixation dispersion	-1.79	27.44	-17.60	21.52	-21.76	40.54
Pupil size	2.68	3.25	6.67	4.18	11.15	6.44

only the dimension of mental demand differentiated between all three task demands whereas effort, frustration, performance and temporal demand only seemed to be sensitive for differences between the lowest and the two higher task difficulties. However, pairwise comparisons of performance and frustration ratings differentiated only marginally between the low and the medium task difficulty condition.

5.3.3 Physiological measures

Cardiovascular measures

Heart rate

One participant in the high task difficulty condition had to be excluded from data analysis because of noisy ECG recordings. This resulted in $N_{low} = 30, N_{medium} = 30$ and $N_{high} = 29$ data samples for analysis. Heart rate increased as a function of task difficulty compared to the driving-only baseline sections (see table 5.7 and figure 5.13a), resulting in a significant main effect of task difficulty, $F(2,86) = 15.47, p < .001, \eta_p^2 = .26$. Post-hoc analyses showed that heart rate only differentiated between the 'low' and the 'medium' respectively the 'high' condition ($p_{low-medium} < .001$ and $p_{low-high} < .001$) but not between the 'medium' and the 'high' condition

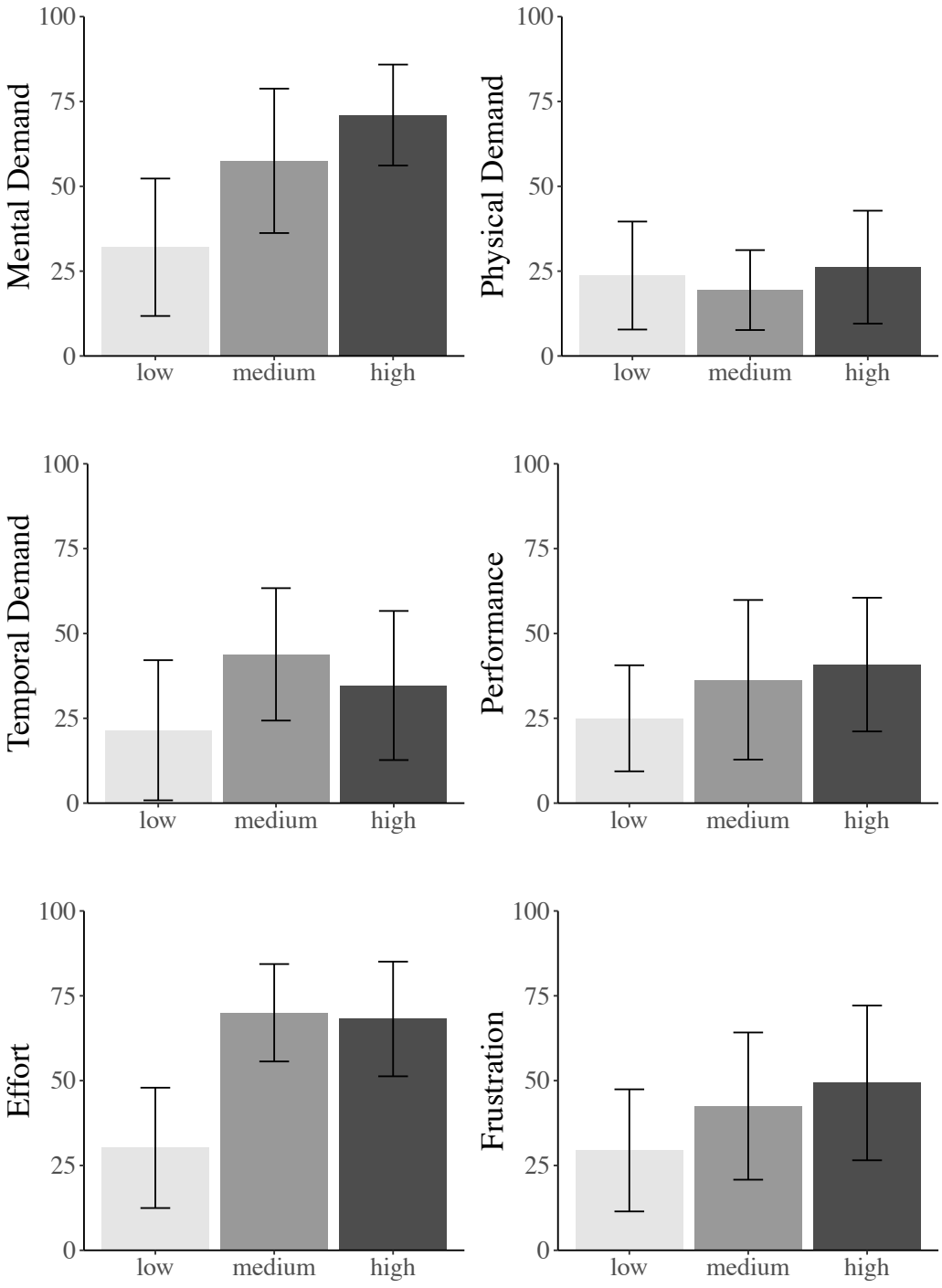
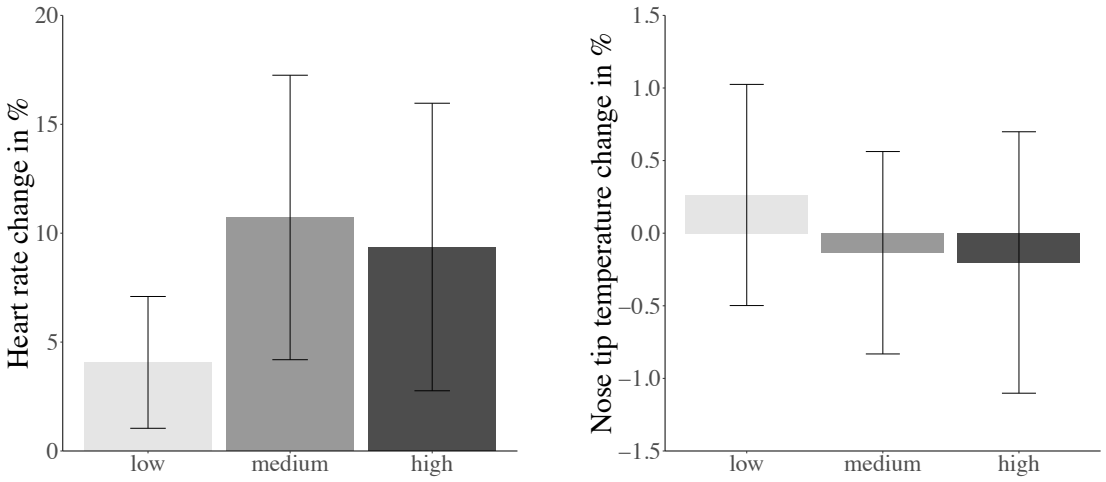


Figure 5.12: Experiment 2 - Mean values and standard deviations of the six NASA-TLX dimensions

($p_{medium-high} = .25$). Additionally, a visual time series analysis (see heart rate section of appendix B.3.2) of heart rate for all three task difficulty conditions revealed that participants in the 'medium' and the 'high' condition exhibited a similar increase within the first minute, followed by a short period of decrease. However, after that period the heart rate of participants in the 'high' condition dropped below the values of the 'medium' condition and continued to decrease whereas heart rate in the 'medium' condition remained relatively stable. Participants in the 'low' condition exhibited a short increase in HR in the beginning of the section and subsequently remained stable after a short period of downwards adaptation.

Nose tip temperature

Nose tip temperature decreased as a function of task difficulty compared to the driving-only baseline sections (see table 5.7 and figure 5.13b), resulting in a significant main effect of task difficulty, $F(2, 87) = 5.46, p < .01, \eta_p^2 = .11$. Post-hoc analyses showed that nose tip temperature only differentiated between the 'low' and the 'medium' respectively the 'high' condition ($p_{low-medium} < .01$ and $p_{low-high} < .001$) but not between the 'medium' and the 'high' condition ($p_{medium-high} = .84$). Visual inspection of the time series of nose tip temperature for all three task difficulty conditions (see nose tip temperature section in appendix B.3.2) showed that for the 'medium' as well as the 'high' condition, temperature values dropped immediately with the introduction of the secondary-task and continued to decrease until the end of the task compared to driving-only baseline sections. No visual differences between these task difficulty conditions was observed. The 'low' condition exhibited only a small decrease of temperature values in the beginning but then re-adapted to the level of preceding driving-only baseline temperatures.



(a) Change in heart rate compared to driving-only baseline in %

(b) Change in nose tip temperature compared to driving-only baseline in %

Figure 5.13: Experiment 1 - Mean values of cardiovascular measures for each task difficulty. Error bars represent the standard deviations

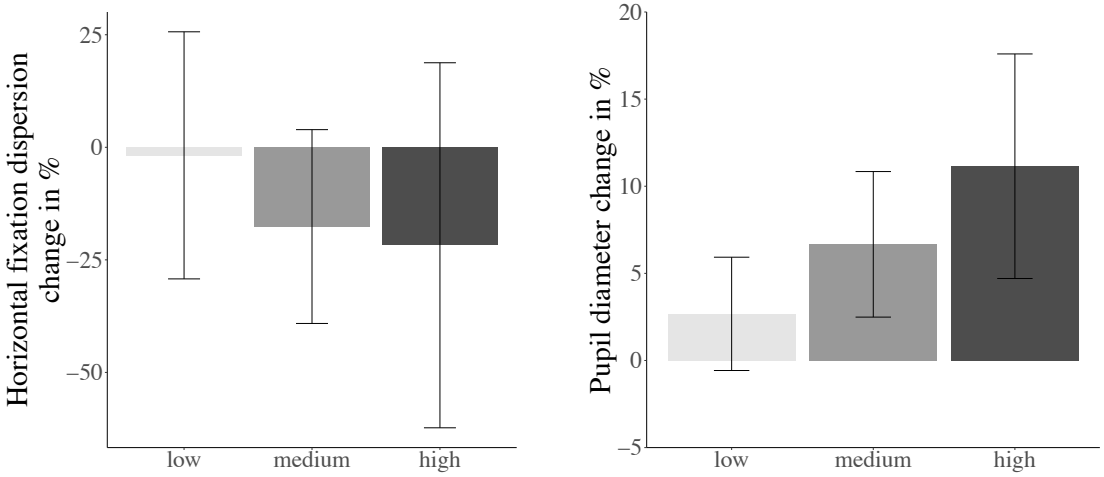
Ocular measures

Horizontal fixation dispersion

One participant in the 'medium' task difficulty condition had to be excluded from data analysis because of missing eye-tracking data ($N_{low} = 30$, $N_{medium} = 29$ and $N_{high} = 30$). Horizontal fixation dispersion decreased as a function of task difficulty compared to the baseline driving sections (see table 5.7 and figure 5.14a), resulting in a significant main effect of task difficulty, $F(2, 86) = 7.79, p < .01, \eta_p^2 = .15$. Post-hoc analyses showed that horizontal fixation dispersion only differentiated between the 'low' and the 'medium' respectively the 'high' condition ($p_{low-medium} < .01$ and $p_{low-high} < .001$) but not between the 'medium' and the 'high' condition ($p_{medium-high} = .64$). Additional inspection of the time series plots for all task difficulty levels revealed generally higher variability over time compared to e.g. heart rate and nose tip temperature (see horizontal fixation dispersion section in appendix B.3.2). However, depending on the selected window size, the average level of horizontal fixation dispersion seems to be different for all task difficulty conditions, with lowest average values in the high task difficulty condition.

Pupil size

Pupil size increased as a function of task difficulty compared to driving-only baseline sections (see table 5.7 and figure 5.14b), resulting in a significant main effect of task difficulty, $F(2, 86) = 35.09, p < .001, \eta_p^2 = .45$. Post-hoc analyses showed that pupil size differentiated between all three task difficulty conditions ($p_{low-medium} < .001$, $p_{low-high} < .001$ and $p_{medium-high} < .001$). Visual time series analysis for all three task difficulty conditions shows high variability, similarly to the results of the horizontal fixation dispersion (see pupil section in appendix B.3.2). All three task difficulty conditions exhibited initial strong increases in pupil size with low latency to the onset of the secondary-task. After this initial increase, pupil sizes adapted to lower levels respective to the task difficulty condition with lowest average declines for the 'high' and strongest for the 'low' condition. In general, all three task difficulties exhibited a decreasing trend in pupil sizes over time.



(a) Change in horizontal fixation dispersion compared to driving-only baseline in %

(b) Change in pupil size compared to driving-only baseline in %

Figure 5.14: Experiment 2 - Mean values of ocular measures for each task difficulty. Error bars represent the standard deviations

Speech signal features

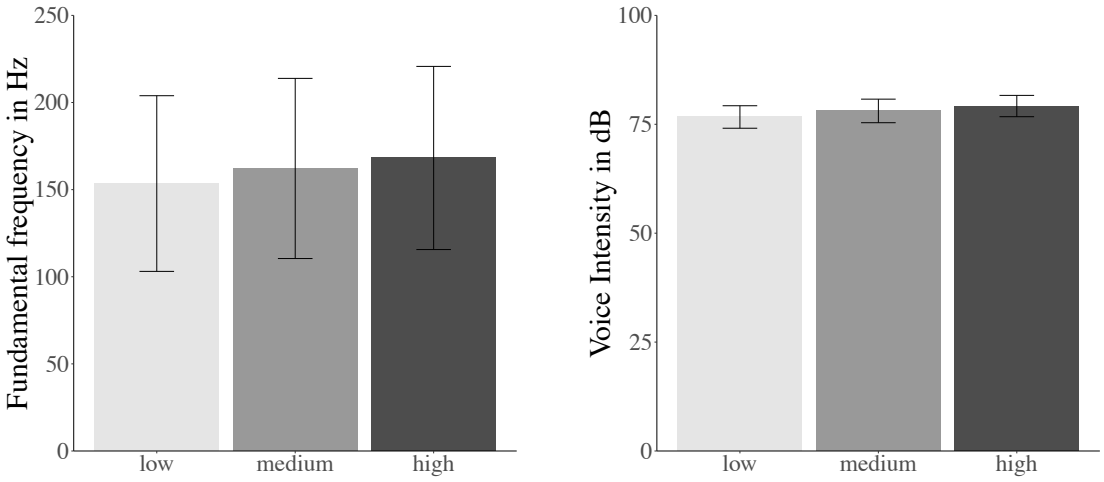
Overall, three participants had to be excluded from data analysis because of corrupted speech recordings. This resulted in the following number of participants for the three task difficulty conditions, $N_{low} = 29$, $N_{medium} = 29$ and $N_{high} = 29$. Originally, it was intended to report the speech signal features as change values compared

Table 5.8: Experiment 2 - Mean (M) absolute values and standard deviations (SD) of the speech signal features for each task difficulty

	low		***medium***		***high***	
	M	*SD*	*M*	*SD*	*M*	*SD*
Mean fundamental frequency in Hz	153.49	50.41	162.15	51.69	168.16	52.56
Mean voice intensity in dB	76.68	2.58	78.07	2.71	79.19	2.45

to the pre-recorded speech baseline. However, analysis of this approach revealed that both fundamental frequency of the voice as well voice intensity of the pre-recorded resting speech baselines exhibited higher mean values than in the experimental conditions. This observation is contrary to the assumptions that heightened cognitive demands increase these voice characteristics. Therefore, it was decided to omit this approach and instead use the absolute values similarly to the approach in experiment 1. This decision is further discussed in the limitations section of this chapter (see section 5.5). For statistical testing gender was added as between-subjects factor to the ANOVA. Fundamental frequency increased as a function of task difficulty compared to the driving-only baseline sections (see table 5.8 and figure 5.15a), resulting in a significant main effect of task difficulty, $F(2, 81) = 3.81, p < .05, \eta_p^2 = .09$. Post-hoc pairwise comparisons showed that fundamental frequency did not differentiate between any of the three task difficulty conditions ($p_{low-medium} = .51$, $p_{low-high} = .15$ and $p_{medium-high} = .72$). Voice intensity also increased as a function of task difficulty compared to the driving-only baseline sections (see table 5.8 and figure 5.15b), resulting in a significant main effect of task difficulty, $F(2, 81) = 8.20, p < .01, \eta_p^2 = .17$. In contrast to fundamental frequency, post-hoc analysis showed, that voice intensity differentiated between all three task difficulty conditions ($p_{low-medium} < .01$, $p_{low-high} < .001$ and $p_{medium-high} < .05$).

The factor gender was highly significant regarding the fundamental frequency ($F(1, 81) = 456.85, p < .001, \eta_p^2 = .85$) but not for the voice intensity ($F(1, 81) = 0.72, p = .40, \eta_p^2 = .01$). Overall, female participants exhibited higher fundamental frequencies compared to male participants (female: $M = 209.57, SD = 27.96$, male: $M = 116.19, SD = 14.84$). No significant interaction of gender and task difficulty was found for either of the two measures, indicating that the effect of task difficulty was the same for both genders.



(a) Voice fundamental frequency in Hz

(b) Voice intensity in dB

Figure 5.15: Experiment 2 - Mean values of speech signal features for each task difficulty. Error bars represent the standard deviations

5.4 Discussion

The aim of this experiment was to validate the results obtained in the first study in a more applied scenario and to evaluate which measures are most suitable for a combination of multiple cognitive workload measures that discriminate between multiple cognitive workload levels. For this reason hypotheses regarding performance, subjective and physiological measures were formulated. First of all the impact of task difficulty on continuous and event-related performance measures will be discussed. Subsequently, the impact of task difficulty on subjective measures will be discussed. On the basis of performance and subjective results it will be inferred whether the three task difficulty conditions invoked three cognitive workload levels. Afterwards the consequences regarding physiology will be interpreted with regard to this and their suitability for a multi-measure approach will be discussed.

5.4.1 Performance measures

Driving performance - Continuous driving measures

Continuous lateral control

No changes were expected for measures of lateral control. However, the mean and standard deviation of the lateral deviation as well as the mean steering acceleration changed significantly with increasing task difficulty. As discussed in subsection

3.4.1, the previous results regarding the effects of cognitive workload on lateral vehicle control are mixed. The results of this experiment are in line with previous results by Liang and Lee (2010), He et al. (2013) and Pavlidis et al. (2016) that showed improved lane maintenance and less smooth steering behavior with increasing cognitive demands. It should be noted that particularly the effect on the mean and standard deviation of lateral deviation is small in absolute terms and is therefore in line with the results of meta-analyses by Horrey and Wickens (2006) and Caird et al. (2008). The steering acceleration, however, almost doubled from the low to the high task difficulty condition, indicating substantial changes in lateral control behavior. In summary, an effect of cognitive task difficulty on lateral vehicle control seems be present but small. Participants tend to show less variable lateral control behavior accompanied with more abrupt steering. Liang and Lee (2010) found similar results to those of this experiment and suggest that decreased lane-keeping variability can be attributed to risk compensational strategies by the driver to broaden safety margins. The more impetuous steering can also be attributed to risk compensation as well as to diminished control under high cognitive demands. This suggests that, whether this behavior can be interpreted as an overall increase or decrease in performance remains up to debate.

Continuous longitudinal control

Minor decreases in drivers' speed were expected as a reaction to increasing task difficulty. Overall, the participants reduced their speed in the presence of a cognitive secondary-task but the decrease was not sensitive for differences in task difficulty. This indicates that compensational speed reduction is more generally associated with cognitive secondary-task activity regardless of the task difficulty.

Driving performance - Critical event performance

Regarding performance related to critical events, three different dependent variables were used. First, the number of crashes, near crashes and no crashes was compared between the three different task difficulty levels. Secondly, brake reaction times were measured to investigate to what extend different task difficulties impact participants' ability to properly react to those critical events. Thirdly it was recorded if participants reacted to the events at all by breaking and whether this differed between different task difficulties.

Collisions

According to the literature described in subsection 3.4.1, it was expected that participants' ability to react to critical events would decrease with increasing task difficulty. However, no differences were found regarding the proportion of crashes, near crashes and no crashes between the three task difficulty conditions. Interestingly, participants in the medium task difficulty condition produced the least number of crashes compared to the low and high condition. Even though this difference did not reach significance, it is interesting that similar numbers of crashes occurred in the low and high condition. According to the literature, it is assumed that general sensitivity to visual stimuli decreases with increasing task demands, thus theoretically resulting in the lowest number of crashes in the low task difficulty condition and highest in the high task difficulty condition (R. Young, 2012). There are several possible explanations for the results that were observed in this experiments. It is possible that the approach of calculating the collisions, as described in subsection 5.2.5, does not produce reliable results. However, the calculations were performed equally for all task difficulty conditions and it is unlikely that a systematic error was introduced. It is also reasonable to question whether the events presented to the participants were sufficiently critical while at the same time possible to avoid by an appropriate reaction. For all task difficulty conditions the percentage of near crash and crash occurrences together was around 20% (low: 21% , medium: 17%, high: 20%). For that reason, it is believed that overall the criticality of the events was within reasonable boundaries. The possibility to predict events after the first event might also have influenced the results. Participants' possibly anticipated the second and third event. It is however unlikely that this introduced a systematic error. Precautionary methodological actions like the dummy block with no event that always followed the first event and the counterbalanced block order were in place to counteract predictability. Most likely most participants were able to compensate the increasing attentional demands through higher compensational effort expenditure (G. Mulder, 1986; Hockey, 1997; M. S. Young et al., 2015) as well as a strategical shift of focus to the driving task to maintain performance. If this was the case, one would expect decreased secondary-task performance as well as physiological reactions to increasing cognitive demands, which will be discussed in the following sections. Regarding the low number of crashes in the medium task difficulty condition, it is possible that participants in the low task difficulty condition were in underload and participants in the high task difficulty condition in overload workload regions

(De Waard, 1996; M. S. Young et al., 2015). In both cases, it can be argued that participants were not in an optimal attentional state, leading to similar numbers of crashes in both conditions for different reasons. This interpretation, however, cannot be verified on the basis of the aforementioned results alone and has to be backed up by further evidence. For this reason, the following discussion will critically examine other results with regard to this assumption.

Brake reaction times

As regarding the number of collisions, it was expected that brake reaction times increase as a function of task difficulty. Results suggest otherwise, though. Participants in the medium task difficulty condition reacted fastest to critical events compared to participants in the other two conditions, while participants exhibited similar brake reaction times in the low and high task difficulty conditions. This reflects a pattern similar to that regarding the number of collisions and corroborates the idea that the participants from the low and high task difficulty groups were in suboptimal workload regions.

Omitted brake reactions

As for the other two measures, an increased number of omitted brake reactions with increasing task difficulty was expected. Again, the results implicate benefits for the medium task difficulty condition. Descriptively there were fewer omitted brake reactions in the low than in the high task difficulty condition, though there was no significant difference. Overall the results show the same pattern as the other event-related measures, i.e. revealing subtle benefits for the medium task difficulty condition.

Summary primary-task performance

Regarding continuous driving performance for measures of lateral as well as longitudinal control only minor changes were found in response to different cognitive task difficulties. Although changes in steering acceleration were somewhat larger compared to speed and lane-keeping variability, no safety relevant declines in performance were observed. Regarding event-related measures, results did not reveal a linear relationship between the cognitive task difficulty and event-related driving behavior. In fact, slight benefits in performance in the medium task difficulty condition provide evidence for a non-linear relationship corroborating the existence of under- and overload workload regions as e.g. described by De Waard (1996), De Waard and Brookhuis (1997) and M. S. Young et al. (2015) and optimal performance between the two regions.

Secondary-task performance

Secondary-task performance exhibited the expected pattern. The performance in the n-back task significantly declined as a function of task difficulty. Hardly any errors were made in the low task difficulty condition, suggesting that the simple repetition of the digits did only involve a negligible amount of attentional resources and could be accomplished almost automatically. Interestingly, in the medium task difficulty condition, the error rate in the n-back task was low, too. This allows for the interpretation that participants were able to compensate higher task demands without a decline in performance. The drop of secondary-task performance in the high task difficulty condition was considerably more substantial, indicating that participants were not able to fully compensate for the increased attentional demands. Additionally, the time series analysis revealed that participants were not able to maintain secondary-task performance in the highest task difficulty condition. Performance decreased over time, probably because participants decreased their task engagement over the course of time. This would be consistent with the experimenters' observations that participants tried keeping up the performance in the beginning of each driving section with secondary-task but more and more neglected the secondary-task over the course of time until a comfortable level was reached.

Summary on performance measures

Considering primary- and secondary-task performance, it can be concluded that the three task difficulty conditions resulted in three different cognitive workload levels (see table 5.9). The observed patterns of primary- and secondary-task performance implicate different effects on continuous driving performance and secondary-task performance compared to event-related driving performance. Whereas results regarding the former suggest a linear relationship, the patterns in event-related performance indicate a non-linear relationship. To validate this assumption the subjectively perceived workload of the participants can be used as an additional indicator.

5.4.2 Subjective measures

An overview of the results of the subjective measures can be found in table 5.9. The SEA ratings indicate that all blocks with critical events were perceived as equally effortful. Furthermore, the ratings showed a significant effect of task difficulty on perceived effort. However, the ratings did not discriminate between the medium and high task difficulty condition. The same pattern was found for the overall NASA-TLX

ratings. Most interestingly, the only single dimension of the NASA-TLX that discriminated between all three task difficulties was the dimension of mental demands. As it was methodologically intended to specifically increase cognitive demands, it is not surprising that the dimension of mental demands is more sensitive for the differences in cognitive task difficulty than other NASA-TLX dimensions like e.g. temporal demands. It is further not surprising, that the less sensitive NASA-TLX dimensions 'dilute' the results of the mental demands dimension, resulting in non-significant differences between the medium and high task difficulty conditions in the overall workload ratings. Effort, frustration and performance ratings were similar for the medium and the high task difficulty condition. This suggests that participants invested the same effort in both conditions even though the perceived mental demands were different. This indicates that participants either limited their effort to an acceptable level or a ceiling level was reached, where further investment of effort would not have resulted in better performance. Additionally, similar ratings on the physical demand dimension for all task difficulty conditions suggest no confounding physical influences on physiological measures. In summary, the overall subjective workload ratings do not directly support the assumption that three different workload levels were induced. Nevertheless, results from both questionnaires (SEA and NASA-TLX) showed higher descriptive overall values for the high task difficulty condition compared to the medium task difficulty condition by trend. Additionally, the dimension of mental demand was sensitive for different cognitive demands imposed on the participants. Therefore, it is reasonable to assume that the experimental manipulation led to three different cognitive workload levels.

Summary of performance and subjective measures

Taken together, performance and subjective measures suggest that three different workload levels resulted from the experimental manipulation (see table 5.9). Primary- and secondary-task performance differed significantly between the task difficulty conditions and the mental demand dimension of the NASA-TLX clearly showed increased ratings with increasing demand. With respect to the event-related performance measures, it has to be noted, that the relationship between task difficulty and performance did not constitute itself in a linear interrelation. The reaction to critical events was best for the medium task difficulty condition compared to the other two conditions, whereas for the lateral continuous performance measures as well as the subjective workload ratings a linear relationship was observable. Both latter measures exhibited

increasing changes with increasing task difficulty. These results imply, that task difficulty impacted the event-related performance in a different way compared to lateral continuous driving measures and subjective measures and suggest that participants in the low task difficulty condition may have been in underload and the participants in the high task difficulty condition in overload workload regions. Therefore the following discussion of the physiological changes will also examine this assumption.

5.4.3 Physiological measures

One central goal of the second experiment was to validate the results of the physiological measures obtained in the first experiment in a more realistic setting. Furthermore, it was aimed at assessing which measures are most promising for multi-measure approaches. Additionally, the physiological results are discussed with respect to the assumption that participants in the low task difficulty condition were in underload and the participants in the high task difficulty condition were in overload workload regions (see subsection 5.4.1). A summary of the physiological measures can be found in table. 5.9

Cardiovascular measures

Heart rate

Heart rate is a popular and robust measure of mental workload and is thought to mainly reflect the amount of mental effort that is invested in a task (Kahneman et al., 1969; Jorna, 1992; L. Mulder, 1992). It was expected that heart rate increases with increasing task difficulty. Except for the study conducted by Mehler et al. (2009), most studies showed that particularly in high workload regions, differential sensitivity of heart rate is limited (Engström et al., 2005; Lenneman et al., 2005; Reimer & Mehler, 2011). The current results are in line with most of the past research. Heart rate increased with increasing task difficulty but failed to discriminate between medium and high task difficulty. Interestingly, the aforementioned study by Mehler et al. (2009) used the same secondary-task (0-back, 1-back, 2-back) as was used in the current study, but heart rate in the study by Mehler et al. (2009) discriminated between all three task difficulties. One main difference between the two studies is that in the experiment of Mehler et al. (2009) the secondary-task was divided into sequences of ten numbers with a pause of 30 seconds between consecutive secondary-task sequences. In the present study, the n-back task was performed continuously

over a period of almost four minutes in each block. Time series analysis (see appendix B.3.2) revealed that after an initial increase in heart rate, an adaptation to a lower level took place over time. Even though the initial increase was the same for the medium and high task difficulty condition, heart rate in the high task difficulty condition dropped below the heart rate in the medium task difficulty condition. This observation suggests that participants in the high task difficulty condition were not able to compensate for the high demands with higher effort investment over a longer period of time. This provides further evidence for the assumption made that participants in the high task difficulty condition reached an overload workload level in subsection 5.4.1, where further effort investment did not result in the maintenance of overall performance and therefore lowered their effort to a level that satisfied their subjective performance goals. It is possible that participants in the study of Mehler et al. (2009), contrary to the participants of this study, were able to invest more effort due to shorter task durations but failed to recognize, that the increased effort did result in maintenance of performance as the high error rate in the 2-back condition of Mehler et al.'s study suggests.

The results regarding heart rate neither confirm nor contradict the assumption that participants in the low task difficulty condition might have been in suboptimal underload regions. Although it can be argued that heart rate would have to drop below or remain on baseline levels to confirm this assumption, it is equally possible that the above baseline heart rates in the low task difficulty condition were a result of changes in respirational patterns due to speech production rather than a result of the cognitive demands.

Nose tip temperature

Nose tip temperature was expected to decrease with increasing task difficulty. This could be confirmed by the current experiment. However, nose tip temperature, similar to heart rate, did not differentiate between the medium and high task difficulty condition, though some differences between nose tip temperature and heart rate were found. These differences are related to the fact that heart rate is subject to parasympathetic and sympathetic influences (Berntson et al., 1991; Backs, 1995; Lenneman & Backs, 2009), whereas nose tip temperature is thought of as being solely influenced by sympathetic activity (Drummond, 1994; Rimm-Kaufman & Kagan, 1996). Nose tip temperature values were above baseline level in the low task difficulty condition and dropped below baseline for the other two conditions. This is contradictory to the results presented by e.g. Or and Duffy (2007) who reported that

nose tip temperature drops when adding a cognitive secondary-task to a driving task. One explanation could be that in the lowest task difficulty condition no increase in sympathetic activity was induced by the relatively simple 0-back task. In higher task difficulty conditions however, sympathetic activity increased due to higher cognitive demands. This explanation is in line with the results by Lenneman and Backs (2009) that showed that a single driving task increases parasympathetic withdrawal compared to resting but only found increases of reciprocally coupled sympathetic activity and parasympathetic withdrawal for high n-back task difficulties. This would explain why no drop in nose tip temperature was found for the lowest task difficulty condition as this measure is insensitive to parasympathetic changes. At the same time this could be an additional explanation for the increase in heart rate compared to baseline in the low task difficulty condition due to parasympathetic withdrawal. Time series analysis revealed differences between the progression of heart rate and nose tip temperature over time for the medium and high task difficulty condition. Whereas the drop in nose tip temperature over time was the same for both conditions, the change in heart rate was different for the medium and high task difficulty condition and different to the progression of nose tip temperature (see appendix B.3.2). Sympathetically invoked changes in nose tip temperature show an almost linear decline over time whereas changes in heart rate are more complex and non-linear. Assuming that the results on nose tip temperature purely reflect sympathetic activity, this provides further evidence that heart rate increases under highly demanding situations due to heightened sympathetic activity (Kahneman et al., 1969; Jorna, 1992; L. Mulder, 1992; Wilson, 1992; L. B. Mulder et al., 2004), but at the same time elucidates the importance of parasympathetically induced compensational processes in cognitively highly demanding situations. Nevertheless, it can of course be argued that the high temporal latency in changes of skin temperature prevent sympathetically invoked adaptive processes to be reflected in changes in skin temperature. To rule out this possibility further studies with longer secondary-task durations would be needed.

Summary on cardiovascular measures

In summary, the results regarding cardiovascular measures validate and extend the results of the first experiment in a more realistic driving scenario. Heart rate proved to be a reliable and robust indicator of cognitive workload, despite lacking sensitivity in higher workload regions. However, additional time series analysis revealed insights into adaptive processes that can be of value for the interpretation of changes in cognitive workload over time. The differences in the progression of heart

rate over time between all three task difficulties further supports the assumption that indeed three different workload levels were induced. These interpretations would not have been possible based on the comparison of the averaged values over the task difficulty conditions alone. The results on nose tip temperature also validate the results from the first experiment as well as empirical evidence of past research presented in subsection 3.6. Nonetheless nose tip temperature appears to be an inferior measure of cognitive workload with regard to sensitivity and temporal characteristics compared to heart rate. Hence, the inclusion of heart rate in a multi-measure approach is more promising than the inclusion of nose tip temperature.

Ocular measures

Horizontal fixation dispersion

A huge amount of past research has consistently reported lower gaze dispersion under cognitive workload in a variety of driving settings and cognitive secondary-tasks (see subsection 3.6). Even though the results from the first experiment contradicted these findings, it was expected that horizontal fixation dispersion decreases with increasing task difficulty as it was assumed that the results from the first experiment were highly influenced by the artificial nature of the driving task (LCT) and that horizontal fixation dispersion would decrease under cognitive workload in a more realistic scenario. Indeed, horizontal fixation dispersion decreased as a function of task difficulty in the current experiment. However, similar to heart rate and nose tip temperature, this measure did not differentiate between the medium and high task difficulty condition. These results are similar to the results obtained by e.g. Reimer et al. (2012) and Niezgoda et al. (2015) using the same secondary-task (0-back, 1-back, 2-back). In both studies no further decrease in horizontal fixation dispersion occurred for the highest task difficulty compared to the medium task difficulty. One possible explanation could be a lower limit for horizontal gaze concentration (Reimer et al., 2012). This lower limit could be related to the visual area, that minimaly needs to be covered to keep up driving performance. According to the SEEV model (Wickens et al., 2003) and research by Horrey et al. (2006), visual attention allocation is highly dependent on the value of a stimulus. Therefore the value of the visual area to be covered to maintain driving performance could be the reason why horizontal fixation dispersion in the high task difficulty condition does not further increase compared to the medium task difficulty condition. However Recarte and Nunes (2000) showed reductions in horizontal fixation dispersion that were three times higher as compared

to the current results. Therefore this assumption does not seem likely to be true. Insensitivity to differences between medium and high task difficulty might additionally be explained with respect to the size of the visual span. Visual span decreases with increasing cognitive demands induced by an auditory task as has been shown by Pomplun et al. (2001). So theoretically the span size should decrease with increasing task difficulty, leading to lower gaze dispersion to compensate for it (described in subsection 3.6.2). Participants in the highest task difficulty condition, however might have been aware that overall performance could not be maintained over time resulting in similar levels of effort as compared to the medium task difficulty condition. The decrease of secondary-task performance, the progression of heart rate over time as well as equal effort ratings for both conditions in the NASA-TLX support the idea that they reduced their engagement in the secondary-task to cope with the overall task demands. Therefore effects on visual span could have been similar for both task difficulties. Additionally, time series analysis confirmed that horizontal fixation dispersion (see appendix B.3.2) is to some extend dependent on the visual task. Even though the secondary-task was continuous over time, variations in horizontal fixation dispersion were apparent and followed similar patterns in all three task difficulty conditions, suggesting that these variations were the result of variations in the driving task demands.

Pupil size

Past research (see subsection 3.6.2) as well as results from the first experiment suggest that pupil size increases with increasing task difficulty. The results from the second experiment fully support this hypothesis and confirm that pupil size is a sensitive indicator of differences in task difficulties. Contrary to the first experiment, this time differences can be attributed to the changes in cognitive demands as they are not confounded by dissimilar inter-stimulus intervals. Additionally, time series analysis revealed a dependency of pupil size on the driving task, similarly to the results of horizontal fixation dispersion. However, the pattern of changes of the pupil over time seems to be more consistent between task difficulties. Furthermore, the absolute values regarding pupil changes for the highest task difficulty condition are close to the maximum of $.5mm$ that is reported in the literature (e.g. Beatty & Lucero-Wagoner, 2000). This suggests that workload levels in the high task difficulty condition were high on an absolute level and is further evidence of a possible overload situation in the highest task difficulty condition.

Summary ocular measures

Both ocular measures appear to be useful in the discrimination of different cognitive workload levels. Considering the results of both experiments, however, pupil size seems advantageous regarding sensitivity and reliability for multi-measure approaches.

Speech signal features

The absolute values of fundamental frequency and voice intensity, were expected to increase as a function of task difficulty. The results of both measures indicate that they are sensitive to increasing task demands. Whereas voice intensity discriminated between all three task difficulties, fundamental frequency did not differentiate between any of the task difficulty conditions. Interestingly, these effects seem to be the same for both genders. However, whereas in the first study both measures exhibited the same differential sensitivity, in the second experiment only the voice intensity seemed feasible to differentiate between different workload levels. Therefore further research is needed to investigate the differential sensitivity of both measures. Furthermore, time series analysis revealed that both measures were stable over time, indicating that the different visual-manual demands from the driving task did not impact the speech signal features. This also implicates robustness against momentary changes in overall workload. With respect to the capability of fundamental frequency and voice intensity for multi-measure approaches, it can be concluded that both measures can be useful general indicators of cognitive workload. However, the challenge to obtain valid baseline values of speech signal features might limit their usefulness. Additionally, confounding influences of e.g. heart activity on speech production described by Orlikoff and Baken (1989) have to be considered.

Summary of physiological measures

The results of the physiological measures from this second experiment mainly replicate the patterns observed in the first experiment (see table 4.6 and 5.9). Again, pupil size and heart rate revealed to be most sensitive to changes in cognitive driver workload. In addition, it was shown again that the majority of physiological measures used in the second experiment lack differential sensitivity in higher workload regions. Only pupil size and voice intensity differentiated between medium and high levels of cognitive driver workload.

Table 5.9: Experiment 2 - Statistical significance for main effect of task difficulty and pairwise comparisons for performance, subjective and physiological measures

	Main effect of task difficulty	**Effect size** η_p^2	**low vs. medium**	**low vs. high**	**medium vs. high**
Primary task performance					
Driving					
Mean lateral deviation in m*	sig.	.06	sig.	sig.	n.s.
Standard deviation of lateral deviation in m*	sig.	.24	sig.	sig.	n.s.
Mean steering acceleration in rad/s*	sig.	.10	sig.	sig.	n.s.
Mean speed in km/h	n.s.	$< .01$	-	-	-
Brake reaction time	sig.	.12	sig.	n.s.	sig.
Number of crashs	n.s.	-	-	-	-
Number of ommited brake reactions	sig.	-	n.s.	n.s.	n.s.
Secondary task performance					
N-back					
Proportion of correct responses	sig.	-	sig.	sig.	sig.
Subjective measures					
SEA - Score	sig.	.34	sig.	sig.	n.s.
NASA-TLX					
Mental Demand	sig.	.44	sig.	sig.	sig.
Physical Demand	n.s.	.03	-	-	-
Temporal Demand	sig.	.16	sig.	sig.	n.s.
Performance	sig.	.11	n.s.	sig.	n.s.
Effort	sig.	.56	sig.	sig.	n.s.
Frustration	sig.	.14	n.s.	sig.	n.s.
Overall workload	sig.	.46	sig.	sig.	n.s.
Physiological measures					
Heart rate	sig.	.26	sig.	sig.	n.s.
Nose tip temperature	sig.	.11	sig.	sig.	n.s.
Horizontal fixation dispersion	sig.	.15	sig.	sig.	n.s.
Pupil size	sig.	.45	sig.	sig.	sig.
Voice fundamental frequency	sig.	.09	n.s.	n.s.	n.s.
Voice intensity	sig.	.17	sig.	sig.	sig.

'sig.': significant (α-level of $p = .05$)
'n.s.': not significant
'*': α-level adjusted to $p = .2$
'-': not calculated

5.5 Limitations of the experiment

The experiment has several methodological limitations. The most important limitation is, that the recorded speech baselines could not be used as intended. Both fundamental frequency values and voice intensity values were higher in the baseline recordings than during the experiment. This can be ascribed to methodological issues during recording. First of all the conditions in which the speech baselines were recorded were different to the experimental conditions. Participants were seated at a desk and did not wear headphones like they did during the experiment. They were instructed to count from zero to nine and speak clearly while doing so. This might have provoked higher voice intensities (louder speech) resulting from the participants attempt to follow the instructions. Furthermore, the speech baseline was recorded before the participants got the possibility to get familiar with the driving simulator. Therefore it cannot be ruled out that participants exhibited anticipatory stress reactions resulting in increased fundamental frequency and voice intensity as both measures are known to be sensitive to psychological stress (Mendoza & Carballo, 1998; Scherer et al., 2002). Alternatively the speech baseline could have been recorded shortly before the experiment after the training to get more valid measures. However, it is unclear whether this would have eliminated a possible anticipatory stress reaction.

In contrast to the first experiment, eye-tracking as well as temperature measurements were not recorded remotely but via eye-tracking glasses and a temperature sensor attached to the participants' nose. This might have influenced the physiological reactions due to increased discomfort. However, the use of individual baselines should counteract these influences. Additionally, the physical demands dimension of the NASA-TLX revealed no differences between the three task difficulties suggesting that if there was a negative influence from the measurement method, it was the same for all task difficulty conditions. In connection to that it also should be noted that a small number of participants finished the experiment prematurely because of signs of simulation sickness. Although participants were instructed to signal any signs of simulator sickness before and during the experiment, it can not be ruled out that some of them experienced symptoms like nausea or headaches during the experiments without reporting them. In future studies a questionnaire assessing simulator sickness like the SSQ (Kennedy, Lane, Berbaum, & Lilienthal, 1993) should be used to avoid confounding influences on physiological measures.

Finally it should be mentioned that similar to the first experiment a location based fixation detection algorithm was used. As described before in section 4.5 of the first experiment, the algorithm would erroneously split one fixation up into several fixations when smooth pursuit eye-movements larger than 80px are made by the participants. Unfortunately this algorithmic approach is up to now without alternative when using eye-tracking data sampled at 60Hz. It is however believed, that the number of smooth pursuit movements in the second experiment is low if present at all.

5.6 Conclusions

Most of the methodological issues from the first experiment were successfully addressed in this second experiment. Furthermore the more realistic experimental setting allows for a broader generalisation of the results. Results of continuous driving performance reflects results from the literature. Event-related driving performance however suggests, that there is no simple interrelationship between cognitive task difficulty and reaction to critical events. This emphasises the need for workload measures that are able to differentiate between under-, optimal and overload workload states. Furthermore most of the results regarding physiological measures of the first experiment could be validated in the second experiment. Heart rate, nose tip temperature and pupil size exhibited similar patterns and the speech signal features proved to be sensitive to changes in cognitive workload although the differential sensitivity was different compared to the first experiment. Additionally horizontal gaze dispersion results replicated findings from earlier research and proved to be a suitable measure of cognitive workload. Finally, this experiment identified and compared physiological parameters regarding their potential to be used for the combination of multiple measures to classify different workload levels in the following chapter.

Chapter 6

Combination of physiological measures

In the previous chapters, the impact of auditory-verbal secondary tasks on cognitive workload was examined. The sensitivity of single physiological measures to multiple levels of cognitive workload was tested and validated in two experiments. The following chapter will now focus on the combination of multiple physiological measures for the classification of drivers' cognitive workload. For this purpose first of all the importance and challenges of driver workload classification will be explained (section 6.1). After that theoretical and methodological concepts for the combination of multiple physiological measures will be introduced and past research on driver cognitive workload classification will be presented (section 6.2). This is followed by the methodological approach which was used to build cognitive workload classifier models on the basis of the data of the second experiment of this thesis (section 6.3). Finally, the results will be presented and discussed (sections 6.3.4, 6.4 and 6.5).

6.1 Introduction

The reliable estimation of mental workload has long been a goal of ergonomics research (M. S. Young et al., 2015). If the psychological state was known by the technological system that the person interacts with, steps to prevent e.g. mental overload or underload could be implemented. Particularly in the context of driving, research focuses on the development of workload evaluation and managing systems to avoid safety critical situations due to suboptimal workload levels (De Waard, 1996; Brookhuis, 2004; Piechulla, Mayser, Gehrke, & König, 2003; Lei, Toriizuka, & Roetting, 2017). In principle these systems try to resolve the asymmetrical relation between the driver and the car regarding the amount of information available to both sides (Hettinger,

Branco, Encarnacao, & Bonato, 2003; Fairclough, 2009). Whereas the driver is to some extend aware of the state of the car (engine speed, motor temperature, status displays and warnings etc.), the car has little to no information about the driver. Only if the car is aware of the state of the driver too, a collaborative interaction can be achieved. State aware systems can be used to directly manipulate workload by e.g. adapting the task load of a driver to the current driving situation (Piechulla et al., 2003; Lei et al., 2017).

Therefore reliable indicators are needed which allow conclusions to be drawn about the driver's psychological state. Compared to e.g. visually-manually induced driver workload, assessing drivers' cognitive workload is more challenging *"... because of the problems associated with observing what a driver's brain (as opposed to hands or eyes) is doing."* (Strayer et al., 2015, p.1301). Nevertheless, past research as well as the results from this thesis, indicate that physiological measures can be used to determine the cognitive workload state of the driver. Particularly the use of ANS regulated physiological responses to increasing workload levels is very promising, since they can be measured non-intrusively without interfering with the driving task. However, a single physiological workload measure is often not sensitive enough to differentiate between different levels of workload. As described earlier (see section 2.2.3 and 3.6), different physiological measures are sensitive to different aspects of mental workload. Whereas e.g. heart rate is considered to be an indicator of mental effort, pupil size changes reflect differences in task difficulty. In addition, different physiological systems are subjected to different influences of the ANS. In high load conditions, heart rate is mainly under sympathetic control whereas pupil size also reflects parasympathetic influences. For this reason it can be assumed that the combination of several ANS measures reflects a more complete picture of autonomic changes in response to different workload levels than single parameters do. A promising approach is to combine several physiological measures using data-driven modelling for a robust assessment of cognitive workload and has gained increasing attention for the assessment of drivers' cognitive workload (see e.g. Liang & Lee, 2014; Solovey, Zec, Garcia Perez, Reimer, & Mehler, 2014; Tjolleng et al., 2017).

The following section will describe a general approach of combining physiological measures to classify psychological states. Subsequently, results of past research regarding the use of physiological measures for the classification of drivers' cognitive workload will be presented. After that the development and test of cognitive workload classifier models on the basis of the data gathered in the second experiment (see chapter 5) will be presented and discussed.

6.2 Theoretical aspects of drivers' state classification

6.2.1 Definition of physiological computing

The process of using advanced computational methods to infer the psychological state of a person from physiological data and to apply countermeasures to avoid critical or undesired psychological states is often referred to as physiological computing. (Allanson, 2002; Allanson & Fairclough, 2004; Fairclough, 2009). Originally physiological computing was introduced *"...to transform bioelectrical signals from the human nervous system into real-time computer input in order to enhance and enrich the interactive experience"* (Allanson & Fairclough, 2004, p.857). The areas of application of physiological computing today can roughly be divided into two main areas (Novak et al., 2012). Affective physiological computing deals with the classification of and adaptation to emotional states (e.g. Picard & Picard, 1997; Kim, Bang, & Kim, 2004; Wagner, Kim, & André, 2005). Cognitive physiological computing deals with the classification of and adaptation to cognitive states like e.g. mental workload or drowsiness. In the context of the classification of drivers' cognitive workload, methods inspired by cognitive physiological computing approaches have recently gained increasing popularity (Putze, Jarvis, & Schultz, 2010; Borghini et al., 2014; Solovey et al., 2014; Liao et al., 2016; Tjolleng et al., 2017). This can be explained by the fact that cognitive workload presents a great risk to the driver (Horrey & Wickens, 2006; Caird et al., 2008; McEvoy, 2015) and is further fuelled by the increasing availability of low-cost physiological measurement systems and the advancements of machine learning algorithms.

6.2.2 Workflow of physiological computing

In their paper, Novak et al. (2012) describe an exemplary workflow to use physiological measures for the classification of physiological states in physiological computing (see figure 6.1). The following subsection will be based on the paper and extended to the specific context of classifying drivers' cognitive workload where necessary. Whereas physiological computing originally involves system adaption on the basis of the classified psychological state, the following subsection will limit itself to the classification task, since system adaptation of physiological computing is beyond the scope of this thesis. In order to build models for the classification of the psychological state, the following steps have to be accomplished.

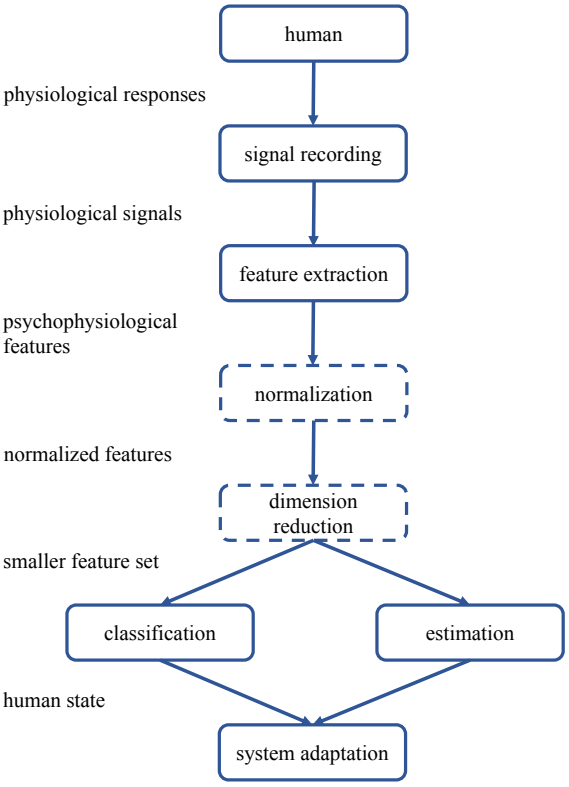


Figure 6.1: Flow of physiological computing adapted from Novak et al. (2012)

Psychological model

According to Novak et al. (2012) it is first of all crucial to select the appropriate psychological model. This is a necessary prerequisite to allow the recording of a data set where the physiological responses can actually be ascribed to the induced psychological state. Often in physiological computing only two 'extreme' psychological states are compared (e.g. bored vs. stressed, resting vs. high workload etc.). This is usually not done on purpose, but is a limitation resulting from the fact that it is hard to design experiments where more than two different psychological states can be reliably induced for one psychological concept. Whereas for example a resting state can easily be distinguished from a high workload state (see e.g Borghini et al. (2014) for an overview), it is harder to distinguish between a medium or high workload state through experimental manipulation. Additionally, Novak et al. (2012) state that the data collection itself needs to ensure that the results are not confounded by other factors like environmental influences or other psychological states. As all this is hard to achieve, additional measures should be applied to validate that the intended psychological state was actually induced. In case of cognitive workload the use of performance and subjective complementary measures is therefore an important possibility to ensure this. Furthermore the recordings of the raw physiological data have to be of high signal quality to minimize the impact of measurement artifacts on classification.

Feature extraction

Once the data set is collected, several pre-processing steps have to be considered according to Novak et al. (2012). First, features have to be extracted from the raw physiological signal (e.g. ECG). Depending on the knowledge about the psychological construct of interest, feature extraction methods reach from brute-force approaches, where as many features as possible are extracted from the raw physiological data to more informed methods where features are extracted that have been shown to be sensitive to the psychological states of interest. A general introduction to feature extraction can be found in Guyon, Gunn, Nikravesh, and Zadeh (2008) and examples of physiological feature extraction for e.g. emotion detection and stress classification are described in Kim et al. (2004), Wagner et al. (2005) and Zhai and Barreto (2006). In the context of drivers' cognitive workload, physiological features like time and frequency domain measures of heart rate, dispersion measures of eye fixations or pupil size measures etc. can be used. For an overview of popular physiological measures

for cognitive workload assessment please see subsection 3.6 of this thesis as well as Borghini et al. (2014). Besides the sensitivity of physiological features to a specific psychological state other criteria like e.g. validity, intrusiveness, operator acceptance etc. should be taken into account (G. Matthews et al., 2015). Additionally, the requirements resulting from the practical circumstances and limitations in the context of application should be considered. Such requirements can for example be if, it is required that the classification is accomplished in real-time or how much computing power there is available.

Normalization

The second step involves the normalization of the data to deal with inter- and intra-subject variability and mathematical issues regarding different data ranges when using algorithms like k-nearest neighbors. For this reason, Novak et al. (2012) recommend using one of the following approaches. The first option is to normalize features by using baseline measures as reference, secondly baseline values can be included as features and thirdly a simple normalization to a range of values (-1 to 1 or 0 to 1) can be applied. More elaborated approaches can be found in e.g. Picard, Vyzas, and Healey (2001) and Kim et al. (2004). In the context of classifying drivers' cognitive workload, the use of physiological baselines is probably the most useful approach as experimental procedures most often require the recording of baseline states to cope with interindividual differences.

Dimension reduction

According to Novak et al. (2012), the next step of pre-processing is dimension reduction as it becomes more difficult to find patterns and relationships in the data when the number of features is large. Additionally, using a large set of features comes with higher computational demands and the danger that the classifier does not generalize well to new data. Again Novak et al. (2012) propose several methods for dimension reduction. The most common and intuitive method is to use statistical analysis of the features regarding their sensitivity to discriminate between different psychological states and to rank them according to their discriminative power using p-values and effect sizes of ANOVAs or correlational analysis with e.g. subjective measures. Further methods are Principal Component Analysis (PCA) that maps the feature space to a dimension reduced space or sequential feature selection. For a general overview over feature extraction methods please see Fodor (2002) and for

examples in the context of physiological computing please see e.g. Kim et al. (2004) and Healey and Picard (2005). In the context of driver cognitive workload, statistical approaches as well as practical aspects like the feasibility of measuring a specific feature can be used to decide which features to include in the classification.

Classification

Following the aforementioned steps proposed by Novak et al. (2012) results in a feature matrix, containing feature vectors (X) labeled with the induced physiological state (y) that can then be used in supervised machine learning approaches. Supervised learning means to learn an unknown function from labeled data. The resulting models can then be used to predict the label of data that was not used to train the algorithm (Mitchell, 2006; Kotsiantis, 2007). In principle supervised learning can be used to predict categorical values (classification) or continuous values (regression). The following explanations however will be limited to the classification approach. For the process of building reliable classifiers it is of great importance to split the data into a training data set and a test data set. It is further important that the same preprocessing steps are applied to both. The training data is used to train a classifier. Here a computational method learns patterns and relationships between the physiological features and the corresponding psychological states (the labels). The test data can then be used to evaluate how good the classifier predicts new data samples or in other words the classification accuracy for unknown data. There is a plethora of machine learning algorithms that can be used for supervised learning tasks (Michalski, Carbonell, & Mitchell, 2013). Commonly used algorithms for classification are e.g. k-nearest neighbors (KNN), decisions trees and random forests (RF) as well as support vector machines (SVM). However, even though they can be used to solve the same problems, the basic principles behind these algorithms differ. KNN is a simple but yet effective method that predicts a new data point by calculating the distance to k nearest neighbors from the training set (Cover & Hart, 1967; Dasarathy, 1990). Decision trees learn a set of basic decision rules from the training data to predict new data (Breiman, Friedman, Stone, & Olshen, 1984). RF are meta/ensemble methods that make use of several decision trees (Ho, 1995). SVMs try to optimally separate two classes through hyperplanes that maximize the margin between the data points from the classes that are closest to each other (Cortes & Vapnik, 1995; Meyer & Wien, 2015).

For all aforementioned algorithms hyperparameters have to be specified that are not learned from the data. It is often not necessary to implement the machine learning

algorithms manually. There is a variety of software that allows easy and comfortable application of machine learning methods to own data like e.g. Python's scikit-learn package (Pedregosa et al., 2011) or Matlab's Statistics and Machine Learning Toolbox (MathWorks ®). The decision which machine learning algorithm with which hyper-parameter values to use for classification cannot easily be determined a-priori. For that reason often several methods are applied and compared with each other, using performance metrics like mean and standard deviation of classification accuracy. As it is beyond the scope of this thesis to elaborate on the multitude of machine learning methods and performance metrics the following articles by Caruana and Niculescu-Mizil (2006) and Kotsiantis (2007) are recommended for a detailed comparison of supervised machine learning algorithms and performance evaluation metrics. Furthermore it is referred to Solovey et al. (2014) and Vaish and Kumari (2014) for a comparison of different machine learning algorithms for classifications on the basis of ECG data.

When splitting the data into a training and a test set, several strategies can be applied according to the experimental procedure of data collection and the number of samples available (Novak et al., 2012). In subject-dependent approaches, the training and the test data set share subjects. This is often used when the data was collected using an experimental within-subjects design with small sample sizes. Subject-independent splitting means that the training and test data set do not share subjects and it is often used with bigger sample sizes. Whereas subject-dependent splitting often yields higher classification accuracies, subject-independent approaches are favorable when the classifier is used for a bigger target group of users because they generalize better. One commonly used method for a reliable estimation of a classifier's accuracy is to use cross-validation for the split of the data and the test of the classifier. For that the data is split into N data folds. $N-1$ folds are used for the training and one fold for the test of the classifier. This procedure is repeated N times until each single fold was used for testing once. The mean classification accuracy over the N folds then gives a reliable estimate of the overall accuracy of the classifier (see Kohavi (1995) fore a review of cross-validation methods).

6.2.3 Examples of drivers' state classification

The use of physiological computing methods for the classification of drivers' states has many applications. Borghini et al. (2014) give an extensive overview over the psychological models and physiological measures used in the context of driving. Methods of physiological computing are frequently and predominately used for the detection

of drowsiness and fatigue (Ji, Zhu, & Lan, 2004; Hu & Zheng, 2009; Zhao, Zheng, Zhao, Tu, & Liu, 2011; Chen, Zhao, Zhang, & Zou, 2015; Wang et al., 2016). This is understandable as fatigue is one of the main causes of severe traffic accidents. According to the AAA-Foundation for Traffic Safety 21% of all fatal accidents in the U.S. between 2009 and 2013 were due to drowsiness at the wheel (Tefft, 2014). However, similar approaches are also used for driver stress detection (e.g. Fernandez & Picard, 2003; Healey & Picard, 2005), driver inattention monitoring (e.g. Dong, Hu, Uchimura, & Murayama, 2011; Sahayadhas, Sundaraj, Murugappan, & Palaniappan, 2015) and driver cognitive workload classification (e.g. Zhang, Owechko, & Zhang, 2004; Kutila et al., 2007; Liang, Reyes, & Lee, 2007; Liang, Lee, & Reyes, 2007; Bořil, Omid Sadjadi, et al., 2010; Miyaji, Danno, Kawanaka, & Oguri, 2008; Putze et al., 2010; Liang & Lee, 2014). Most of the approaches regarding the classification of drivers' cognitive workload differentiate between two very different workload levels (Zhang et al., 2004; Kutila et al., 2007; Liang, Reyes, & Lee, 2007; Liang, Lee, & Reyes, 2007; Bořil, Omid Sadjadi, et al., 2010; Liang & Lee, 2014). Additionally, they use within-subject designs with small sample sizes that limit the generalization of the models to a broader population (Zhang et al., 2004; Kutila et al., 2007; Liang, Reyes, & Lee, 2007; Liang, Lee, & Reyes, 2007; Jin et al., 2012; Liang & Lee, 2014; Tjolleng et al., 2017). Furthermore most approaches incorporate additional driving measures to physiological measures (Zhang et al., 2004; Kutila et al., 2007; Liang, Reyes, & Lee, 2007; Liang, Lee, & Reyes, 2007; Miyaji et al., 2008; Jin et al., 2012; Liang & Lee, 2014). Another general limitation of this kind of research is that the psychological models underlying the experimental data collection differ widely. This can be ascribed to different operational definitions that are applied for each study (see chapter 2).

Subsequently, two recent studies on classification of drivers' cognitive workload that use physiological measures will be presented in-depth. Even though it is hard to compare classification results due to different methodological approaches, the selected studies are suitable for a comparison of the classification task that is at the center of this chapter, since they use a similar task combination as it was done in the second experiment of this thesis.

Tjolleng et al. (2017) published a study that used artificial neural networks (ANNs) to classify three different levels of cognitive workload based on ECG features. For their study 15 male participants drove in a low fidelity driving simulator while performing a cognitively demanding secondary task (N-back) with three levels of task difficulty (0-back, 1-back and 2-back). A within-subjects design was used for data

collection. Each participant performed four consecutive two minute driving trials, one trial for each level of secondary task difficulty and an additional baseline trial without secondary task. As dependent measures three time-domain ECG features and three frequency-domain ECG features were extracted from the ECG raw signal. No further performance or subjective workload measures were collected. For the classification three pre-processing steps were applied individually for each participant. First of all the two most sensitive ECG features were extracted based on the visual inspection of differences between the three task difficulty levels. Secondly the four trials were grouped into three different workload levels according to the similarity of the change in ECG feature and labeled accordingly. When e.g. the baseline and the 0-back trial were most similar to each other, they were grouped to one workload level (low) and the 1-back and 2-back trials were labeled as medium respectively high workload level. Finally, the medians of the two selected ECG measures' were normalized by defining them to take the value 'one'. For the classification, the data of the 15 participants was randomly split into training and test data at a ratio of 70% to 30%. Whether the split was subject-dependent or subject-independent is unfortunately not explicitly mentioned. The design and the number of samples, however, suggest that the split was done subject-dependent. As machine learning method a three layer feed-forward network with back propagation ANN was used. The data split was repeated 100 times to validate the classifier, resulting in a mean classification accuracy of $M = 95\%, SD = 2.77$ for the training data and $M = 82\%, SD = 8.58$ for the test data.

Solovey et al. (2014) conducted two experiments to examine whether robust models can be built across individuals instead of individual models for each participant. In both experiments, the data was collected in real-world driving environments and different machine learning algorithms were compared for the classification task. The features used for the classification task were identical for both experiments. As physiological features, measures of heart rate and skin conductance were incorporated. Additionally several continuous driving performance measures (speed, steering position and acceleration) were included. For the first experiment $N = 26$ participants drove on an interstate highway while performing a cognitively demanding secondary 2-back task (Mehler et al., 2011). Overall the participants completed 24 task periods of 30 seconds length in dual-task conditions, each followed by a 90 second driving-only section. For the classification, they labeled data from the middle of the driving-only periods (24 x 30 seconds) as *normal workload* and the data from the dual-task sections (24 x 30 seconds) as *elevated workload*. Feature vectors were then extracted

by using different sliding windows with different overlapping factors for the calculation of e.g. mean, minimum, maximum values as well as standard deviations. No normalization of the features was applied. Based on the resulting feature matrix they built and tested individual classification models for each participant using a ten-fold cross-validation. Overall the classification with the best performing machine learning algorithm yielded $M = 75.7\%; SD = 10.9\%$ classification accuracy with all features included. Using heart rate alone as feature for the classification resulted in $M = 74.1\%; SD = 12.4\%$ classification accuracy, showing that including performance measures does not improve the accuracy much. In their second experiment they aimed at developing more generalized models. For that reason $N = 146$ (age and gender balanced) participants drove on an interstate highway while performing a cognitively demanding secondary task (N-back) with three levels of task difficulty (0-back, 1-back and 2-back) described in Mehler et al. (2011). Overall they completed three dual-task blocks, one for each task difficulty condition. Each block consisted of four 30 second sequences of the secondary task. Inbetween three dual-task blocks participants got a 150 second recovery period. The construction of the feature matrix was the same as described for the first of their experiments. However, only the 2-back dual-task blocks were used for the *elevated workload* condition. Based on the resulting feature matrix, they built and tested subject-independent classification models using a ten-fold cross-validation. Classification accuracy was highest using heart rate features only ($\sim 90\%$). The classification accuracy was equally high when using all physiological features but deteriorated with the integration of driving performance measures ($\sim 80\%$). Driving measures alone yielded the lowest classification accuracy ($\sim 60\%$).

The two different approaches of classifying drivers' cognitive workload by Tjolleng et al. (2017) and Solovey et al. (2014), resulted in promising classification accuracies. Whereas the focus of the study by Solovey et al. (2014) was to build subject-independent general models, Tjolleng et al. (2017) focused on differentiating between more than two workload levels. Both aims are combined in the following sections by building general classification models based on the data collected in experiment two of this thesis (chapter 5).

6.3 Method

The following section describes the procedure that was used for the development of cognitive workload classifier models built on the basis of the data collected in the

second experiment of this thesis (see chapter 5). The procedure is based on the steps introduced in subsection 6.2.2.

6.3.1 Psychological model

The psychological model used here was the cognitive driver workload model (see chapter 3). As shown in the second experiment of this thesis, three different cognitive workload levels were induced through three different task difficulty conditions. Therefore it can be assumed that the physiological responses are the result of three different psychological states (trinary state). They are defined as *low workload*, *medium workload* and *high workload*. Additionally a binary psychological state distinction was introduced. Data from the low task difficulty condition was labeled as *low workload* and data from the medium and high workload condition were grouped together as *elevated workload*. This was done to examine how well two 'extreme' workload states can be distinguished in comparison with earlier studies that only used two different workload levels.

6.3.2 Feature extraction

Feature extraction from the raw data of ocular measures, cardiovascular measures and speech signal measures are described in subsection 5.2.5. On the basis of past research as well as experiments one and two of this thesis, six features were selected due to their sensitivity to changes in cognitive driver workload as well as practical considerations. However, only the features for which driving-only baseline values were available were considered for the workload classification task. This resulted in four different features: heart rate ($f1$), nose tip temperature ($f2$), pupil size ($f3$), and horizontal fixation dispersion ($f4$). The numbers associated to the features reflect the mean percentage of change of the features compared to driving-only baseline, averaged over all blocks for every participant (see subsection 5.2.6). This resulted in one feature vector for each participant consisting of the four different physiological features combined in a feature matrix with n rows and four columns. To each feature vector a corresponding psychological state label y is available in the target vector (see equation 6.1)

$$\begin{bmatrix} f_{11} & f_{21} & f_{31} & f_{41} \\ f_{12} & f_{22} & f_{32} & f_{42} \\ \vdots & \vdots & \vdots & \vdots \\ f_{1n} & f_{2n} & f_{3n} & f_{4n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{6.1}$$

6.3.3 Normalization and dimension reduction

No normalization was necessary as the range of values was similar for all features. Dimension reduction was also not necessary because the number of features was low anyway. In the context of classifying drivers' cognitive workload it is however of great interest to minimize the effort of measuring features with respect to practical considerations. Whereas it is easy to collect multiple data sources in laboratory settings, real world applications are most likely limited with respect to the availability of different physiological data sources. Therefore the goal is to minimize the number of features, if possible. To accomplish that a scatterplot matrix was created for the combination of the two cases of classification to visually inspect pairwise feature distributions grouped by psychological state (see figure 6.4). This enables the identification of the most promising feature combinations when the goal is to minimize the number of features. With respect to the three-leveled workload cases, it became apparent that the distribution of features originating from the low workload condition seemed to be relatively distinct from those of the medium and high workload conditions (see subfigure 6.2 and 6.4a). Particularly pairs of features including pupil size exhibited this pattern. To some extend the same was found for heart rate, except for the combination of heart rate and nose tip temperature. The distinction between the medium and high workload condition was not particularly clear although the combination of pupil size and heart rate seemed to be the most promising feature combination to distinguish the two conditions. These observations are in line with the results from both experiments of this thesis. Heart rate as well as pupil sizes exhibited higher effect sizes compared to the other physiological measures in statistical testing. For that reason it was decided to use pupil size and heart rate as a feature subset, in addition to the full feature set, to examine the suitability of the combination of these two features for cognitive workload classification. The resulting feature matrix is shwon in equation 6.2.

$$\begin{bmatrix} f_{11} & f_{31} \\ f_{12} & f_{32} \\ \vdots & \vdots \\ f_{1n} & f_{3n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{6.2}$$

6.3.4 Classification

For the classification only feature vectors without missing values were used, resulting in 82 complete feature vectors and their corresponding target state (workload level).

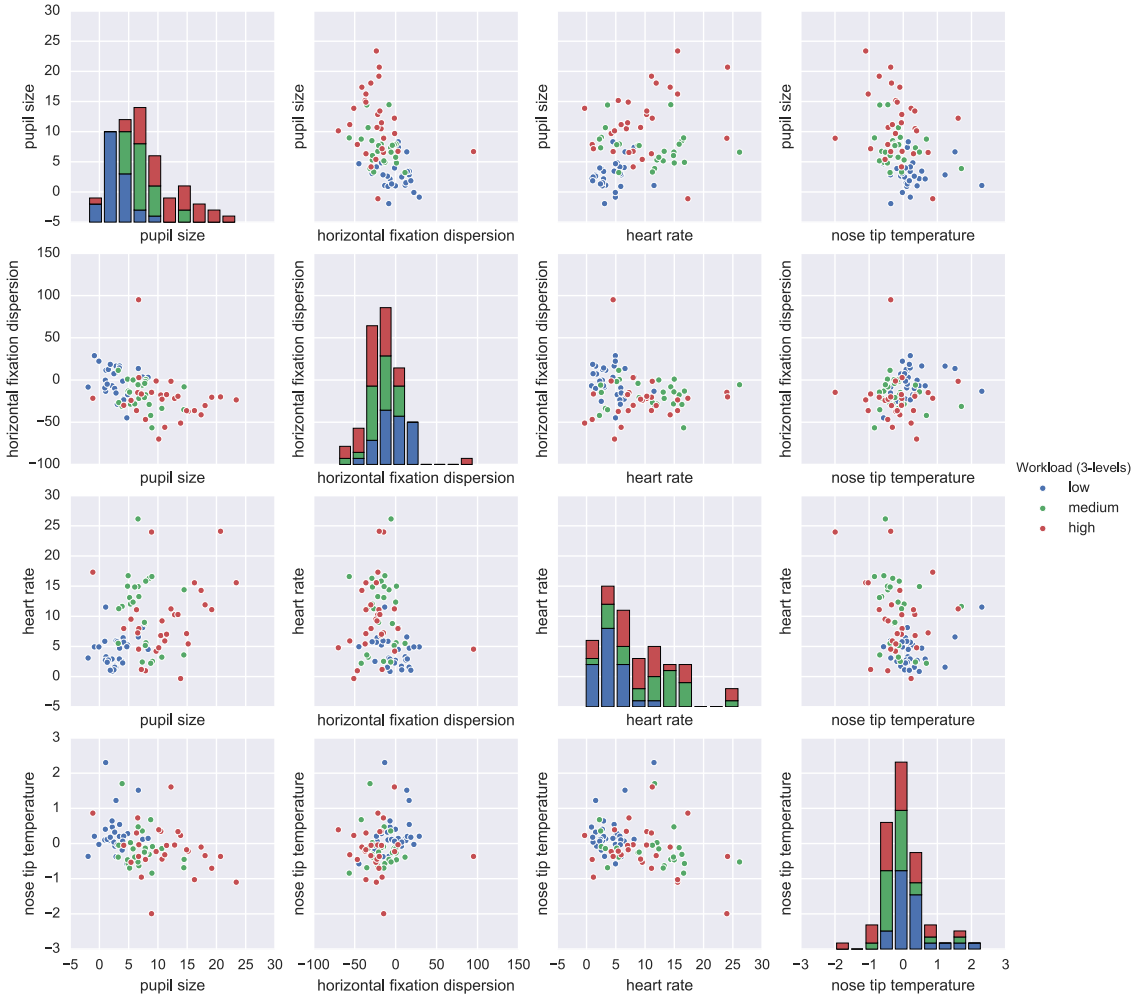


Figure 6.2: Scatterplot matrix for all physiological features and psychological target states - Three workload levels

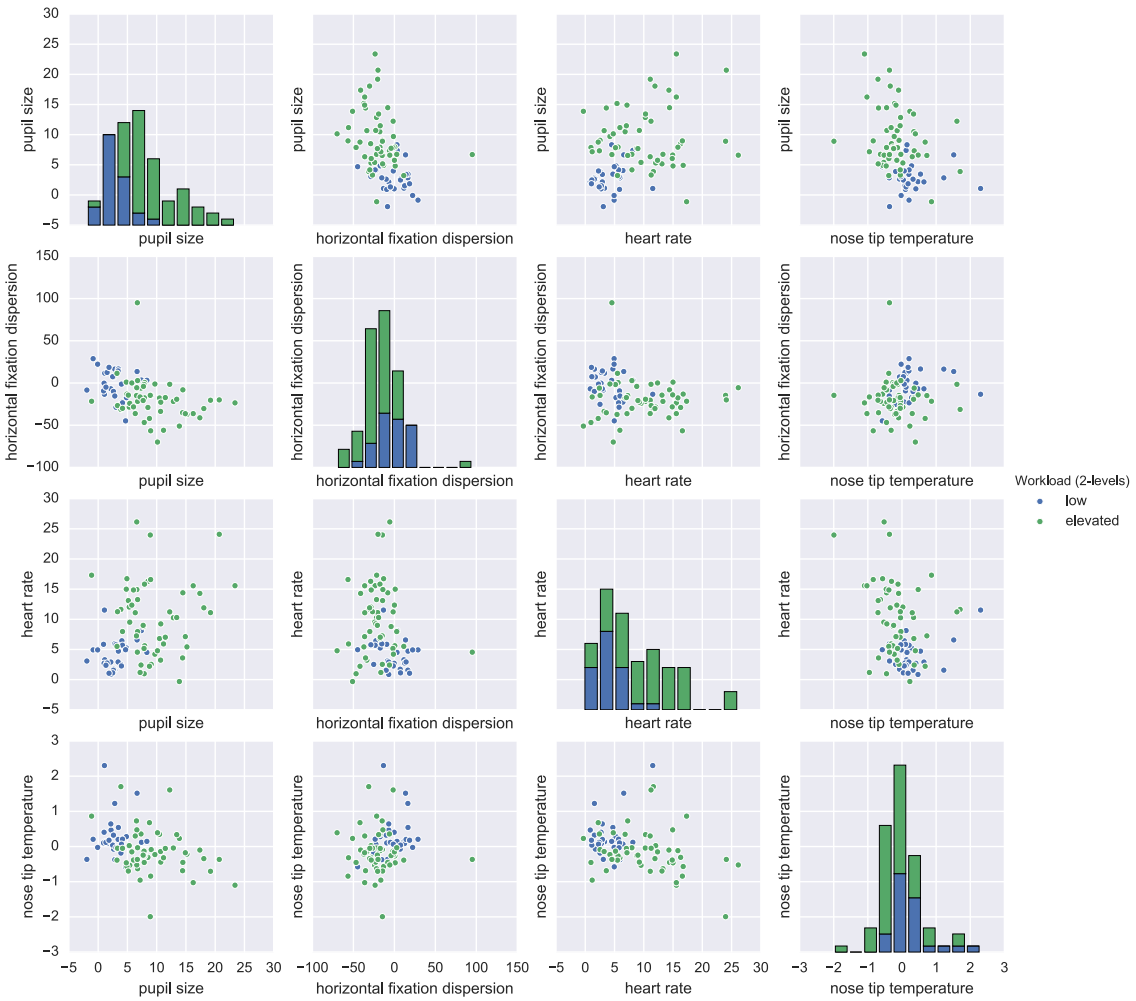
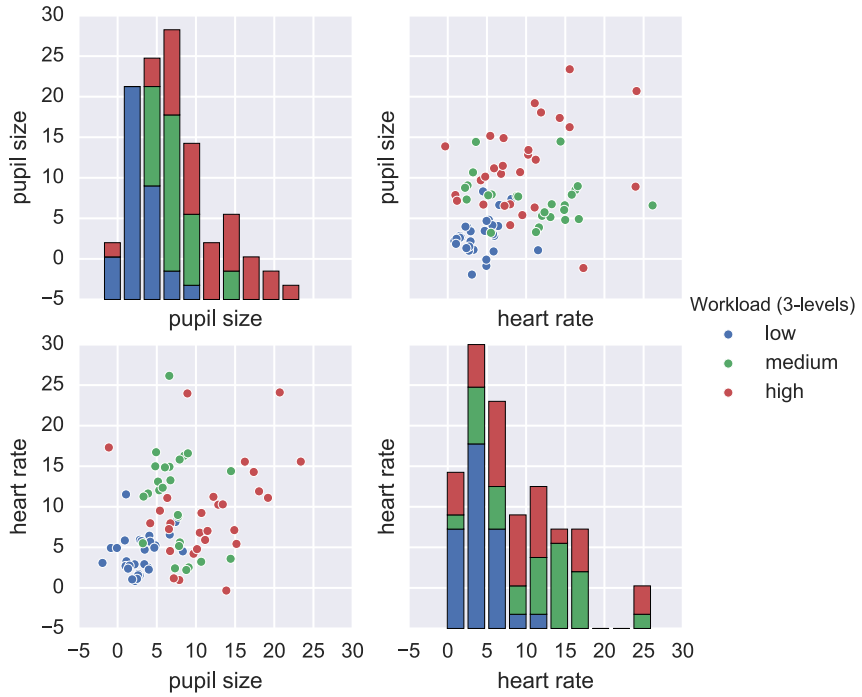
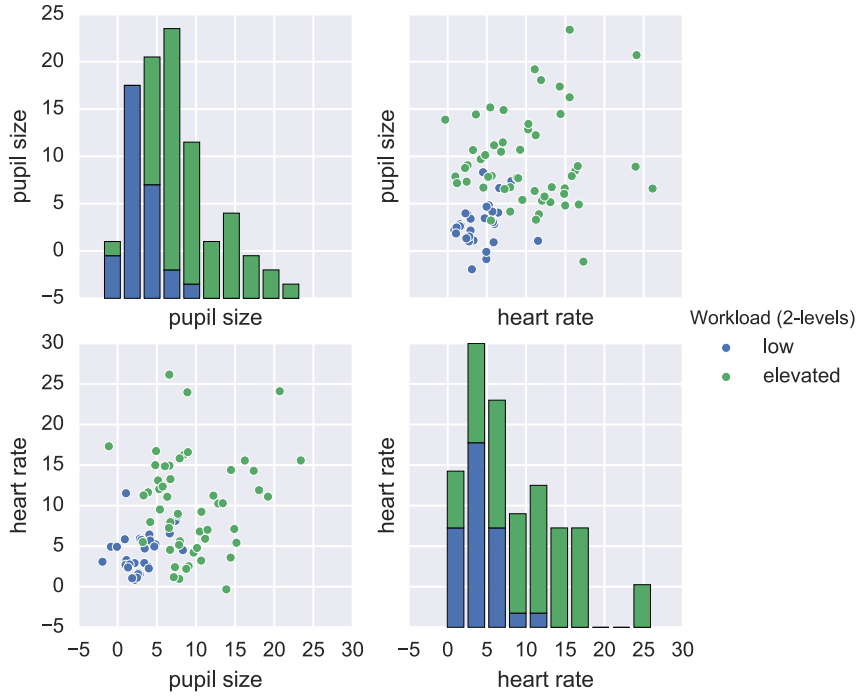


Figure 6.3: Scatterplot matrix for all physiological features and psychological target states - Two workload levels



(a) Three workload levels



(b) Two workload levels

Figure 6.4: Scatterplot matrix for subset of physiological features

With two different target vectors (two vs. three workload states) as well as two different feature sets (four features vs. two features) this resulted in four different models. The full feature set and the subset of features were used for both the classification of the binary as well as the trinary psychological state. All following calculations were accomplished using Python (version 3.5.2) (see appendix C.1). The source code as well as the detailed results can be found in appendix C.2.

The following steps were applied in the process of building classification models for all four cases: first, the data set of 82 feature vectors was split into a training and a test set at a ratio of 80% to 20%. A stratified split was used to ensure the same probabilities of the target states in the training as well as the test set. For the training set an extensive 5-fold cross validated grid search was implemented to find the best values for the hyperparameters for the machine learning algorithms of interest and then tested on the test set. As machine learning algorithms k-nearest neighbors (KNN), support vector machines (SVM) as well as a random forest classifier (RF) were chosen. For each of the three algorithms a grid was created with different combinations of values for the hyperparameters that can be found in appendix C.2. For the SVM as well as the RF method, class weights to balance the impact of different target state frequencies in the data were applied by using the class_weight=balanced option in the corresponding scikit-learn methods (Pedregosa et al., 2011). This was particularly necessary for the two-class workload classification, as the number of feature vectors in the elevated workload condition was almost two times higher than in the low workload condition. The result of this first step was a set of hyperparameter values for each machine learning algorithms that achieved highest classification accuracy. For details on the hyperparameter sets and the corresponding performance metrics please see appendix C.2.

After the best performing sets of hyperparameters were determined, a 10-fold cross validation was carried out with all three machine learning algorithms with their respective 'best' values of the hyperparameters from the grid search regarding the whole data set. Mean values and standard deviations of the classification accuracy can be found in table 6.1. In general the subset of features achieved higher classification results than the full set of features and the binary decision task yielded higher accuracy than the trinary decision task. This results in the lowest classification accuracy for the model with all features and the three-level workload target states and highest for the model with the subset of features and two-level workload target states. For each of the four different cases the different machine learning algorithms yielded almost

Table 6.1: Mean and standard deviation of the cross validation accuracy

	All Features		**Heart rate & Pupil size**	
	Workload 3-levels	Workload 2-levels	Workload 3-levels	Workload 2-levels
KNN	**.68** (.23)	**.88** (.08)	**.76** (.09)	**.90** (.07)
SVM	**.69** (.14)	**.84** (.07)	**.76** (.11)	**.83** (.14)
RF	**.67** (.17)	**.88** (.09)	**.74** (.18)	**.92** (.07)

Note: ***M*** *and (SD)*

similar results, except for the subset of features and the binary decision task, where SVM accuracy was lower compared to the other algorithms.

6.4 Discussion

The aim of this chapter was to combine promising physiological measures that were identified in chapter 4 and chapter 5 for the classification of drivers cognitive workload. Based on the data generated in the second experiment of this thesis, four different workload classification models were built. The results are discussed in this section.

The most complex model, using four physiological features and three different target states (workload levels), yielded a classification accuracy of almost $\sim$ 70%. However, the combination of the four physiological measures performed worse compared to the combination of only two physiological measures (heart rate and pupil size) resulting in classification accuracies of up to 76%. The same effect was found for the binary classification task that yielded $\sim$ 88% with the full feature set and up to 92% with the subset of two features. These differences can be explained when considering the results of experiment one and two as well as past research. Heart rate and pupil size have consistently shown to be sensitive indicators of cognitive workload with a good reliability. As discussed earlier, changes in nose tip temperature are somewhat related to changes in heart rate in response to cognitive workload. Under cognitive workload both measures are predominately sympathetically innervated and coupled through the cardiovascular system (L. Mulder, 1992; Drummond, 1994; Rimm-Kaufman & Kagan, 1996). For that reason it is possible that changes in heart rate (in)directly influence nose tip temperature. Increases in heart rate might have a moderating effect on changes of nose tip temperature. Additionally nose tip temperature exhibits slower responses to changes in workload and is not sensitive to

parasympathetic influences. Therefor it is not surprising that the inclusion of nose tip temperature did not increase differential sensitivity. Horizontal fixation dispersion on the other hand did also not increase classification accuracy. This might be due to the fact that this measure added variability to the feature matrix as compared to the other measures (see table 5.7). As mentioned earlier this is most probably related to inter- and intra-individual of visual attention allocation. For that reason a normalization to equal data ranges for all features might have been beneficial. However, additional models could have been built and statistically compared with each other to test whether the inclusion of either nose tip temperature or horizontal fixation dispersion would yield different results compared to the inclusion of both features. Furthermore, to statistically test different models against each other would have been useful in general and constitutes a limitation of the approach used in this thesis.

In summary it can be stated, that the combination of heart rate and pupil size is a promising approach for the classification of drivers' cognitive workload. Furthermore the combination of heart rate and pupil size comes with an additional practical advantage: both can be measured with remote technological equipment, like low-cost eye-trackers (e.g. The Eye Tribe, Tobii EyeX) and video-based heart rate detection algorithms (see e.g. Poh et al., 2010). It is therefore reasonable to assume that both measurements could be integrated in one low-cost device.

With respect to the classification accuracy, the results are very promising. The models are all subject-independent and the results are well above chance level (binary states: 50% and trinary state: 33%). The models also performed well compared to past studies. Tjolleng et al. (2017) achieved 82% accuracy in a trinary classification task of drivers' cognitive workload. However, the models depend highly on individual pre-processing of the data. Furthermore, the workload levels were assigned individually depending on the descriptive evaluation of physiological reactions without further validation by e.g. subjective measures. Although this approach probably increases accuracy, it limits the generalization of the models to a broader user group and casts doubt upon the use of this approach in real-world settings. Nevertheless using individual characteristics as additional features in workload classification is of great interest as classification accuracy can often be improved by this as cognitive capabilities vary between individuals and can have an influence on physiological traits. Past research has e.g. shown that individuals' working memory capacity impacts cognitive workload (Ross et al., 2014). In their study Ross and colleagues could show that higher working memory capacity was associated with better lane-keeping and event detection performance. This study was replicated by a master student at the

chair of human machine systems at TU Berlin under supervision of the author of this thesis and additionally collected physiological measures (Scholl, 2016). The study replicated the results by Ross et al. (2014) and additionally revealed that participants with higher working memory capacity exhibited higher heart rates by trend compared to participants with lower working memory capacity. This indicates that physiological responses to the same task difficulties can differ individually. Including such attributes like an individual's working memory capacity into classification can be an alternative to hand-picked features and post-hoc assignments to different levels of workload for each driver. Factors like working memory capacity could be evaluated once and then be used for many different models without the need to individually determine the most sensitive physiological features.

With respect to the study by Solovey et al. (2014), the results for the binary classification task are highly comparable. In both cases subject-independent models were built that classified two different workload levels. Solovey and colleagues found that a general model based on heart rate features alone yielded a classification accuracy of $\sim 90\%$ in the discrimination of single (driving-only) vs. dual-tasks (driving + 2-back). The present approach yielded a similar classification accuracy based on pupil size and heart rate but discriminated between two dual-task conditions (driving + 0-back vs. driving + 1-/2-back). In both cases additional features did not increase the classification accuracy, showing that good general models can be built on the basis of few physiological features. However, both studies differed with respect to the implementation of the secondary task and the time intervals used for feature generation. Whereas the secondary task in the study presented in this thesis was a continuous task over almost four minutes per block without breaks, the study by Solovey et al. (2014) implemented the n-back version described in Mehler et al. (2011) over a period of two minutes per block. In the latter implementation a sequence of four separate 10-digit n-back sequences within each two minute block was used, implying that short breaks between each sequence were granted. For a detailed description of the secondary task structure please see Mehler et al. (2011) and Solovey et al. (2014).

Physiological adaptational processes might hence differ significantly over time on task between the two studies. As discussed in subsection 5.4.3, heart rate exhibited noticeable changes over time on task in the second experiment of this thesis. If or to what extend adaptational processes were observed by Solovey et al. (2014) within the 10-digit sequences and/or over the two minute blocks is unclear. Hence the physiological data underlying the classifications might differ between this thesis and the study by Solovey and colleagues.

When comparing the results of this study and Solovey et al. (2014) with respect to the differences in classification accuracy, the choice of time intervals that were used for feature generation should also be considered. Solovey et al. (2014) varied the window sizes for the calculation of the features and found that the longest window size (30 seconds) they tested was the best choice, while smaller window sizes led to lower classification accuracy. This is not surprising since shorter window sizes are more prone to outliers. For the approach of this thesis, features were calculated over four minutes and averaged over three blocks. Therefore the resulting feature vectors in this thesis represent the averaged physiological reactions to the imposed cognitive workload levels, whereas the feature vectors from Solovey et al. (2014) represent multiple physiological reactions to each imposed cognitive workload level. This constitutes a tremendous difference between the approach used in this thesis and the approach by Solovey et al. (2014) with respect to the applicability of the resulting models. Whereas the models from this chapter classify averaged workload, the models from Solovey and colleagues classify momentary workload. Both approaches are based on the assumption, that the workload level is the same for each feature vector in the corresponding workload condition. In this regard the models built in this chapter would presumably perform insufficiently for online classification, as they don't account for momentary changes in cognitive workload. This could be tested by generating features with a method similar to that used by Solovey et al. (2014) and use those features as test vectors for the models that were built in this chapter. However, if one abandons the assumption that the workload levels are stable for each task difficulty condition but instead assumes that they change over time due to changes in task engagement, it would be interesting to examine which model would perform better with regard to online classification. This however involves the even bigger challenge of objectively labeling the momentary workload levels. One solution could be to use the models as input for a workload adaptive task manager and compare the performance outcome as well as subjective evaluations of this systems.

In summary, the approach that was presented in this chapter resulted in promising models for the classification of drivers' cognitive workload. It should however be noted that the specific procedure that was used here, represents only one possible solution among many valid alternatives and is limited in its applicability. Therefore the data set that was generated from the second experiment of this thesis could be used to explore other features, pre-processing methods and machine learning algorithms. Additionally, it would be interesting to define different target states with respect to the assumption made that participants in the low task difficulty condition were in

underload and the participants in the high task difficulty in overload workload regions (sections 5.3 and 5.4), to identify optimal workload levels.

6.5 Conclusions

Overall it can be concluded that driver workload classification can be accomplished by using methods of physiological computing. Using a combination of physiological measures to build subject-independent models that can be generalized proved to be a promising approach yielding high classification accuracies for binary as well as trinary workload target states. Furthermore it was shown that using a combination of only two easy-to-measure features (heart rate and pupil size) are sufficient in this context. However, the models that were built are limited in their ability to be applied in real world driving, as they are not ideally suited for online workload classifications. However, they could be used to compare other cognitive secondary tasks or speech interaction applications with the workload levels that were induced by the n-back task when evaluated in the same experimental setting and design as described in section 5.2. Other methodological approaches for the model building process should be explored using the data that was collected in the scope of the second experiment of this thesis. Therefore all steps in the workflow of physiological computing should be considered.

Chapter 7

Discussing cognitive workload in the context of drivers' state classification

7.1 Accomplishments of the aims of the thesis

In the previous chapters two experiments were presented examining the sensitivity and validity of physiological measures to multiple levels of cognitive driver workload. Models for driver cognitive workload classification were successfully built on the basis of the physiological measures that were identified. The discussion in each chapter mostly covered specific aspects of the two experiments as well as of the classification. The following sections will now discuss more general observations made across all three chapters (4, 5 and 6) with respect to the aims formulated in the introduction of this thesis.

7.1.1 Reviewing the concept of drivers' cognitive workload

The first aim of this thesis was to review the concept of drivers' cognitive workload. It is striking that there is still no definition of workload that is commonly agreed upon and confusion with other related concepts like distraction is abound. Even though it is desirable to solve this problem it is unlikely that this happens any time soon due to the fuzzy nature of workload and other underlying concepts. Therefore it is of utmost importance when using operational definitions of workload to extensively describe them to enable other researchers to interpret the results accordingly.

Furthermore performance, subjective and physiological consequences of cognitive workload were examined together and in relation to each other as they represent different aspects of workload that need to be considered when interpreting cognitive

workload resulting from a specific secondary-task. Particularly the 'correct' interpretation of performance and subjective measures to determine the cognitive workload state of an operator is a prerequisite to understand physiological reactions, which in turn is necessary to build reliable models for workload classification.

With respect to overall performance in dual-task settings, both primary- as well as secondary- task performance have to be considered. When the primary-task (driving) is instructed to be the main focus of the driver, it is expected that secondary-task performance reflects the change in attentional demands. Therefore secondary-task performance can serve as a workload indicator. In both experiments secondary-task performance exhibited differential sensitivity to the three different task difficulties and is therefore a very useful component for the overall workload assessment. However, in both experiments it would have been beneficial to evaluate performance regarding the secondary-task in a single-task setting to get more detailed insights into the changes of overall performance, thus into changes of attentional resource demands in dual-task conditions. From an applied point of view secondary-task performance is probably of no relevance as it is almost impossible to operationalize in real-world scenarios.

In experiment 1+2 consequences of auditory-verbal secondary-tasks on performance in part confirmed past research but also revealed some important aspects of changes in driving performance that need to be considered in the future. In general measures of continuous driving performance like lane-keeping and speed only seem to be minimally affected by cognitive workload on an absolute level which is in line with past research (see subsection 3.4.1). However, whether lane-keeping generally improves or worsens can not simply be inferred by the nature of the secondary-task but is also dependent on the primary-task. Whereas in the first experiment lane-keeping worsened with increasing secondary-task difficulty, it improved in the second experiment due to different driving task demands. This underlines that conclusions regarding consequences of cognitive workload cannot simply be drawn by analyzing secondary-task demands but also need to consider the specific demands put on the driver from the driving task. Only if the attentional demands of primary- and secondary-tasks are evaluated together, conclusions about the resulting workload can be drawn. With this in mind it is not surprising that studies on common secondary-tasks e.g. hands-free telephoning sometimes report contradictory results regarding driving performance as well as crash risk.

Event-related driving measures in the second experiment did not show the expected pattern. Reaction times, omitted break reactions and collisions did not linearly increase as a function of task difficulty like lane-keeping measures but showed

benefits in the medium task difficulty compared to the other task difficulties. First of all these results indicate that under special circumstances performing a medium difficulty auditory-verbal secondary-task in addition to driving can be beneficial with respect to some driving measures when compared to lower or higher task difficulties. Secondly this implies that optimal performance regarding event-related measures was reached for a different task difficulty condition as for e.g. lane-keeping variability that reached its 'optimum' at the lowest task difficulty condition. With respect to the interpretation of changes in performance the results reveal a general problem. In workload literature it is assumed that workload is optimal when performance is optimal (see subsection 2.1.4). To decide on the level of workload more than one driving performance measure has to be taken into account. However, when using event-related as well as continuous driving measures like in the second experiment of this thesis, sometimes the results from different measures diverge which can be ascribed to different attentional resources involved for continuous vs. event-related vehicle control measures (see chapter 4 3 for details). As a result it is hard to decide whether driving performance is optimal or not. Optimal performance must therefore be defined for each task combination with respect to the desired outcome. It might for example be tolerable to allow for little decreases in lane-keeping performance but even more crucial that event-related performance is best when it comes to safety related optimal performance. The discussion on performance measures of cognitive driver workload so far reveals that the use of complementary non-performance measures is often required to differentiate between different workload levels. For that reason both experiments in this thesis additionally collected subjective workload measures to infer if the experimental manipulations resulted in three different workload levels as intended. It has been shown in both cases that this approach was beneficial. Particularly the interpretation of the single dimensions of the NASA-TLX allowed for a more holistic interpretation of the induced workload levels with respect to the contribution of different aspects of workload to the overall workload ratings.

7.1.2 Sensitivity and reliability of physiological measures

The second aim of the thesis was to find suitable physiological measures to be integrated in multi-measure approaches of drivers' cognitive workload. Therefore traditional cardiovascular and ocular measures were combined with facial temperature features and speech signal measures. The selection of the physiological measures for the first experiment was primarily driven by their capability to be collected non-intrusively. Secondly they were chosen with respect to their sensitivity to changes

in cognitive workload. There are other psychological measures that are potentially more sensitive to cognitive workload than the ones chosen. Measurement methods like Electroencephalography (EEG) or Functional Near-Infrared Spectroscopy (FNIRS), directly assessing brain activity instead of ANS activity, have the potential to be advantageous regarding sensitivity but are less feasible for applied settings since the methods require a physical connection to the human body (see e.g. Ayaz et al., 2012; Borghini et al., 2014; Peck, Afergan, Yuksel, Lalooses, & Jacob, 2014). Additionally there are other non-intrusive measures that could have been taken into account as physiological workload measures. The use of modern thermography cameras for example allow to additionally record measures that have shown to be sensitive to workload and effort like e.g. perspiration at the perinasal area (Shastri et al., 2009; Pavlidis et al., 2016) or breathing patterns by analyzing air streams from the nose (Fairclough & Mulder, 2011). Additionally, other facial temperature features like forehead temperature or temperature in periorbital regions could have been included, too (Or & Duffy, 2007; Reyes et al., 2009). The measures examined in this thesis only represent a selection of possible physiological workload indicators that, according to the literature, appear to be the most promising. Most of the selected physiological measures exhibited differential sensitivity to different cognitive workload levels in the first experiment. However, only pupil size differentiated between the medium and the high cognitive workload condition. Nose tip temperature, heart rate, as well as fundamental frequency of the voice and voice intensity did not differentiate between those two workload levels indicating that their sensitivity to differences in the attentional demands is limited. Whereas pupil size seems to be more sensitive to differences in parasympathetic and sympathetic changes induced by the different secondary-task demands, the other measures predominantly reflect sympathetic changes in response to the task demands.

The use of eye-movement measures revealed a major dependence on dynamic aspects of the task demands because they are highly related to drivers' strategic visual resource allocation. Eye-movements are primarily driven by visual aspects of the task, therefore the large variations of fixation duration and count can be accounted for by the assumption that strategies to handle the overall task demands probably change over the course of time. This means fixation count and duration is not feasible when cognitive workload is measured and averaged over prolonged time intervals. When used over short time periods they might however be useful as predictors for performance decrements as indicated by the results of Tsai et al. (2007). Horizontal fixation dispersion showed the same sensitivity as e.g. heart rate to the increasing

workload levels but was contradictory to past research in its direction. Horizontal fixation dispersion increased with increasing workload instead of decreasing. This reflects in part the problems that eye-movements are highly task dependent. In the case of the first experiment, gaze dispersion was highly influenced by the artificial nature of the LCT where most driving relevant features for visual attention allocation appeared on the left and the right of the driving course instead of on the center of the road. Nevertheless, horizontal fixation dispersion was included in the validation study as large amounts of research consistently reported gaze concentration with cognitive secondary-tasks in more realistic scenarios. Research on speech signal features for cognitive driver workload measurement is surprisingly limited so far. The results from the first experiment however revealed that fundamental frequency of the voice and voice intensity can very well be used to discriminate between different cognitive workload levels. Although the effect sizes are smaller than that of heart rate and pupil size, the results are promising. Speech interaction systems inherently analyze the speech signal anyway, so there is no additional effort needed to measure it. It has however to be mentioned that even though accuracy of speech signal analysis has improved tremendously over the last ten years, it remains the most complex signal to be analyzed as compared to the other measures used in this thesis. Standards have to be implemented for the calculation of fundamental frequency and voice intensity in the context of workload measurement to make results comparable. So far there are multiple differing algorithms for the calculation of the aforementioned measures depending on the quality of the speech signal and the software that was used. Best practice guidelines for the analysis of speech signals in the context of psychological state assessment like they already exist for e.g. ECG (see e.g. L. Mulder, 1992) would be beneficial.

Overall the aim of finding physiological measures that are sensitive to multiple levels of drivers' cognitive workload that can be used for multi-measure approaches and that are non-intrusive was accomplished. More traditional measures like heart rate and pupil size as well as measures that are so far not yet thoroughly researched like speech signal and facial temperature features revealed a general sensitivity to different levels of cognitive workload. As only pupil size proved to differentiate 'perfectly' between different levels of cognitive workload, the results also underline the need for combined multi-measure approaches for cognitive workload classification. Additionally, eye-movement measures revealed its limits for the measurement of cognitive workload over prolonged time periods. Nevertheless they are promising in situations where dynamic workload assessment over shorter time periods is necessary.

The second experiment addressed several limitations from the first experiment and aimed at validating the identified physiological measures in a more applied scenario. Prior research regarding pupil size, heart rate, nosetip temperature and fundamental frequency could be replicated with respect to the general and differential sensitivity to multiple levels of cognitive workload, indicating a good reliability of those measures. Voice intensity even exhibited better differential sensitivity than in the first experiment. Furthermore, fixation dispersion decreased with increasing cognitive workload, which is consistent with past research but opposite to the results from the first experiment supporting the assumption that demands of the driving task highly influence eye-movement related measures. As the driving scenario used for the second experiment was much more naturalistic compared to the first driving scenario, results on fixation dispersion from the second experiment are more representative and hold the potential to be used in real-world driving applications.

Time series analysis of the progression of the different physiological measures over time tremendously increased insights into adaptational workload processes for all measures and is therefore very useful for the interpretation of the results. Whereas the statistical analyses only looked at mean values of the different workload measures, the use of time series allows for the interpretation of instantaneous, peak and accumulated workload (Xie & Salvendy, 2000a, 2000b). Therefore in future studies it would be interesting to look at those different aspects of workload in more depth when dynamic changes of workload are of interest.

Methodological limitations of experiment 1 & 2

To evaluate the methodological quality of empirical research four different criteria have to be considered (Cook, Campbell, & Day, 1979; Bortz & Döring, 2007). The following subsection will discuss construct validity, statistical validity as well as internal and external validity of the two experiments described in chapter 4 and 5.

Construct validity

In experimental settings the construct validity is mainly determined by the appropriateness and operationalization of the dependent and independent variables (Bortz & Döring, 2007). In both experiments the independent variable of task difficulty was operationalized by using cognitive tasks that have been extensively studied in various fields of research (Kirchner, 1958; Gronwall, 1977; Jaeggi et al., 2010). The different levels of the PASAT as well as the n-back have consistently shown to impose different

cognitive task demands on humans and to induce different cognitive workload levels (e.g. Tsai et al., 2007; Horrey, Lesch, & Garabet, 2008; Mehler et al., 2009; Lenneman & Backs, 2009; Mehler et al., 2012).

Furthermore, the operationalization of the different workload measures is based on extensive literature review and implemented dependent variables that have been shown to be valid indicators of the respective constructs to be measured. Where possible, the impact of individual influences was minimized by the calculation of differences between individual driving-only baselines and driving with secondary-task segments. Moreover, the overall workload assessment was accomplished by integrating performance, subjective as well as physiological measures. Therefore, overall construct validity can be assumed to be high.

Statistical validity

Statistical validity can be assumed to be high when the choice of measurement tools and statistical analysis is appropriate (Bortz & Döring, 2007). An alpha error level of p = .05 was chosen for statistical analysis. This corresponds to the general scientific conventions and reduces the likelihood of the occurrence of type I and type II errors. For discarding alternative hypotheses in favor of the null hypothesis, a robust significance level of $p = .2$ was used to reduce the probability of type II errors (Bortz, 2005). For testing main and interaction effects common ANOVAs were used. ANOVAs come with advantage that they are relatively robust against violations of the distribution assumptions (Bortz, 2005). Nevertheless, transformation of variables, corrections in case of sphericity violations or non-parametric tests were applied and mentioned accordingly when necessary. Since the test of assumptions of e.g. normality involves subjective visual inspection of distributions it cannot be ruled out that in rare cases other researchers would have come to a different conclusions regarding the violations of assumptions. Finally, post-hoc comparisons always used adjustment methods to cope with cumulating chance of commiting type I errors. In summary, statistical validity can therefore be assumed to be high.

Internal validity

Internal validity is high when changes in the dependent variables can conclusively be attributed to the impact of variations of the independent variables (Bortz & Döring, 2007). Generally speaking the internal validity of both experiments can be describes as high. For both experiments the respective experimental procedure was the same

for all participants and the experiments were conducted in the same location under almost the same conditions. Furthermore most of the instructions were standardized and in written form whereever possible. Additionally, a standard was agreed upon for necessary general oral instructions to avoid experimenter related influences. Participants were asked not to talk about the experiment with potential participants. One limiting factor of internal validity, however, is that simulation sickness was not formally assessed in either studies. Therefore in the first experiment that used a within-design, it is possible that for some participants the physiological results were influenced by simulation sickness side-effects. Due to the randomized assignment of the participants to the experimental conditions in the second experiment (between-design), a systematic effect on the physiological results is unlikely. Additionally, in both experiments the participants were explicitly instructed to report any signs of simulation sickness related symptoms at any time during the experiment.

Another limiting factor in both experiments could be that it was not controlled for the time of day and therefore the possibility confounding influences on physiological measures due to different phases in the circadian cycle of the participants can not be excluded. However, for both experiments experimental sessions were scheduled at the same time slots for each day of testing. For the first experiment it can be assumed that participants were distributed roughly equally over the different time slots, hence no systematic impact of time of day is expected. For the second experiment a systematic influence is also unlikely as the distribution of time slots per experimental condition was similar.

External validity

External validity is established, when the results from the experiment can be generalized to different samples, times and situations (Bortz & Döring, 2007). The experimental setting in the first study is clearly limited regarding its external validity. The data was collected in a laboratory setting using a low-fidelity driving simulator. However, for the purpose of testing the sensitivity of multiple physiological measures to different cognitive workload levels it can be assumed to be sufficient. The LCT as well as the PASAT are validated research instruments in the context of driver workload assessment mimicking the attentional resources involved in speech interaction while driving. Nevertheless, due to the influence of the artificial nature of the driving task on eye-movement related measures, the generalization of the fixation dispersion results is clearly limited.

The second experiment used a more realistic driving task as compared to the first experiment, mimicking real-world driving situations (see subsection 5.2.1). Additionally, the secondary-task was adapted because of some possible limitations of the PASAT (described in subsection 4.5). The projection of the driving scene was closer to real-world driving in the second experiment, covering almost 180° of visual angle (see subsection 5.2.1). Therefore it can be assumed that the possibility to generalize the results to driving of the second experiment is higher compared to the first experiment. Both experiments were simulator studies in a laboratory setting. Even though this allows for tighter control of the experiment, it limits the transfer of the results to real-world driving. Nevertheless, meta analytic studies on workload related performance decrements by Horrey and Wickens (2006) and Caird et al. (2008) revealed no differences between simulator and real-world driving studies. Additionally, Reimer and Mehler (2011) could show that results regarding physiological measures in simulator studies are highly comparable to the results in real-world driving conditions.

The sample in both experiments mainly consisted of young and healthy students with only limited driving experience. Therefore, the possibility to generalize the results to the broader population is limited and should be tested in a comparative study with different age groups and educational levels.

7.1.3 Combination of physiological measures

The fourth aim of this thesis was to combine the most promising physiological measures using machine learning algorithms. Overall the classification models performed well. However, whereas the classification accuracy for the binary workload decision task was high, the accuracy for the trinary workload decision task needs to be improved to use the models in applied scenarios.

First of all, the models represent the classification of workload on the basis of averaged features over prolonged time spans for each participant and therefore the averaged physiological reactions to the defined workload states. However, the workload target states are defined on the basis of the holistic assessment of the mean performance and subjective changes averaged for each task difficulty. Therefore, the target states are defined based on the grand mean over all participants for each task difficulty condition but the feature vectors represent the individual's physiological reaction to the different task difficulties. This means that an individual's physiological reaction to the same defined workload level can be different to the physiological reaction from another participant. For example could one person's physiological reaction to the medium workload condition be similar to another's physiological reaction to

the high workload condition. This could simply be due to individual differences in attentional capacity or effort invested in the task. Taken that into account it is understandable that the distinction between the medium and high cognitive workload levels is challenging. Subject-dependent feature selection and target state definitions could be used to avoid this or models for each individual could be built. This however would limit the possibility to generalize the models to a broader population. So depending on the application of the classification models different strategies have to be applied. For this thesis, the aim was to test whether generalizable models for cognitive workload classification can be built based on the combination of multiple physiological measures. By using subject-independent training and test sets as well as 10-fold cross validation, the models can be assumed to generalize well at least for the researched sample. Nevertheless, overall bigger and more diverse sample sizes would have been beneficial to generalize the results to a broader population. Additionally, the collected data set holds the potential to be used for different purposes. Regression and classification models could be built to e.g. predict reactions to collisions for critical events. This would most probably involve smaller time windows for feature generation as well as including eye-movement related features with information about visual attention allocation.

7.2 Transfer to real-world driving

The experiments of this thesis and the classification models built clearly showed that it is possible to detect different cognitive workload levels by using physiological measures. The transfer of those results to real-world driving however holds several challenges. A prerequisite for the real-world use of models for workload classification based on physiological measures is the collection of robust data outside of highly controlled laboratory environments (Jacucci, Fairclough, & Solovey, 2015). Only if the data is reliable, the classification models are brought to their full use. If those models perform badly due to unreliable data, the acceptance of adaptive systems will suffer. Therefore, the influence of environmental influences like differing luminance levels etc. on video-based measures and online outlier detection represents a major technological challenge. Assuming that this problem can be solved, a second even bigger challenge is to delineate different psychological states from each other. As mentioned in sections 3.6 and 2.2.3 physiological measures are suspect to changes due to different emotional and cognitive states. Therefore the same psychological reactions can be a result of different psychological states. Assuming that physiological

measures like heart rate and pupil size are more general indicators of ANS related changes, it becomes apparent that, in order to increase diagnosticity, additional steps have to be considered. One approach to that problem could be the development of models that differentiate between different psychological states. Therefore e.g. adding features related to the activity of the driver to differentiate between different kinds of mental workload, the inclusion of facial expressions analysis to detect the current emotional states, or fatique related features like blink frequency should be considered. To accomplish both aforementioned challenges, it is necessary to massively collect data in real-world driving to build robust, reliable and diagnostic psychological state classification models.

The next challenge that has to be dealt with is the implementation of countermeasures to dynamically adapt to suboptimal psychological states. In the context of cognitive driver workload, task managers could be implemented that adapt the task load of a driver to optimal levels. In its mildest form warnings could be presented to inform drivers about their suboptimal psychological states encouraging them to manage their task load themselves. This however might even increase the momentary workload of the driver through additional information. Alternatively, task demands could be lowered implicitly by restricting the use of specific infotainment functions or the ADAS could overtake certain aspects of driving to lower the overall task load.

For the last couple of decades humans explicitly interacted with technology and adapted to technical system and not the other way around. Even though this is changing rapidly right know due to big data and artificial intelligence, there are major challenges to designing forms of interaction, with humans in the loop that are no longer capable of explicitly predicting the 'behavior' of the technological system (Böhle, Coenen, Decker, & Rader, 2013). Therefore topics like who is ultimately in control of the adaptive system and to what extend the driver is aware of how the system works, are central aspects (Fairclough, 2010). Additionally, the collection of physiological data and the manipulation of psychological states have big implications with respect to privacy. Psychological data has to be understood as highly sensitive data. Even when collected for purposes of e.g. workload adaption, psychological features like pupil size and heart rate could be used to assess the driver's health (Böhle et al., 2013) or even for identification purposes (e.g. Biel, Pettersson, Philipson, & Wide, 2001; Bednarik, Kinnunen, Mihaila, & Fränti, 2005). For this reason regulations have to be implemented to assure that drivers have full transparency on where and how their data is stored, how it is used, who has access to it and that they have full control over their data (Fairclough, 2009, 2014).

Chapter 8

Conclusions and Outlook

The last chapter will summarize the conclusions of this thesis and give an outlook on research that could be conducted in the future based on the results presented in this thesis.

8.1 Conclusions

Driving has changed massively over the past decades and will continue to change in the future due to new assistance systems, increasing availability of information and communication technology as well as new interaction modalities. Speech interaction as an alternative to traditional interaction techniques is already standard in modern cars and might become the main interaction scenario in cars in the future. Therefore, the overall aim of the thesis was to examine the cognitive workload implications of auditory-verbal secondary-tasks and provide methods to measure drivers' cognitive workload non-intrusively as that potentially prevents dangerous situations before they appear. To this end non-intrusive physiological cognitive workload measures were successfully identified and validated in two experiments and afterwards combined to build models of workload classification. Both experiments considered performance and subjective measures in addition to physiological measures for the holistic assessment of the cognitive workload induced by auditory-verbal secondary-tasks with three difficulties.

The first experiment revealed that besides traditional workload measures like pupil size and heart rate, speech signal features and facial temperature features are sensitive to multiple cognitive workload levels induced by auditory-verbal secondary-tasks. However, pupil size exhibited best differential sensitivity which is consistent with past research. Eye-movement related measures like fixation count and duration as well as horizontal fixation variability also proved to be sensitive to different cognitive

workload levels but are limited in their generalizabilty, since they are highly dependent on the visual dynamics of the driving task. Therefore fixation duration and count were excluded as general indicators of cognitive driver workload for this thesis but should nevertheless be considered as measures of dynamic workload changes over time.

The second experiment aimed at validating the results found in the first experiment in a more realistic driving scenario that included potentially hazardous events. With respect to different aspects of driving, it became apparent that performance consequences of cognitive workload can diverge regarding continuous and event-related indicators and should be considered carefully when interpreting the workload of a driver. This will be important when determining optimal workload levels. Most of the results regarding physiological measures replicated the results found in the first experiment and therefore underlined the potential for their use as cognitive workload measures but also emphasized that the combination of several measures are useful for a reliable estimation of cognitive workload.

Finally, the data from the second experiment was used to build generalizable models of cognitive workload classification based on the most promising physiological measures. Results validated the assumption that high classification accuracies can be achieved when two very different workload levels are distinguished. Classification of all three workload levels however resulted in lower classification accuracies and should be improved in future research. Furthermore, it could be shown that using only heart rate and pupil size resulted in better performing models compared to models that additionally use nose tip temperature and horizontal fixation dispersion suggesting that more is not always better. These results are very promising, since both measures can be collected contactless with low-cost technological equipment.

Overall, the results of this thesis added insight to the current state of science on drivers' cognitive workload. Particularly with respect to consequences of drivers' cognitive workload regarding performance and physiology as well as classification of multiple cognitive workload levels, this thesis contributes significantly to the understanding of drivers' cognitive workload. However, several limitations have been discussed and the transfer of these results to the real-world of driving faces multiple challenges that have to be addressed before reliable workload detection and adaption systems can be implemented into cars. The last section of this thesis therefore describes possible areas for future research in the area of drivers' cognitive workload.

8.2 Outlook

In this thesis a selection of physiological measures were examined regarding their sensitivity to multiple cognitive workload levels. However this selection of measures, and the extraction of features from those measures, is far from complete. Future research could therefore explore a variety of directions. First of all physiological measures that can be collected with modern thermography cameras are of great interest. Thermal imaging offers the possibility to detect several physiological processes at once. Besides facial temperature features, which were in part covered in this thesis, other processes like respiration could be analyzed regarding their suitability as cognitive workload measures. Secondly, this thesis only explored two different features of the speech signal. Future research could be conducted on additional speech signal features known to be usefull for emotion recognition (for a review see e.g. Ververidis & Kotropoulos, 2006) and test them with respect to their sensitivity to cognitive workload.

For the second experiment of this thesis multiple data sources were collected for the classification of cognitive workload levels. The modelling approaches were however far from exhaustive. For this reason the generated data set could be used for a variety of different analysis. Future research could explore different feature extraction methods, normalization approaches, machine learning algorithms and learning strategies with respect to their classification accuracy. Additionally, the data can be used for the prediction of reactions to the critical events. For this reason, the author of this thesis intends to make the data set available to other researchers.

On a broader scale, physiological computing faces a variety of challenges. Within this thesis the system adaptation to the psychological state was not covered. However, this part of physiological computing is equally important as the detection of physiological states. Several challenges with respect to the design of the interaction, acceptance of the drivers as well as data privacy and ethics have to be mastered. Designing implicit interactions between technological systems and humans represents a huge change from the traditional explicit interaction paradigm. Even though this form of interaction is penetrating our daily lives more and more, it is still relatively new and needs to be researched to the same extend as explicit interaction has been so far. In this context, future research should also explore other factors that might be relevant to the design of state-aware technological systems. On a broader scale it might also be useful to incorporate information about other traffic participants and the traffic infrastructure for the system adaptation through car-to-car and car-to-x technologies. This possibly allows for adaptation strategies that are not only

beneficial for the individual but for the whole traffic ecosystem.

Another important direction of research is the transfer of the results to real-world driving. To accomplish this, future research has to collect data in real traffic environments with real speech interaction applications. Real-world driving environments might impose qualitatively and quantitatively different attentional demands on drivers as compared to driving simulator settings in this thesis. The same is true for actual speech interaction systems. Whereas in this thesis speech interaction was modeled through cognitive tasks that completely relied on auditory-verbal resources for input and output, speech interactions in real life would most probably involve some sort of visual feedback as response to speech inputs. Therefore, it should be researched whether these potential differences in task demands result in different physiological responses. Furthermore, workload adaptive systems have to be built and tested with respect to performance and subjective workload consequences but also with respect to usability and user-experience of the drivers. Only then the models can be fine-tuned to the needs of the drivers and allow applications of physiological computing to reach their greatest possible benefit.

The highlighted path of potential research in this section only represents a selection of possibilities. With the introduction of semi-autonomic and autonomic vehicles on the horizon, additional research needs to be conducted to continuously improve drivers' safety.

Appendix A

Experiment 1

A.1 Documents

A.1.1 NASA-TLX - German version

VP-Nummer:

Fragebogen NASA TLX

Beanspruchungsstruktur

Geben Sie bitte an, welche relative Bedeutung für die empfundene Gesamtbeanspruchung bei der eben durchgeführten Aufgabe die sechs Beanspruchungsdimensionen

- Geistige Anforderung
- Körperliche Anforderung
- Zeitliche Anforderung
- Aufgabenausführung
- Anstrengung und
- Frustration

für Sie hatten. Lesen Sie dazu bitte zunächst die folgenden Erläuterungen:

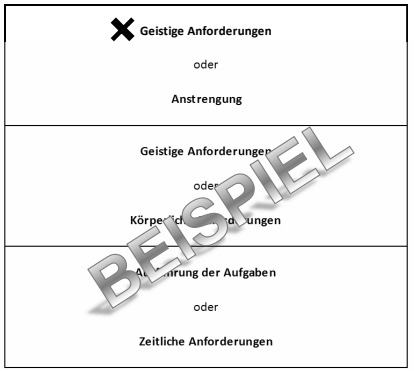
Geistige Anforderungen	Wie viel geistige Anstrengung war bei der Informationsaufnahme und bei der Informationsverarbeitung erforderlich (z.B. Denken, Entscheiden, Rechnen, Erinnern, Hinsehen, Suchen ...)? War die Aufgabe leicht oder anspruchsvoll, einfach oder komplex, erfordert sie hohe Genauigkeit oder ist sie fehlertolerant?
Körperliche Anforderungen	Wie viel körperliche Aktivität war erforderlich (z.B. ziehen, drücken, drehen, steuern, aktivieren ...)? War die Aufgabe leicht oder schwer, einfach oder anstrengend, erholsam oder mühselig?
Zeitliche Anforderungen	Wie viel Zeitdruck empfanden Sie hinsichtlich der Häufigkeit oder dem Takt mit dem Aufgaben oder Aufgabenelemente auftraten? War die Abfolge langsam und geruhsam oder schnell und hektisch?
Ausführung der Aufgaben	Wie erfolgreich haben Sie ihrer Meinung nach die vom Versuchsleiter (oder Ihnen selbst) gesetzten Ziele erreicht? Wie zufrieden waren Sie mit Ihrer Leistung bei der Verfolgung dieser Ziele?

VP-Nummer:

Anstrengung	Wie hart mussten Sie arbeiten, um Ihren Grad an Aufgabenerfüllung zu erreichen?
Frustration	Wie unsicher, entmutigt, irritiert, gestresst und verärgert (versus sicher, bestätigt, zufrieden, entspannt und zufrieden mit sich selbst) fühlten Sie sich während der Aufgabe?

Im folgenden werden jeweils zwei der sechs Beanspruchungsdimensionen in verschiedenen Kombinationen gegenübergestellt. Geben Sie jeweils an, welche Beanspruchungsdimension für die Gesamtbeanspruchung, die Sie empfunden haben, bedeutsamer war. Es geht also zunächst nicht darum, wie hoch die Beanspruchung in den einzelnen Dimensionen war, sondern wie wichtig die jeweilige Dimension für das Gesamtempfinden war!

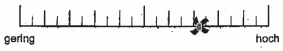
Beispiel: Wenn für Sie die geistigen Anforderungen, die die Aufgabe gestellt hat, bedeutsamer für das Beanspruchungserleben waren, als die Anstrengung, die Sie aufbringen mussten, kreuzen Sie bitte so an:



Beanspruchungshöhe

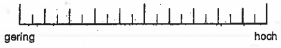
Geben Sie jetzt bitte an, wie hoch die Beanspruchung in den einzelnen Dimensionen war. Markieren sie dazu auf den folgenden Skalen bitte, in welchem Maße Sie sich in den sechs genannten Dimensionen von der Aufgabe beansprucht oder gefordert gesehen haben:

Beispiel:



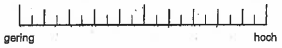
Geistige Anforderungen

Wie viel geistige Anstrengung war bei der Informationsaufnahme und bei der Informationsverarbeitung erforderlich (z.B. Denken, Entscheiden, Rechnen, Erinnern, Hinsehen, Suchen ...)? War die Aufgabe leicht oder anspruchsvoll, einfach oder komplex, erfordert sie hohe Genauigkeit oder ist sie fehlertolerant?



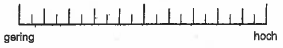
Körperliche Anforderungen

Wie viel körperliche Aktivität war erforderlich (z.B. ziehen, drücken, drehen, steuern, aktivieren ...)? War die Aufgabe leicht oder schwer, einfach oder anstrengend, erholsam oder mühselig?



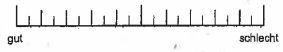
Zeitliche Anforderungen

Wie viel Zeitdruck empfanden Sie hinsichtlich der Häufigkeit oder dem Takt mit dem Aufgaben oder Aufgabenelemente auftraten? War die Abfolge langsam und geruhsam oder schnell und hektisch?



Ausführung der Aufgaben

Wie erfolgreich haben Sie Ihrer Meinung nach die vom Versuchsleiter (oder ihnen selbst) gesetzten Ziele erreicht? Wie zufrieden waren Sie mit Ihrer Leistung bei der Verfolgung dieser Ziele?

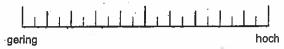


Anstrengung

Wie hart mussten Sie arbeiten, um Ihren Grad an Aufgabenerfüllung zu erreichen?

Frustration

Wie unsicher, entmutigt, irritiert, gestresst und verärgert (versus sicher, bestätigt, zufrieden, entspannt und zufrieden mit sich selbst) fühlten Sie sich während der Aufgabe?



Kontrollieren Sie bitte, ob Sie zu allen Fragen Angaben gemacht haben. Bei Unklarheiten wenden Sie sich bitte an die anwesenden Versuchsleiter.

A.1.2 Demographic questionaire - German version

VP-Nummer:
Datum:

Liebe/rTeilnehmer/In

Im Folgenden bitten wir dich einige Angaben zu deiner Person zu machen. Alle Angaben werden anonymisiert ausgewertet und ausschließlich für die Auswertung dieses Experiments verwendet!

Vielen Dank

Demographischer Fragebogen	
Alter in Jahren	
Geschlecht	☐☐w ☐☐m
Nationalität	
Was machen Sie beruflich?	☐☐Schüler(in)
	☐☐Student(in)
	☐☐Auszubildende(r)
	☐☐Angestellte(r)
	☐☐Selbständig
	☐☐Sonstiges:
	☐☐keine Angabe
Welches Fach studieren Sie?	
Haben Sie einen gültigen Führerschein (Auto)?	☐☐ja ☐☐nein
Wie regelmäßig fahren Sie mit dem Auto?	☐ täglich
	☐ mehrmals pro Woche
	☐ mehrmals im Monat
	☐ mehrmals im Jahr
Wieviele Kilometer legen sie im Jahr durchschnittlich mit dem Auto zurück	☐☐ < 5000km
	☐ 5001 – 10000km
	☐ 10001 – 20000km
	☐ > 20000km

A.1.3 General introduction - German version

Probandeninformation

Sehr geehrte/r Versuchsteilnehmer/In,

wir danken Ihnen sehr für Ihre Bereitschaft an unserem Experiment teilzunehmen.

Mit unserem Versuch möchten wir die mentale Beanspruchung beim Autofahren messen. Dafür werden mittels folgender Methoden verschiedenste physiologische Parameter, während einer simulierten Autofahrt aufgezeichnet.

- **Elektrokardiogramm** (EKG)
  - Messung erfolgt über 3 im Thoraxbereich befestigte Elektroden
- **Blickbewegungsmessung**
  - Messung erfolgt berührungsfrei über eine Eye-Tracking Anlage
- **Sprachaufzeichnung**
  - Messung erfolgt über ein an der Kleidung befestigtes Funkmikrofon
- **Messung der Gesichtstemperatur**
  - Messung erfolgt berührungsfrei über eine Thermographiekamera

Während Ihrer Fahrt im Fahrsimulator besteht Ihre Hauptaufgabe darin, gemäß der Beschilderung am Fahrbahnrand die Fahrspur zu wechseln und solange zu halten, bis ein neues Schild mit einer neuen Markierung erscheint. Parallel dazu werden Ihnen nacheinander die Zahlen (1-9) akustisch präsentiert. Ihre Aufgabe besteht darin fortlaufend die jeweils letzten zwei präsentierten Zahlen miteinander zu addieren.

Der Versuch ist unterteilt in drei Teilfahrten mit einer Dauer von jeweils acht Minuten. Nach jedem Durchgang wird es eine Pause von zwei Minuten geben, in der sie gebeten werden einen Fragebogen auszufüllen. Im Anschluss daran startet die nächste Fahrt.

Eine ausführliche Erläuterung der einzelnen Aufgaben sowie der Ablauf der Teilfahrten erfolgt im Anschluss an diese Probandeninformation. Hier werden Sie ebenfalls die Möglichkeit haben die einzelnen Teilaufgaben zu üben.

Bei Fragen und Anmerkungen wenden Sie sich bitte direkt an den Versuchsleiter.

Noch einmal vielen Dank für Ihre Bereitschaft!

A.1.4 Informed consent form - German version

Technische Universität Berlin

Prof. Dr. Matthias Rötting - Institut für Psychologie und
Arbeitswissenschaften - Sekr. Fr 2-7/1 Franklinstr. 28-29
D-10587 Berlin

Einverständniserklärung

Ich _____
(Name, Vorname)

Geburtsdatum _____

Erkläre, dass ich die Probandeninformation zur Studie:

„Fahrsimulatorstudie –
Einfluss mentaler Beanspruchung auf physiologische Parameter beim Autofahren“

und diese Einverständniserklärung zur Studienteilnahme erhalten habe.

- ✓ Ich wurde für mich ausreichend mündlich und/oder schriftlich über die wissenschaftliche Untersuchung informiert. Zu dem Ablauf und den möglichen Risiken konnte ich Fragen stellen. Die mir erteilten Informationen habe ich inhaltlich verstanden.
- ✓ Mir ist bewusst, dass ich die Teilnahme an der Untersuchung jederzeit und ohne Angabe von Gründen abbrechen kann.
- ✓ Ich erkläre mich bereit, dass im Rahmen der Studie Daten über mich gesammelt und anonymisiert aufgezeichnet werden. Es wird gewährleistet, dass meine personenbezogenen Daten nicht an Dritte weitergegeben werden. Bei der wissenschaftlichen Veröffentlichung wird aus den Daten nicht hervorgehen, wer an dieser Untersuchung teilgenommen hat. Meine persönlichen Daten unterliegen dem Datenschutzgesetz.
- ✓ Ich weiß, dass ich jederzeit meine Einverständniserklärung, ohne Angabe von Gründen, widerrufen kann, ohne dass dies für mich nachteilige Folgen hat.
- ✓ Mit der vorstehend geschilderten Vorgehensweise bin ich einverstanden und bestätige dies mit meiner Unterschrift.

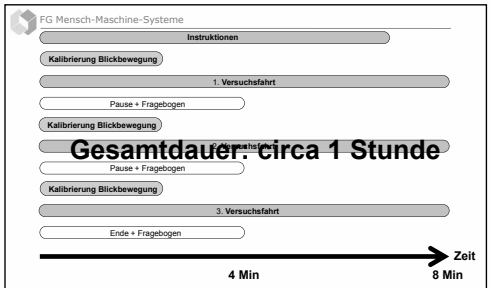
_____	_____	_____
(Ort, Datum)	(Unterschrift des Probanden)	(Unterschrift des Untersuchenden)

A.1.5 Task instructions - German version

Instruktionen

Allgemeine Hinweise zum Versuchsablauf

- Bitte **bewegen Sie ihren Kopf nach Möglichkeit nicht** bzw. nur so wenig wie möglich.
- Den Kopf bitte möglichst an der Stütze anlehnen.
- Bitte das Lenkrad **nicht ruckartig einschlagen** und auch **nicht mehr als 90° nach links und rechts** drehen.



Gesamtdauer: circa 1 Stunde

Lane Change Task (LCT)
Instruktionen

Einführung

- Ihre Aufgabe besteht darin, gemäß der Beschilderung am Seitenrand, die Fahrspuren zu wechseln
- Hierbei sind einige Bedingungen zu beachten, welche nachfolgend verdeutlicht werden.

Anfang

- In dieser Kurve beginnen Sie Ihre Fahrt und fahren anschließend auf die Teststrecke
- Nach Verlassen der Kurve bitte die **mittlere Spur** einnehmen
- Bitte **treten Sie das Gaspedal komplett durch** nachdem Sie die Kurve verlassen haben und **halten es bitte** so **bis zum erreichen der nächsten Kurve**
- Der Versuchsbeginn wird durch **zwei START Schilder** am Fahrbahnrand signalisiert

Start / Versuchsbeginn

- Ab erreichen der **START Schilder** beginnt der Versuch und **Ihre Daten werden aufgezeichnet**
- Bitte nehmen Sie die **mittlere Fahrspur** ein !
- Bitte **treten Sie das Gaspedal komplett durch** und **halten** es so **bis zum erreichen der nächsten Kurve !**

Spurwechsel (1)

- Solange die Schilder am Seitenrand weiß sind, ist **noch kein Spurwechsel** von Ihnen verlangt
- Behalten Sie also bitte Ihre Spur möglichst mittig bei

Spurwechsel (2)

- Sobald Sie auf den Schildern einen Pfeil erkennen können, wechseln Sie bitte **sofort** auf die angegebene Spur (hier rechts)
- Auch beim Spurwechsel gilt: **Gaspedal voll durchgetreten lassen !**

Nach dem Spurwechsel

- Sie haben nun die Spur gewechselt und **behalten diese bei**
- Bitte halten Sie die **Spur** möglichst **mittig**
- Sobald für Sie auf dem nächsten Schild ein Pfeil erkennbar ist, wechseln Sie bitte wieder **sofort** auf die von Ihnen verlangte Fahrspur

Ende einer Teilstrecke

- Am Ende jeder Teilstrecke beginnt eine Kurve welche Sie zum nächsten Streckenabschnitt führt
- Hier spielt sowohl **Geschwindigkeit, als auch Spureinhaltung** eine **untergeordnete Rolle**
- Durchfahren Sie die **Kurve** mit **genügend Sicherheit !**
- **Nach Verlassen der Kurve das Gaspedal bitte wieder komplett durchtreten und die mittlere Fahrspur einnehmen !**

Geschwindigkeit

- Bitte drücken Sie das Gaspedal auf den geraden Streckenabschnitten **immer voll durch**, sodass die Maximalgeschwindigkeit gehalten wird

Fahrbahnbegrenzung

- Bitte auf gar keinen Fall die Strecke verlassen !

Fragen zum LCT

Zusammenfassung - LCT

- Gaspedal auf den geraden Streckenabschnitten immer komplett durchtreten !
- Anweisungen der Schilder am Fahrbahnrand befolgen (Fahrspuren wechseln) !
- Kurven mit verminderter Geschwindigkeit durchfahren !
- Nach Verlassen der Kurve bitte die mittlere Fahrspur einnehmen und das Gaspedal wieder voll durchtreten !
- Bitte auf gar keinen Fall die Strecke verlassen !

ES FOLGT NUN EINE 2 MINÜTIGE TESTFAHRT !
Bitte machen Sie sich dabei mit dem Fahrzeugverhalten vertraut.

PASAT Instruktionen

Einführung

Bei dem **PASAT** Test werden Ihnen über die Kopfhörer **akustisch Zahlen von 1 bis 9** präsentiert.

Dabei können zwei verschiedene Aufgaben an Sie gestellt werden

1. **Wiederholen** Sie bitte fortlaufend die gehörten Zahlen
2. **Addieren** sie bitte fortlaufend die jeweils letzten beiden gehörten Zahlen

1. Aufgabe - Wiederholen

- Bei **einem von insgesamt drei** Versuchsdurchläufen müssen die akustisch dargebotenen Zahlen **direkt wiedergegeben** werden
- **Beispiel:** **9 7 1 5**
  - **Ergebnis:** **9 7 1 5**

ES FOLGT NUN EIN BEISPIEL !!! BITTE KOPFHÖRER AUFSETZEN !!!

Beispiel - 1

Sie hören jetzt beispielhaft Zahlen, die Sie bitte **direkt wiedergeben**

Sie haben gehört: **4 7 8 7 1 3**

Lösung: **4 7 8 7 1 3**

2. Aufgabe - Addieren

- Ihre Aufgabe besteht bei **2 von 3** Versuchsdurchläufen darin, die jeweils **letzten beiden** Ihnen präsentierten **Zahlen** miteinander zu addieren und **laut und deutlich wiederzugeben**

- **Beispiel:** **2 5 6 3**
  - **Ergebnis:** **-** **7** **11** **9**
    (2+5) (5+6) (6+3)

ES FOLGT NUN EIN BEISPIEL !!! BITTE KOPFHÖRER AUFSETZEN !!!

Beispiel - 2

Sie hören jetzt beispielhaft Zahlen, von denen Sie bitte die **letzten beiden gehörten Zahlen miteinander addieren**

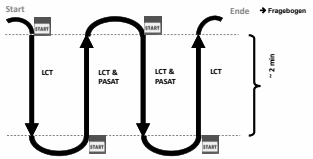
Sie haben gehört: **6 1 7 8 1 3**

Lösung: **- 7 8 15 9 4**
(6+1) (1+7) (7+8) (8+1) (1+3)

Fragen zum PASAT Test

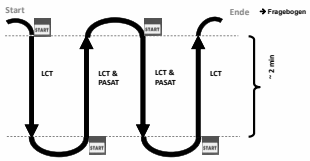
Exemplarischer Versuchsdurchlauf

Alle Versuchsdurchläufe folgen folgendem Schema:



Exemplarischer Versuchsdurchlauf

Alle Versuchsdurchläufe folgen folgendem Schema:



FRAGEN?

Kalibrierung Blickbewegung

Bevor der erste Durchlauf startet muss die Blickbewegungsanlage kalibriert werden. Dafür wird Ihnen gleich ein grauer Kreis mit rotem Punkt in der Mitte präsentiert. Bitte fixieren Sie den kleinen roten Punkt in der Mitte und folgenden Sie dem Kreis mit den Augen, während sich dieser über den Bildschirm bewegt.

Wichtig:

- Bitte folgen Sie dem Punkt **nur** mit den Augen
- Bitte bewegen Sie sich **möglichst nicht** während und **nach** der Kalibrierung

Bevor gleich der erste Durchlauf beginnt, bitten wir Sie sich für 3min zu entspannen! Der Versuchsleiter wird Ihnen dann mitteilen, wenn es weiter geht!

WICHTIG FÜR DIE VERSUCHE !

Nachfolgend werden nun die 3 Versuchsdurchläufe beginnen, in denen Sie die **Fahraufgabe (LCT)** und parallel den **PASAT** ausführen sollen. Vor Beginn jedes Durchlaufs wird Ihnen mitgeteilt welche Aufgabenvariante des PASAT von Ihnen verlangt wird.

- Ihr Hauptaugenmerk soll bei jedem Durchlauf auf die **Fahraufgabe** gerichtet sein!
- Bitte führen Sie alle Fahraufgaben **so schnell und korrekt** wie möglich aus!
- Bearbeiten Sie den PASAT **so gut es Ihnen möglich** ist ohne die Fahraufgabe zu vernachlässigen!

1. Versuchsdurchlauf

Bitte Zahlen nur wiedergeben !

Bevor gleich der nächste Durchlauf beginnt, bitten wir Sie sich für 3min zu entspannen! Der Versuchsleiter wird Ihnen dann mitteilen, wenn es weiter geht!

PAUSE

Bitte füllen Sie den Fragebogen aus.

Kalibrierung Blickbewegung

WICHTIG FÜR DIE VERSUCHE !

ZUR ERINNERUNG

- Ihr Hauptaugenmerk soll bei jedem Durchlauf auf die **Fahraufgabe** gerichtet sein!
- Bitte führen Sie alle Fahraufgaben **so schnell und korrekt** wie möglich aus!
- Bearbeiten Sie den PASAT **so gut es Ihnen möglich** ist ohne Ihre Leistung in der Fahraufgabe zu beeinträchtigen!

PAUSE

Bitte füllen Sie den Fragebogen aus.

2. Versuchsdurchlauf

Bitte addieren Sie die Zahlen in diesem Durchlauf !

Bevor gleich der nächste Durchlauf beginnt, bitten wir Sie sich für 3min zu entspannen! Der Versuchsleiter wird Ihnen dann mitteilen, wenn es weiter geht!

Kalibrierung Blickbewegung

3. Versuchsdurchlauf

Bitte addieren Sie die Zahlen in diesem Durchlauf !

WICHTIG FÜR DIE VERSUCHE !

ZUR ERINNERUNG

- Ihr Hauptaugenmerk soll bei jedem Durchlauf auf die **Fahraufgabe** gerichtet sein!
- Bitte führen Sie alle Fahraufgaben **so schnell und korrekt** wie möglich aus!
- Bearbeiten Sie den PASAT **so gut es Ihnen möglich** ist ohne Ihre Leistung in der Fahraufgabe zu beeinträchtigen!

Ende der Versuche

Bitte füllen Sie den letzten Fragebogen aus.

Vielen Dank für Ihre Teilnahme.

A.2 Data pre-processing

A.2.1 Session info R statistics

- R version 3.3.1 (2016-06-21), `x86_64-apple-darwin13.4.0`
- Locale: `en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: calibrate 1.7.2, car 2.1-3, colorspace 1.2-6, corrplot 0.77, digest 0.6.10, ez 4.4-0, ggplot2 2.2.1, gridExtra 2.2.1, heplots 1.3-3, lattice 0.20-33, MASS 7.3-45, plyr 1.8.4, PMCMR 4.1, reshape 0.8.6, schoRsch 1.3
- Loaded via a namespace (and not attached): assertthat 0.1, grid 3.3.1, gtable 0.2.0, lazyeval 0.2.0, lme4 1.1-12, magrittr 1.5, Matrix 1.2-7.1, MatrixModels 0.4-1, mgcv 1.8-12, minqa 1.2.4, munsell 0.4.3, nlme 3.1-128, nloptr 1.0.4, nnet 7.3-12, parallel 3.3.1, pbkrtest 0.4-6, quantreg 5.26, Rcpp 0.12.6, reshape2 1.4.2, scales 0.4.1, SparseM 1.7, splines 3.3.1, stringi 1.1.1, stringr 1.1.0, tibble 1.2, tools 3.3.1

A.2.2 Praat settings

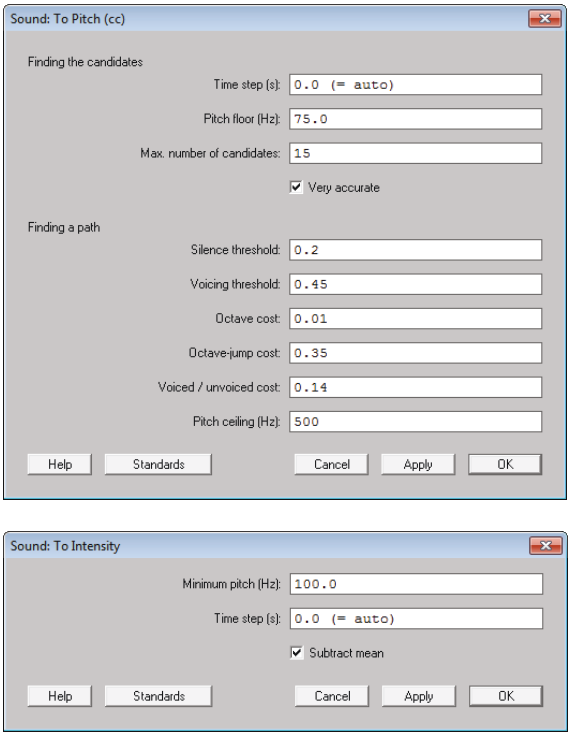


Figure A.1: Praat settings for F0 and voice intensity estimation

Appendix B

Experiment 2

B.1 Documents

B.1.1 Ethics statement - German version

TU Berlin | Sekr. MAR 3-2 | Marchstraße 23 | 10587 Berlin

Fakultät V
Verkehrs- und Maschinensysteme
Institut für Psychologie und Arbeitswissenschaft

Kognitionspsychologie und Kognitive Ergonomie

Dr. Stefan Brandenburg
Vorsitzender Ethikkommission

Sekr. MAR 3-2
Raum MAR 3.080
Marchstraße 23
10587 Berlin

Telefon +49 (0)30 314-24838
Telefax +49 (0)30 314-25289
stefan.brandenburg@tu-berlin.de

Antragsteller: Stefan Ruff

Eingangsdatum des Antrages: 09.03.15

Antragsnummer: RUF_01_20150309

Datum der Beschlussfassung: 03.06.2015

Berlin, 03. Juni 2015

Stellungnahme der Ethikkommission zu einem Forschungsantrag

Die Ethikkommission des Instituts für Psychologie und Arbeitswissenschaft (IPA) der TU - Berlin, im Folgenden Kommission genannt, hat Ihr Forschungsvorhaben begutachtet.

Unter Einhaltung der von Ihnen vorgegebenen Richtlinien werden die wichtigsten Vorkehrungen zur Minimierung des Probandenrisikos getroffen. Es besteht ein angemessenes Verhältnis zwischen dem Nutzen und dem Risiko des Untersuchungsvorhabens.

Die Freiwilligkeit der Versuchsteilnahme wird im geprüften Forschungsvorhaben sichergestellt. Weiterhin werden die Bestimmungen zum Datenschutz eingehalten.

Als Ergebnis der Begutachtung durch die Kommission, wird die Untersuchung „PHYREF – Physiologische Reaktionen auf Fahrerbeanspruchung“ als ethisch unbedenklich eingestuft.

Mit freundlichen Grüßen,

Stefan Brandenburg
- Vorsitzender -

B.1.2 General information - German version

Prof. Dr. Ing. Rötting
TU Berlin - Sekretariat MAR 3-1
Marchstrasse 23, 10587 Berlin

Prof. Dr. Vollrath
TU Braunschweig
Gaußstr. 23, 38106 Braunschweig

Informationsblatt

PHYREF – Physiologische Reaktionen auf Fahrerbeanspruchung

Bitte lesen Sie sich die folgenden Anmerkungen zur Durchführung der Studie aufmerksam durch.

Über die Studie

Die Studie wird durch die Technische Universität Berlin - Fachgebiet Mensch-Maschine-Systeme am Standort des Lehrstuhls für Ingenieur- und Verkehrspsychologie der TU Braunschweig durchgeführt. Die Kontaktadresse des Hauptverantwortlichen des Fachgebietes Mensch-Maschine-Systeme der TU Berlin sowie die Adresse des Hauptverantwortlichen des Lehrstuhls für Ingenieur- und Verkehrspsychologie der TU Braunschweig finden Sie in der Kopfzeile dieses Informationsblatts.

Ziel der Studie ist es, mehr Erkenntnisse über physiologische Veränderungen (Herzaktivität, Hauttemperaturveränderungen, Spracheigenschaften, Blickverhalten und Pupillengröße) des Menschen während des Führens eines Fahrzeuges unter verschiedenen Anforderungen zu gewinnen. Dafür werden Sie im Rahmen dieser Studie gebeten verschiedene Aufgaben während einer simulierten Autofahrt zu bewältigen.

Zunächst bekommen Sie die Möglichkeit sich mit dem Fahrsimulator vertraut zu machen. Darauf folgt ein Training der für Sie relevanten Aufgaben, welche Sie zusätzlich zu ihrer Fahraufgabe erledigen sollen.

Bevor der eigentliche Versuch beginnt werden wir die notwendigen Messinstrumente anbringen (genaue Beschreibung findet sich später im Dokument) und deren Funktionstüchtigkeit überprüfen.

Während der sich anschließenden Versuchsfahrt werden sich immer wieder Phasen abwechseln in denen Sie entweder gebeten werden nur zu Fahren oder zu Fahren und zusätzlich die Ihnen vorher vorgestellte Aufgabe zu bewältigen. Unabhängig davon in welcher Phase Sie sich befinden, werden Sie gebeten so natürlich und sicher wie möglich zu fahren. Die gesamte Fahrzeit beträgt in etwa 40 Minuten. Des Weiteren wird der/die Versuchsleiter/in Ihnen während der Fahrt und am Ende der Fahrt mündliche Fragen stellen bzw. Fragebögen zur Beantwortung vorlegen.

Der Versuch wird in etwa 120 Minuten Ihrer Zeit in Anspruch nehmen. Vor, während und nach dem Versuch steht Ihnen die Versuchsleiterin bzw. der Versuchsleiter jederzeit unterstützend zur Seite.

Prof. Dr. Ing. Rötting
TU Berlin - Sekretariat MAR 3-1
Marchstrasse 23, 10587 Berlin

Prof. Dr. Vollrath
TU Braunschweig
Gaußstr. 23, 38106 Braunschweig

Datenerhebung

Im Rahmen dieser Untersuchung werden verschiedene Daten erhoben. Nachstehend erfolgt eine kurze Erläuterung der einzelnen Daten und es wird ggf. auf Risiken der jeweiligen Datenaufzeichnung hingewiesen. Generell gilt, dass die erhobenen Daten in keinem Fall mit ihrem Namen verbunden werden, so dass kein Rückschluss auf Sie persönlich möglich ist. Alle digital erhobenen Daten werden nach Abschluss des Versuchs auf verschlüsselten Servern anonym gespeichert.

Demografische Merkmale & Gewohnheiten & Krankheiten

Mittels eines Fragebogens werden wir persönliche Daten von Ihnen erfassen, die es uns ermöglichen Sie bestimmten Personengruppen zuzuordnen. Da Faktoren wie Alter oder Geschlecht möglicherweise einen Einfluss auf die erhobenen Daten haben können, wird die Ausprägung dieser Faktoren zur Kontrolle erfasst. Zudem werden wir Ihnen Fragen zu ihren Lebensgewohnheiten (v.a. zum Genussmittelkonsum wie Zigaretten, Kaffee, Alkohol und Drogen) und eventuellen Krankheiten stellen, um Risiken die für Sie durch die Teilnahme entstehen könnten auszuschließen.

Fragen und/oder Fragebögen

Hierbei handelt es sich um weitere Fragen und/oder Fragebögen bevor, während, oder auch nach der Studie, die spezifisch für diese Studie relevant sind.

Verhaltensdaten

Hierbei handelt es sich z.B. um Tastendrücke, Rechnereingaben, Fahrverhalten, Antwortverhalten. Diese Daten sind notwendig, um Ihr Verhalten und Ihre Antworten hinsichtlich Reaktionsgeschwindigkeit, Richtigkeit, Präferenzen etc. auswerten zu können.

Elektrokardiogramm (EKG; Herzaktivität)

Mit einem EKG wird die Herzaktivität gemessen, so dass z.B. die Herzschlagfrequenz analysiert werden kann. Die Erfassung erfolgt an der Hautoberfläche mithilfe von mehreren Elektroden an unterschiedlichen Stellen des Oberkörpers. Um den Kontakt zwischen Elektroden und Haut zu optimieren werden männliche Probanden gegebenenfalls gebeten, die Haare an der jeweiligen Stelle selbstständig durch einen bereitgestellten Einwegrasierer zu entfernen. Danach wird zunächst die Haut mit Alkohol gereinigt. Mit Elektrodengel wird die Leitfähigkeit weiter verbessert. Die verwendeten Stoffe sind klinisch getestet und lassen sich nach Abschluss des Experiments leicht abwaschen. Durch die Vorbereitung und den Druck der Elektroden kann es zu leichten Rötungen auf der Haut kommen, die in der Regel nach einigen Stunden wieder verschwinden. Bitte teilen Sie uns mit, falls Sie an bestimmten Hautallergien oder Hautüberempfindlichkeiten leiden.

Blickbewegungserfassung mittels Eye-Tracking

Mithilfe eines Eye-Tracking Systems kann kamerabasiert z.B. Ihre Blickrichtung oder Ihre Pupillengröße erfasst werden. Das System wird vor dem Versuch individuell angepasst. Die Versuchsleiterin bzw. der Versuchsleiter wird Ihnen den Ablauf erklären.

Prof. Dr. Ing. Rötting
TU Berlin - Sekretariat MAR 3-1
Marchstrasse 23, 10587 Berlin

Prof. Dr. Vollrath
TU Braunschweig
Gaußstr. 23, 38106 Braunschweig

Hauttemperaturmessung

Die Messung der Hauttemperatur erfolgt über einen kleinen Temperatursensor der auf ihrer Nasenspitze platziert und mithilfe eines Pflasters befestigt wird. Dafür wird vor der Anbringung, die Haut mit Alkohol gereinigt.

Audioaufzeichnungen

In diesem Experiment werden akustische Hinweise präsentiert oder es sind akustische Reaktionen erforderlich, die zu Auswertungszwecken mit Mikrofon(en) aufgezeichnet werden.

Videoaufzeichnungen

Zur Auswertung von z.B. Bewegungsdaten und zur Überprüfung eines reibungsfreien Experimentalablaufs filmen wir das Experiment mit mehreren Kameras und speichern die Aufnahmen. Die Kameras werden über die komplette Versuchsdauer Aufnahmen frontal von vorn, aus der Vogelperspektive und aus dem Fußraum aufzeichnen.

Ihre Rechte

Ihre Teilnahme ist freiwillig. Sie haben als Teilnehmer jederzeit das Recht, den Versuch abzubrechen ohne dafür Gründe benennen zu müssen. Wenn Sie die Studie vorzeitig abbrechen, haben Sie Anspruch auf eine anteilige Vergütung für die bis dahin absolvierte Zeit.

Sie haben jederzeit das Recht, zu verlangen, dass die Audio- und/oder Videoaufzeichnungen eingestellt werden. Falls Sie dies wünschen, werden die bis dahin aufgezeichneten Daten umgehend gelöscht. In diesem Fall wird der Versuch abgebrochen.

Nutzung und Anonymisierung der Daten

Die Technische Universität Berlin arbeitet nach den gesetzlichen Bestimmungen für den Datenschutz. Die erhobenen Daten werden elektronisch in anonymisierter Form gespeichert. Hierfür wird für Sie eine Versuchspersonennummer erstellt, die keinen Rückschluss auf ihre Person zulässt. Eine Codierliste, die eine Zuordnung von Namen und Versuchspersonennummer ermöglicht, wird getrennt von allen Daten passwortgeschützt aufbewahrt. Die Möglichkeit der Wiederherstellung des Namensbezugs zu den Daten erfolgt ausschließlich, um Ihnen die nachträgliche Löschung Ihrer Daten zu ermöglichen.

Die aufgezeichneten Daten werden ausschließlich zu wissenschaftlichen Zwecken verwendet und können auch zu wissenschaftlichen Zwecken in anonymisierter Form an andere Wissenschaftler weitergegeben und gegebenenfalls veröffentlicht werden.

Weitere Anmerkungen

Es kann vorkommen, dass z.B. wegen technischer Probleme, das Experiment vorzeitig abgebrochen werden muss. In diesem Fall haben Sie Anspruch auf eine anteilige Vergütung für die bis dahin absolvierte Zeit.

Prof. Dr. Ing. Rötting
TU Berlin - Sekretariat MAR 3-1
Marchstrasse 23, 10587 Berlin

Prof. Dr. Vollrath
TU Braunschweig
Gaußstr. 23, 38106 Braunschweig

Vergütung

Für Ihre Teilnahme erhalten Sie pro Stunde wahlweise 8 Euro, oder 1 vom Lehrstuhl für Ingenieurs- und Verkehrspsychologie der TU Braunschweig anerkannte Versuchspersonenstunde. Ein zusätzlicher Bonus von 2 Euro / Stunde kann bei Übertreffen einer Mindestgesamtleistung im Versuch erworben werden. Aus experimentellen Gründen kann das spezifische Kriterium hier nicht ausgeführt werden.

Wenn Sie noch Fragen zum Informationsblatt haben, dann wenden Sie sich bitte an den Versuchsleiter. Vielen Dank!

B.1.3 General instructions - German version

Instruktionen - Versuchsablauf

Die Fahrt dauert insgesamt in etwa 40 Minuten. Es werden sich immer wieder Abschnitte abwechseln auf denen Ihre einzige Aufgabe das Fahren ist und Abschnitte in denen Sie zusätzlich zur Fahraufgabe, die im Training geübte Nebenaufgabe zu bewältigen haben. Die Abschnitte mit den Nebenaufgaben werden immer kurz vorher durch die Ansage „Aufgabe Start" angekündigt!

Ihr Hauptaugenmerk **sollte auf der Fahraufgabe** liegen. Bemühen Sie sich bitte trotzdem in den Abschnitten mit zusätzlicher Nebenaufgabe diese so gut wie möglich zu bearbeiten, ohne dabei die Fahraufgabe zu vernachlässigen. Sollten Sie bei der **Nebenaufgabe durcheinander kommen, so setzen Sie einfach so bald wie möglich wieder** ein. Denken Sie daran, dass Sie bei Überschreiten einer Mindestgesamtleistung zusätzlich 2 Euro pro Stunde erhalten können.

Bitte beachten Sie die Straßenverkehrsordnung (Ampeln, Vorfahrtsregeln usw.) und fahren Sie so, wie sie es auch im Straßenverkehr tun würden. Bitte halten sie die Geschwindigkeitsbegrenzungen ein und überholen Sie nicht.

Während der Fahrt werden Sie zwischendurch mehrmals aufgefordert einen Fragebogen auszufüllen. Dies wird Ihnen auf dem Bildschirm angekündigt. Bitte halten Sie dann an, füllen den neben Ihnen liegenden Fragebogen aus und geben Sie eine Rückmeldung an den Versuchsleiter, sobald Sie fertig sind.

Sollte Ihnen während der Fahrt schlecht werden oder sich anderes Unwohlsein wie starke Kopfschmerzen o.ä. äußern, geben Sie bitte sofort dem Versuchsleiter Bescheid.

Zum Ende der Versuchsfahrt, der Ihnen ebenfalls über den Bildschirm angezeigt wird, bleiben Sie bitte sitzen und warten auf den Versuchsleiter. Dieser wird Ihnen noch einen abschließenden Fragebogen aushändigen, bevor die Messinstrumente entfernt werden. Danach haben Sie es geschafft!

Aus wissenschaftlichen Gründen bitten wir Sie, keine Informationen über den Versuch an potentielle weitere Versuchspersonen weiterzugeben, da dies die Ergebnisse beeinflussen könnte! Dies gilt selbstverständlich nur so lange die Untersuchungen laufen.

Sollten Sie nach Beendigung des Versuchs noch Fragen haben, dann wenden Sie sich gerne an den Versuchsleiter oder schreiben Sie eine Email an sru@mms.tu-berlin.de!

Vielen Dank für Ihre Teilnahme!

B.1.4 N-back instructions example for 2-back - German version

Instruktionen

Während des Experiments wird Ihnen eine Aufgabe, der n-back Task, gestellt, der Ihnen im Folgenden erklärt wird.

Wenn Sie Fragen haben, stellen Sie Sie bitte immer sofort.

Die Variante, die im Experiment genutzt wird, ist der 2-back Task. Hierbei wiederholen Sie nach jeder Ziffer, die vorletzte zur aktuellen Ziffer genannte Ziffer.

Beispiel:

VL sagt:	3	2	5	7	9
Sie sagen:	Nichts	Nichts	3	2	5

Die Variante, die im Experiment genutzt wird, ist der 2-back Task. Hierbei wiederholen Sie nach jeder Ziffer, die vorletzte zur aktuellen Ziffer genannte Ziffer.

Beispiel:

Bereit für eine Übung? Dann klicken Sie zur nächsten Folie.

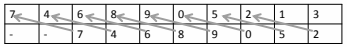
Übung:

7	4	6	8	9	0	5	2	1	3

Lösung:

7	4	6	8	9	0	5	2	1	3
-	-	7	4	6	8	9	0	5	2

Lösung:



Bisher haben Sie die vorgelesenen Ziffern gleichzeitig gesehen. In der Aufgabe wird dies nicht der Fall sein. Daher folgt nun eine Übungsaufgabe, bei der Ihnen die Ziffern nur vorgelesen werden.

Antworten Sie so wie zuvor, indem Sie die jeweils vorletzte Ziffer wiederholen.

2-back Übung

Haben Sie das Prinzip des 2-back Tasks verstanden? Falls nicht, stellen Sie Ihre Fragen jetzt der Versuchsleitung.

Wenn Sie die Aufgabe verstanden haben, können Sie nun mit dem Experiment beginnen.

Wenden Sie sich an die Versuchsleitung, wenn sie bereit sind!

B.1.5 N-back training - German version

VL-Unterlagen: n-back Training

Kurztrainings:

Die folgenden Trainings sollen vom Probanden absolviert werden. Nach jedem Durchgang sind die korrekten Antworten zu zählen. Wenn mindestens die Hälfte (5,5,4 für 0,1,2-back respektive) richtig beantwortet wurde, kann die VP auf Wunsch mit dem Versuch beginnen.

Wenn nicht erfüllt: „Ok, ich starte noch einen Durchgang.“.

Wenn Kriterium erfüllt: „Fühlen Sie sich sicher oder möchten Sie noch einen Trainingsdurchgang?“

Wenn nach 5 Durchgängen nicht erfüllt/Wunsch nach mehr Training, wieder bei trainingkurz1.wav anfangen. Wenn nach 10 Durchgängen nicht erfüllt, VP danken und Versuch abbrechen (Analog zu Mehler et al: „Es gibt einige Personen, die mit dieser speziellen Aufgabe Schwierigkeiten haben. Leider können wir diese Personengruppe in unserem Experiment nicht berücksichtigen. Vielen Dank für Ihre Zeit und Teilnahme bisher, wir werden Sie für die absolvierte Dauer vergüten.“)

trainingkurz1.wav

Ziffer	0-back		1-back		2-back	
	Soll	Korrekt	Soll	Korrekt	Soll	Korrekt
4	4		-		-	
1	1		4		-	
5	5		1		4	
4	4		5		1	
8	8		4		5	
3	3		8		4	
1	1		3		8	
8	8		1		3	
2	2		8		1	
9	9		2		8	

trainingkurz2.wav

Ziffer	0-back		1-back		2-back	
	Soll	Korrekt	Soll	Korrekt	Soll	Korrekt
6	6		-		-	
7	7		6		-	
3	3		7		6	
9	9		3		7	
8	8		9		3	
2	2		8		9	
7	7		2		8	
8	8		7		2	
2	2		8		7	
4	4		2		8	

trainingkurz3.wav

Ziffer	0-back		1-back		2-back	
	Soll	Korrekt	Soll	Korrekt	Soll	Korrekt
8	8		-		-	
4	4		8		-	
2	2		4		8	
9	9		2		4	
4	4		9		2	
1	1		4		9	
3	3		1		4	
2	2		3		1	
9	9		2		3	
6	6		9		2	

trainingkurz4.wav

Ziffer	0-back		1-back		2-back	
	Soll	Korrekt	Soll	Korrekt	Soll	Korrekt
4	4		-		-	
1	1		4		-	
7	7		1		4	
9	9		7		1	
4	4		9		7	
7	7		4		9	
8	8		7		4	
2	2		8		7	
5	5		2		8	
1	1		5		2	

trainingkurz5.wav

Ziffer	0-back		1-back		2-back	
	Soll	Korrekt	Soll	Korrekt	Soll	Korrekt
8	8		-		-	
9	9		8		-	
3	3		9		8	
7	7		3		9	
9	9		7		3	
5	5		9		7	
2	2		5		9	
6	6		2		5	
8	8		6		2	
9	9		8		6	

B.1.6 Informed consent - German version

Einverständniserklärung

Ich _____
(Name, Vorname)

Geburtsdatum _____

erkläre, dass ich die Probandeninformation zur Studie:

PHYREF – Physiologische Reaktionen auf Fahrerbeanspruchung

und diese Einverständniserklärung zur Studienteilnahme erhalten habe.

- ✓ Ich wurde für mich ausreichend mündlich und schriftlich [siehe Informationsblatt] über die wissenschaftliche Untersuchung informiert. Zu dem Ablauf und den möglichen Risiken konnte ich Fragen stellen. Die mir erteilten Informationen habe ich inhaltlich verstanden.
- ✓ Mir ist bewusst, dass die Teilnahme an der Untersuchung freiwillig ist und ich jederzeit und ohne Angabe von Gründen meine Teilnahme abbrechen kann.
- ✓ Ich erkläre mich bereit, dass im Rahmen der Studie Daten über mich gesammelt und anonymisiert aufgezeichnet werden. Es wird gewährleistet, dass meine personenbezogenen Daten nicht an Dritte weitergegeben werden. Bei der wissenschaftlichen Veröffentlichung wird aus den Daten nicht hervorgehen, wer an dieser Untersuchung teilgenommen hat. Meine persönlichen Daten unterliegen dem Datenschutzgesetz.
- ✓ Ich weiß, dass ich jederzeit meine Einverständniserklärung, ohne Angabe von Gründen, widerrufen kann, ohne dass dies für mich nachteilige Folgen hat.
- ✓ Mit der vorstehend geschilderten Vorgehensweise bin ich einverstanden und bestätige dies mit meiner Unterschrift.

_____	_____	_____
(Ort, Datum)	(Unterschrift des Probanden)	(Unterschrift des Untersuchenden)

B.1.7 Informed consent form audio/video - German version

Technische Universität Berlin

Fachgebiet Mensch-Maschine-Systeme

M.Sc. Stefan Ruff

Ansprechpartner für eventuelle Rückfragen:

Stefan Ruff

Telefon: 030/314 79518

Einwilligungserklärung für Bild- und Tonaufnahmen

Technische Universität Berlin

Titel der Studie: PHYREF – Physiologische Reaktionen auf Fahrerbeanspruchung

Ich (Name des Teilnehmers /der Teilnehmerin in Blockschrift)

_____

bin mündlich und schriftlich vom Versuchsleiter darüber informiert worden, dass im Rahmen der Studie eine Video- und Tonaufnahme gemacht wird und die Aufzeichnung Voraussetzung für die Teilnahme am Versuch ist.

Die Aufnahme dient dazu, Sprachsignale aus der Tonaufnahme zu extrahieren und auf verschiedene Charakteristika während des Fahrverlaufes hin zu untersuchen.

Ich bin darüber informiert, dass die Aufzeichnung und Auswertung der Video und Tonaufnahme anonymisiert erfolgt, d. h. unter Verwendung einer Versuchspersonennummer und ohne Angabe meines Namens. Es besteht die sehr geringe Wahrscheinlichkeit, dass eine an der Datenauswertung beteiligte Person mich erkennt. Aus diesem Grund unterliegen alle an der Auswertung beteiligten Personen einer absoluten Schweigepflicht und dürfen unter keinen Umständen vertrauliche Informationen an Dritte weitergeben.

Mir ist bekannt, dass ich mein Einverständnis zur Aufbewahrung bzw. Speicherung dieser Daten widerrufen kann, ohne dass mir daraus Nachteile entstehen. Die Video und Tonaufnahme werden auf einem passwortgeschützten Server digital hinterlegt. Ich bin darüber informiert worden, dass ich jederzeit eine Löschung meiner Aufnahmen unter Angabe meines Namens verlangen kann. Die Aufnahmen werden aber in jedem Fall nach Abschluss der Auswertung vernichtet.

Mit der beschriebenen Handhabung der erhobenen Aufnahmen bin ich einverstanden.

Zusatz für Demonstrationen *Ich gebe mein Einverständnis, dass meine >Video / Bild / Tonaufnahme< zu Demonstrationszwecken in teilnehmerbegrenzten Veranstaltungen (z. B. Lehrveranstaltungen) abgespielt werden. Zutreffendes bitte ankreuzen: O JA O NEIN.*

Die Einverständniserklärung für die Video und Tonaufnahme ist freiwillig. Ich kann diese Erklärung jederzeit widerrufen. Im Falle einer Ablehnung oder eines Rücktritts entstehen für mich keinerlei Kosten oder anderweitige Nachteile; eine Teilnahme an der Studie ist dann allerdings nicht möglich.

Ich hatte genügend Zeit für eine Entscheidung. Ich habe alles gelesen und verstanden und erkläre mich hiermit bereit, dass eine Video und Tonaufnahme von mir gemacht wird.

Eine Ausfertigung dieser Einwilligungserklärung habe ich erhalten.

Ort, Datum & Unterschrift des Teilnehmers: _____

Name des Teilnehmers in Druckschrift: _____

Ort, Datum & Unterschrift des Versuchsleiters: _____

Name des Versuchsleiters in Druckschrift: _____

Bei Fragen oder anderen Anliegen kann ich mich an folgende Personen wenden:

Versuchsleiter:	Projektleiter:
Name:	Stefan Ruff
Anschrift:	Marchstr. 23, D-10587 Berlin
Telefonnummer:	+49 (0)30 314-79518
Email:	stefan.ruff@mms.tu-berlin.de

Vom Autor angepasste Vorlage der Ethikkommission der Deutschen Gesellschaft für Psychologie für die Einwilligungserklärung für Bild- und Tonaufnahmen
01. Juli 2013

B.1.8 Demographic questionaire - German version

Fragebogen zu demographischen Angaben

Alter: _____ Jahre

Biologisches Geschlecht: w □ m □ anderes □

Nationalität: _____

Beruf: _____ Wenn Student, Studienfach: _____

Körpergröße: _____ cm

Körpergewicht: _____ kg

Vielen Dank ! (Bitte weiter auf nächster Seite)

Fragebogen zu Gewohnheiten und Befinden

1. Wie ist Ihr Allgemeinbefinden derzeit?

Sehr gut □ gut □ mittelmäßig □ eher schlecht □ schlecht □

2. Wie haben sie heute Nacht geschlafen?

sehr gut				sehr schlecht
1	2	3	4	5
❍	❍	❍	❍	❍

3. Wann sind Sie heute aufgestanden?_____ Uhr

4. Wieviel Stunden haben Sie die letzte Nacht geschlafen? _____Stunden

5. Trinken Sie regelmäßig Kaffee oder schwarzen Tee? ja □ nein □
   Wenn ja, wieviel pro Tag? _____Tassen

6. Wieviel Tassen schwarzen Tee oder Kaffee haben Sie heute schon getrunken? _____ Tassen

7. Wann haben Sie heute die letzte Tasse schwarzen Tee oder Kaffee getrunken? Um ____Uhr

8. Rauchen Sie regelmäßig? ja □ nein □
   Wenn ja, wie viele Zigaretten pro Tag? _____ Zigaretten

9. Wie viele Zigaretten haben Sie heute schon geraucht? _____ Stück

10. Wann haben Sie heute die letzte Zigarette geraucht? Um _____ Uhr

11. Trinken Sie regelmäßig Alkohol? ja □ nein □
    Wenn ja, welche Art von Alkohol _____ und wieviel durchschnittlich pro Tag in ____Litern

12. Haben Sie in den letzten 24 Stunden Alkohol getrunken? ja □ nein □
    Wenn ja, Art des Getränks: _____ Mengen: ____ Liter

13. Konsumieren Sie regelmäßig Drogen? ja □ nein □
    Wenn ja, Art der Drogen: _____ Menge: _____

14. Haben Sie in den letzten 24 Stunden Drogen konsumiert? ja □ nein □
    Wenn ja, Art der Drogen: _____ Menge: _____

15. Nehmen sie regelmäßig Medikamente zu sich?ja □ nein □
    Wenn ja, Art des Medikaments:_____Menge:____

16. Haben Sie in den letzten 24 Stunden Medikamente zu sich genommen? ja □ nein □
    Wenn ja, Art des Medikaments: _____ Menge: ____

17. Haben Sie Probleme mit dem Herz-Kreislaufsystem? ja □ nein □
    Wenn ja, welche? _____

18. Haben Sie andere gesundheitliche Probleme? ja □ nein □
    Wenn ja, welche? _____

19. Treiben sie regelmäßig Sport? ja □ nein □
    Wenn ja, Art des Sports: _____und wie häufig _____pro Woche?
    Wenn ja, wie lang durchschnittlich: _____Stunden?

20. Haben Sie heute Sport gemacht? ja □ nein □
    Wenn ja, Art des Sports: _____
    Wie lange? _____ Stunden
    Bis wann? Bis _____ Uhr

21. Wann haben Sie Ihre letzte Mahlzeit zu sich genommen? ____Uhr

22. Gründe für die Versuchsteilnahme:

Ich nehme an diesem Versuch teil, weil:

_____

_____

Vielen Dank ! (Bitte weiter auf nächster Seite)

Fragebogen_Demographie_Befinden_Fahrgewohnheiten_StefanRuff_032015

Fragebogen zu Ihren Fahrgewohnheiten

1. Besitzen Sie ein eigenes Fahrzeug? Ja ❍ Nein ❍

Wenn Ja, Marke ..... und Typ ..... Ihres derzeit meist genutzten Wagens.

2. Gibt es in Ihrem Haushalt weitere Autos? Ja ❍ Nein ❍

Wenn Ja, Marke ..... und Typ .....

3. Nutzen Sie Ihr Fahrzeug beruflich? Ja ❍ Nein ❍
4. Sind Sie Berufskraftfahrer? Ja ❍ Nein ❍
5. Seit wieviel Jahren haben Sie Ihren Führerschein für Pkw? Seit ..... Jahren.
6. Wie schätzen Sie Ihren Fahrstil ein?

sehr ruhig						sehr dynamisch
-3	-2	-1	0	1	2	3
❍	❍	❍	❍	❍	❍	❍

7. Wieviel Kilometer fahren Sie jährlich mit dem Auto?

unter 10.000	10.000 - 20.000	mehr als 20.000
❍	❍	❍

Bsp.(1): Sie fahren täglich mit dem Auto zur Arbeit (z.B. 20km*200 Tage=4.000km) und einmal jährlich in Urlaub (ca. 2000km). Darüber hinaus nutzen Sie ihr Auto nicht. Ihre jährliche Fahrleistung beträgt also unter 10.000km.

Bsp.(2): Sie fahren täglich mit dem Auto zur Arbeit (z.B. 20km*200 Tage=4.000km) und besuchen 4mal im Jahr ihre Eltern in Stuttgart (4*1300km). Einmal jährlich fahren sie mit dem Auto in Urlaub (ca.2000km). Ihre jährliche Fahrleitung beträgt also 10.000 bis 20.000km.

Bsp.(3): Sie nutzen ihr Auto täglich sowohl beruflich als auch privat und fahren pro Woche über 400km. Ihre jährliche Fahrleistung beträgt also über 20.000km.

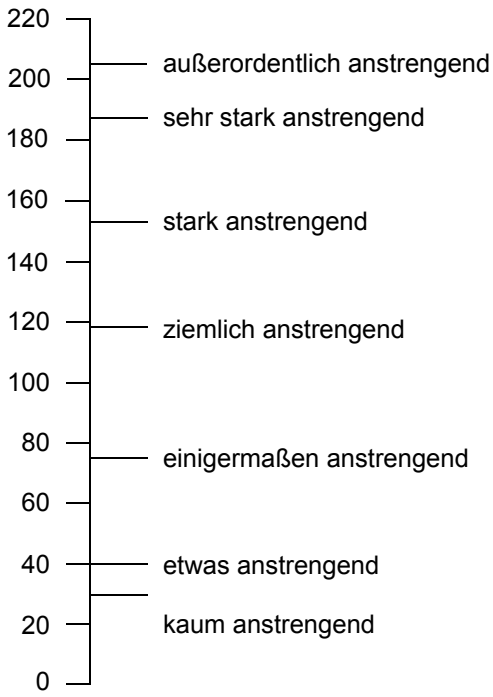
Vielen Dank! (Ende)

B.1.9 SEA Scale - German version

Subjektiv erlebte Anstrengung

Bitte kreuzen Sie frei auf der gesamten Skala an. Die Worte und Zahlen dienen dabei nur zur Orientierung.

Wie anstrengend war der gerade absolvierte Streckenabschnitt mit der Nebenaufgabe insgesamt?



Vielen Dank !

➢ bitte wenden

B.2 Data pre-processing

B.2.1 Praat settings

Intensity: [to intensity] with pitch minimum = 75Hz and time step of 0.001s

F0: [to pitch (ac)] time step=0.001s, pitch floor= 75, max number of candidates = 15, very accurate="on", silence threshold= 0.03, voicing threshold= 0.45, octave cost= 0.01, octave jump cost = 0.35, voiced/unvoiced cost = 0.14, pitch ceiling = 500

B.2.2 Praat script

```
list = Create Strings as file list: "Filelist", "Praat_Data/*.wav"
n = Get number of strings
for y to n
 selectObject: list
 filename$= Get string: y
 fileID=Read from file: "Praat_Data/" + filename$
 writeInfoLine: fileID
 tmin = Get start time
 tmax = Get end time

 To Pitch (ac): 0.001,75, 15,"on", 0.03, 0.45,0.01, 0.35, 0.14, 500
 Rename: "pitch"

 selectObject: fileID
 To Intensity: 75, 0.001
 Rename: "intensity"

for i to (tmax-tmin)/0.01
    time = tmin + i * 0.01
    selectObject: "Pitch pitch"
    pitch = Get value at time: time, "Hertz", "Linear"
    selectObject: "Intensity intensity"
    intensity=Get value at time: time, "Cubic"
    appendFileLine:"Praat_Data/output/" +
    filename$ +".txt", fixed$ (time*1000, 1), " ",
```

```
    fixed$ (pitch, 2), " ", fixed$ (intensity, 1)
 endfor

 removeObject:fileID
 removeObject:"Pitch pitch"
 removeObject:"Intensity intensity"

endfor

removeObject: list
```

B.2.3 R packages

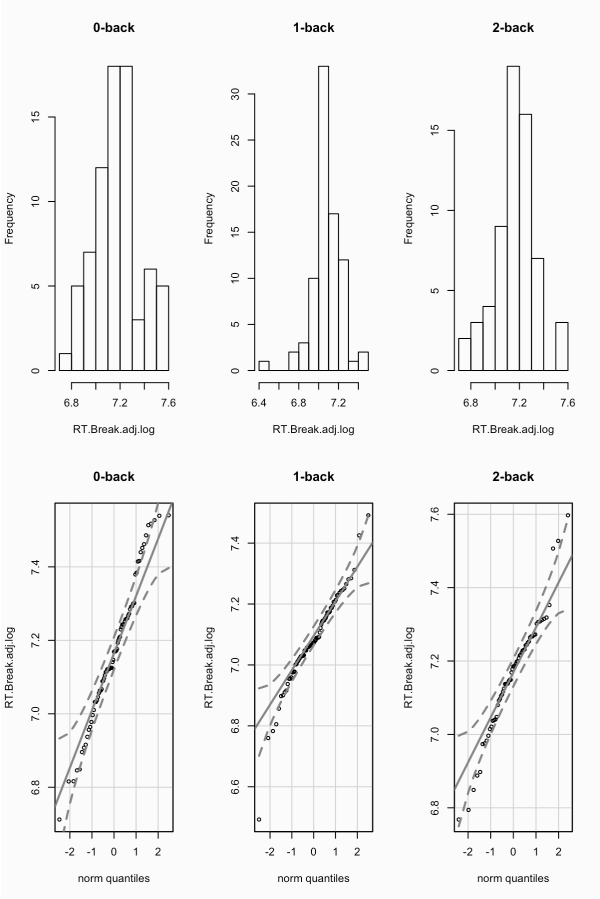
- R version 3.3.1 (2016-06-21), `x86_64-apple-darwin13.4.0`
- Locale:
  `en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, graphics, grDevices, grid, methods, stats, utils
- Other packages: calibrate 1.7.2, car 2.1-3, caret 6.0-71, colorspace 1.2-6, corrplot 0.77, cowplot 0.6.3, digest 0.6.10, DTK 3.5, ez 4.4-0, ggplot2 2.2.1, gridExtra 2.2.1, gvlma 1.0.0.2, heplots 1.3-3, knitr 1.14, lattice 0.20-33, LIStest 2.1, MASS 7.3-45, mgcv 1.8-12, multcomp 1.4-6, MVN 4.0.2, mvtnorm 1.0-5, nlme 3.1-128, plyr 1.8.4, R.matlab 3.6.1, reshape 0.8.6, reshape2 1.4.2, schoolmath 0.4, schoRsch 1.3, signal 0.7-6, sjPlot 2.3.0, stringr 1.1.0, survival 2.39-4, TH.data 1.0-7, tidyr 0.6.0, xtable 1.8-2, zoo 1.7-13
- Loaded via a namespace (and not attached): abind 1.4-5, arm 1.9-3, assertthat 0.1, blme 1.0-4, boot 1.3-18, broom 0.4.1, class 7.3-14, cluster 2.0.4, coda 0.19-1, codetools 0.2-14, coin 1.1-2, cvTools 0.3.2, data.table 1.10.0, DBI 0.5-1, DEoptimR 1.0-8, diptest 0.75-7, dplyr 0.5.0, DT 0.2, e1071 1.6-7, effects 3.1-2, flexmix 2.3-13, foreach 1.4.3, foreign 0.8-66, fpc 2.1-10, GGally 1.3.0, gtable 0.2.0, haven 1.0.0, htmltools 0.3.5, htmlwidgets 0.7, httpuv 1.3.3, iterators 1.0.8, kernlab 0.9-25, laeken 0.4.6, lazyeval 0.2.0, lme4 1.1-12, lmtest 0.9-34, magrittr 1.5, Matrix 1.2-7.1, MatrixModels 0.4-1, mclust 5.2.2, merTools 0.3.0, mime 0.5, minqa 1.2.4, mnormt 1.5-5,

modelr 0.1.0, modeltools 0.2-21, moments 0.14, munsell 0.4.3, mvoutlier 2.0.8, nloptr 1.0.4, nnet 7.3-12, nortest 1.0-4, parallel 3.3.1, pbkrtest 0.4-6, pcaPP 1.9-61, pls 2.6-0, prabclus 2.2-6, psych 1.6.12, purrr 0.2.2, quantreg 5.26, R.methodsS3 1.7.1, R.oo 1.21.0, R.utils 2.5.0, R6 2.1.3, RColorBrewer 1.1-2, Rcpp 0.12.6, robCompositions 2.0.3, robustbase 0.92-7, rrcov 1.4-3, sandwich 2.3-4, scales 0.4.1, sgeostat 1.0-27, shiny 0.13.2, sjmisc 2.3.0, sjstats 0.8.0, sp 1.2-4, SparseM 1.7, splines 3.3.1, sROC 0.1-2, stats4 3.3.1, stringdist 0.9.4.4, stringi 1.1.1, tibble 1.2, tools 3.3.1, trimcluster 0.1-2, vcd 1.4-3, VIM 4.6.0

B.3 Results

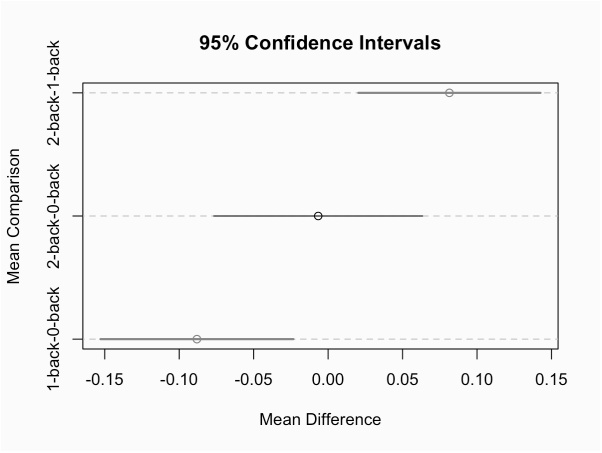
B.3.1 Brake reaction time

```
## [1] "############### RT.Brake.adj.log ###############"
## [1] "Cells with zero Standard Deviation: "
## [1] "TEST FOR NORMALITY WITH SHAPIRO-WILKS"
## Deviation from normality for Condition:  1-back p= 0.0053053235032856
```




```
## [1] "############### RT.Brake.adj.log ###############"
## [1] "LEVENE TEST for HOMOGENEITY of VARIANCE: "
## Deviation from Homogeneity of Variance for Factor Condition   p= 0.0398
217537941122
##
## Call:
## lm(formula = value ~ Cond, data = data)
##
## Coefficients:
## (Intercept)   Cond1-back   Cond2-back
##    7.169664    -0.088101    -0.006651
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = lm(value ~ Cond, data = data))
##
##                      Value p-value                   Decision
## Global Stat        8.169e+00 0.08557    Assumptions acceptable.
## Skewness           8.097e-01 0.36820    Assumptions acceptable.
## Kurtosis           6.385e+00 0.01151 Assumptions NOT satisfied!
## Link Function      3.573e-13 1.00000    Assumptions acceptable.
## Heteroscedasticity 9.742e-01 0.32363    Assumptions acceptable.
## [1] "############### RT.Break.adj.log ###############"
## $`--- ANOVA RESULTS     ------------------------------------`
##   Effect        MSE df1 df2    F     p petasq getasq
## 1   Cond 0.01279898   2  70 4.83 0.011   0.12   0.12
##
## $`--- SPHERICITY TESTS  ------------------------------------`
## [1] "N/A"
##
## $`--- FORMATTED RESULTS ------------------------------------`
##   Effect                                Text
## 1   Cond F(2,70) = 4.83, p = .011, np2 = .12
##
## $`NOTE:`
## [1] "Reporting unadjusted p-values."
##
## [1] "############### RT.Break.adj.log ###############"
## [1] "Post-Hoc for mildly unbalances data"
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = x ~ f)
##
## $f
##                       diff         lwr         upr     p adj
## 1-back-0-back -0.088101345 -0.15027090 -0.02593179 0.0027855
## 2-back-0-back -0.006650967 -0.07295306  0.05965112 0.9695814
## 2-back-1-back  0.081450379  0.01627898  0.14662178 0.0098523
##
```

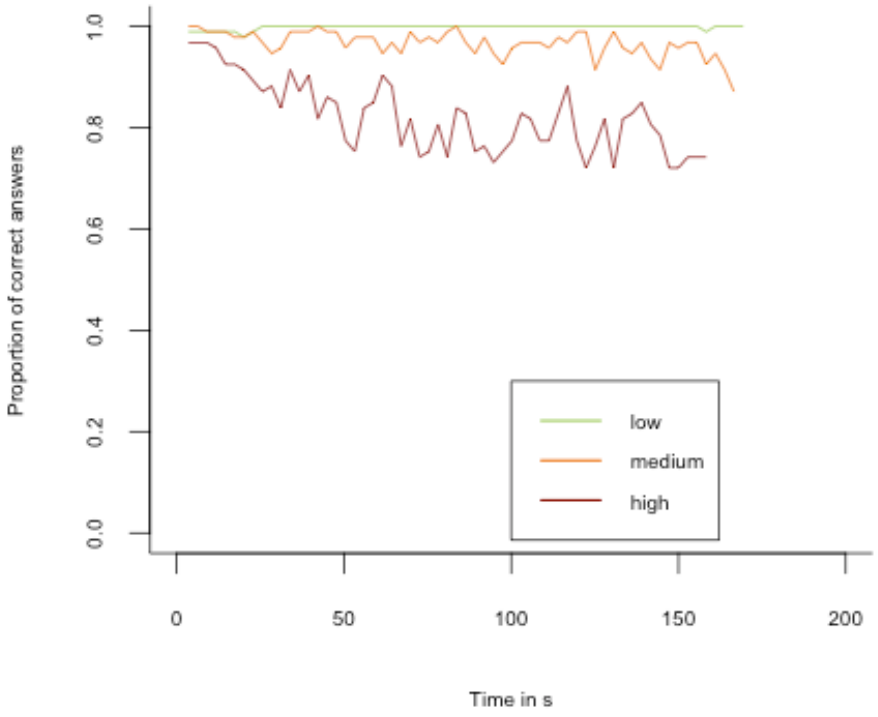
```
## [1] Post-Hoc for Unequal sample sizes and unequal variances between Gro
ups: Dunnet-Tukey-Kramer”


## [[1]]
## [1] 0.05
##
## [[2]]
##                         Diff    Lower CI    Upper CI
## 1-back-0-back -0.088101345 -0.15286832 -0.02333437
## 2-back-0-back -0.006650967 -0.07677321  0.06347128
## 2-back-1-back  0.081450379  0.02029788  0.14260288
```



B.3.2 Sliding means n-back and physiology

Time series of n-back performance and physiological measures

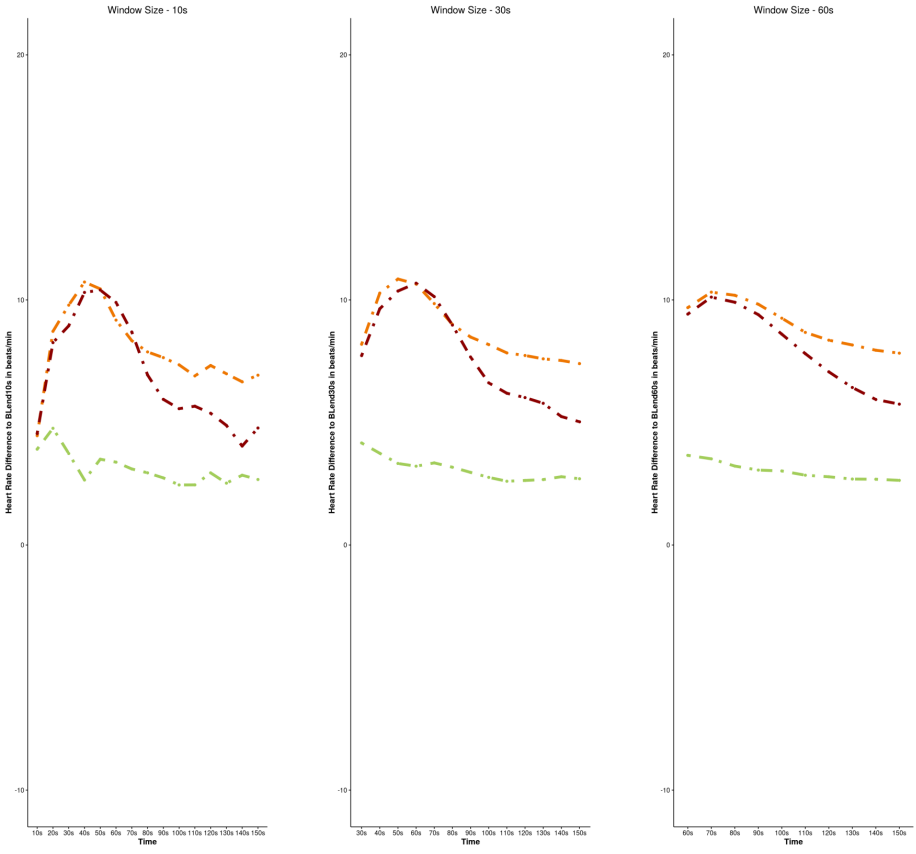
1. N-back over time:
   Mean values of correct answers per subtask over all participants and conditions and blocks



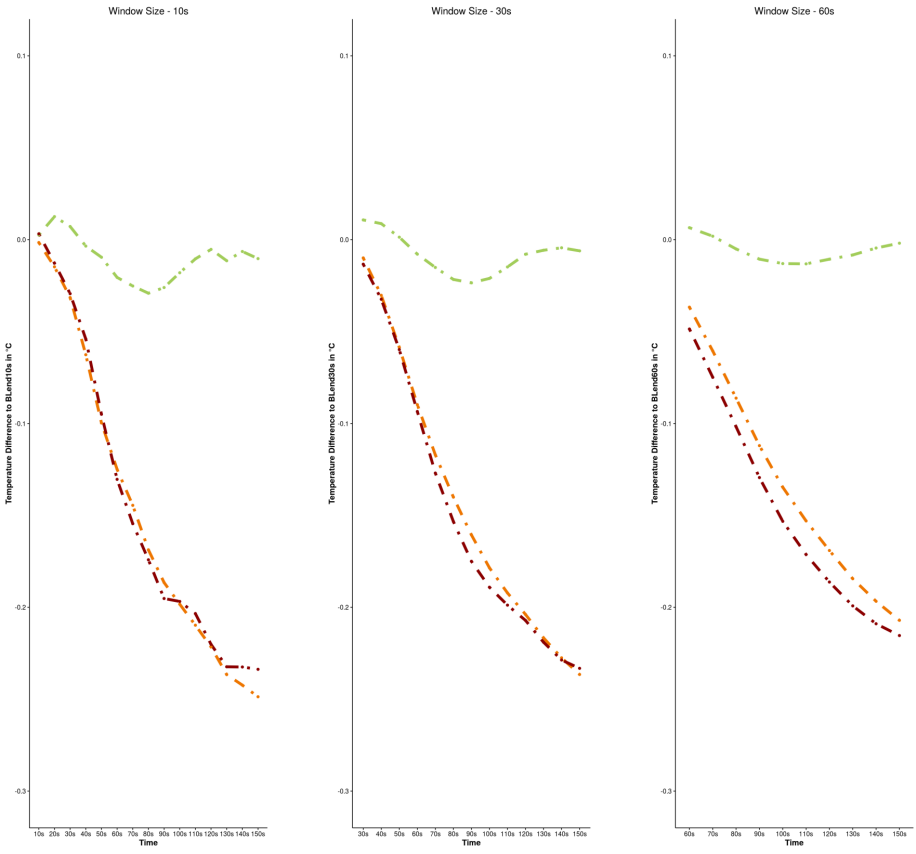
Physiological measures:

Sliding mean absolute values over driving + n-back task sections till the critical event, compared to baseline driving were calculated. Window size was varied between 3 different sizes (10s/30s/60s). Physiological values were averaged over the window size period and compared to the average of the respective last 10s/30s/60s of the baseline driving sections. This was done to ensure to compare the means between equal length intervals. The aforementioned procedure was done for each block and participant. After that the time series were averaged over all blocks and participants for each task difficulty condition (low: 0-back, medium: 1-back, high: 2-back).

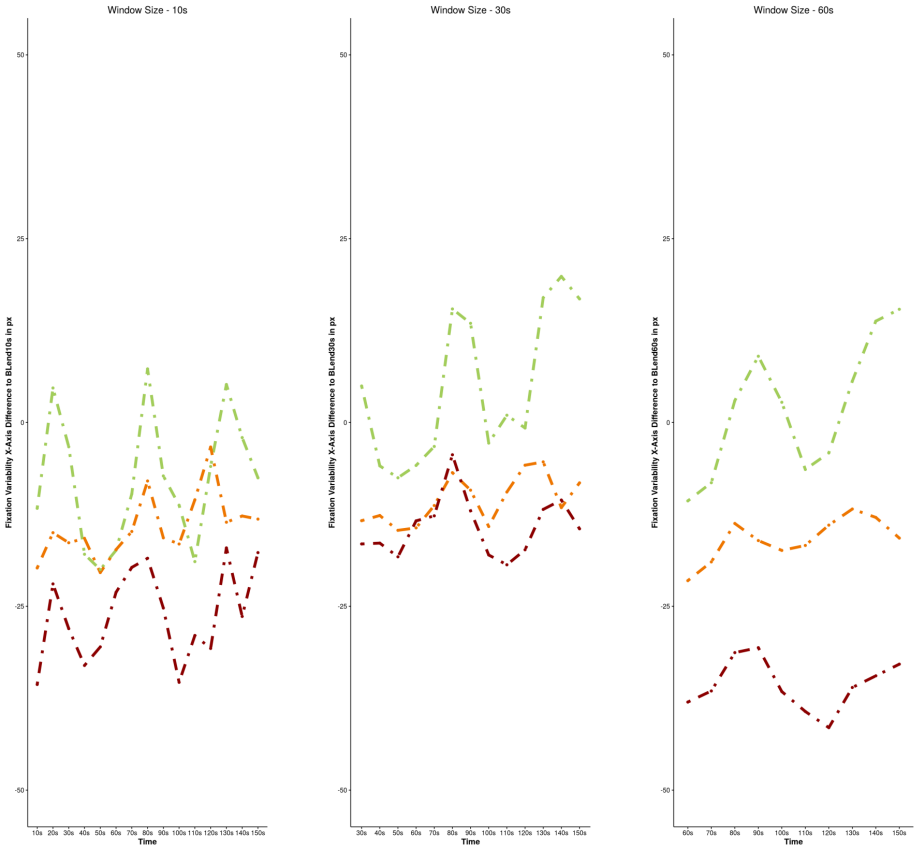
1. Heart rate



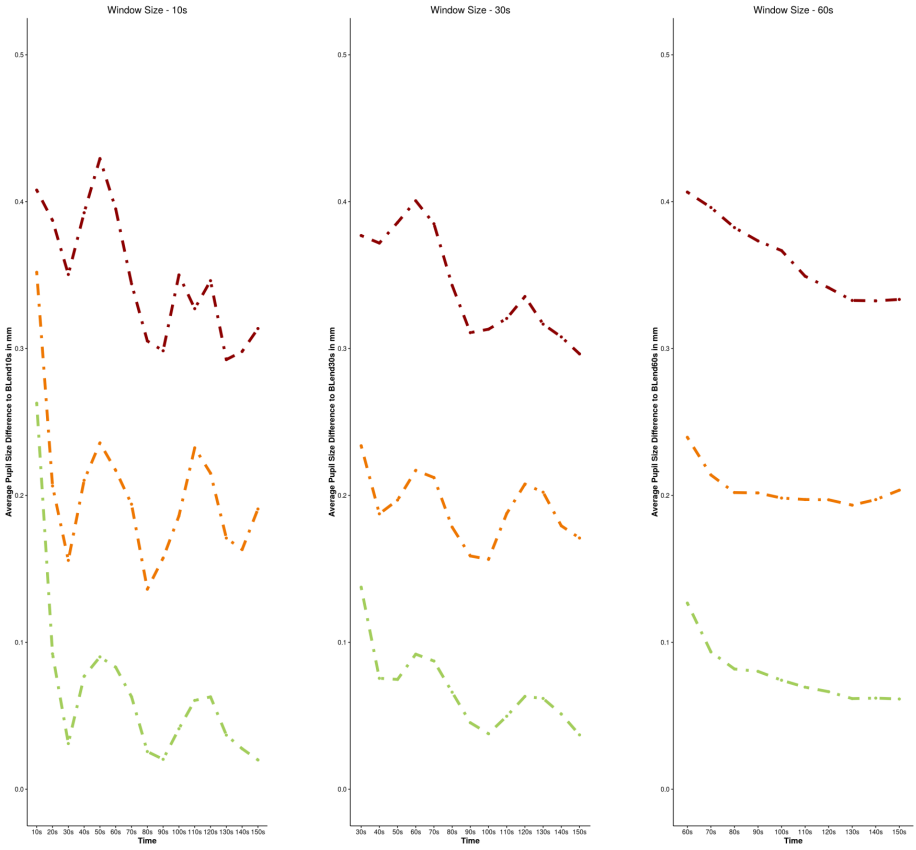
2. Nose tip temperature



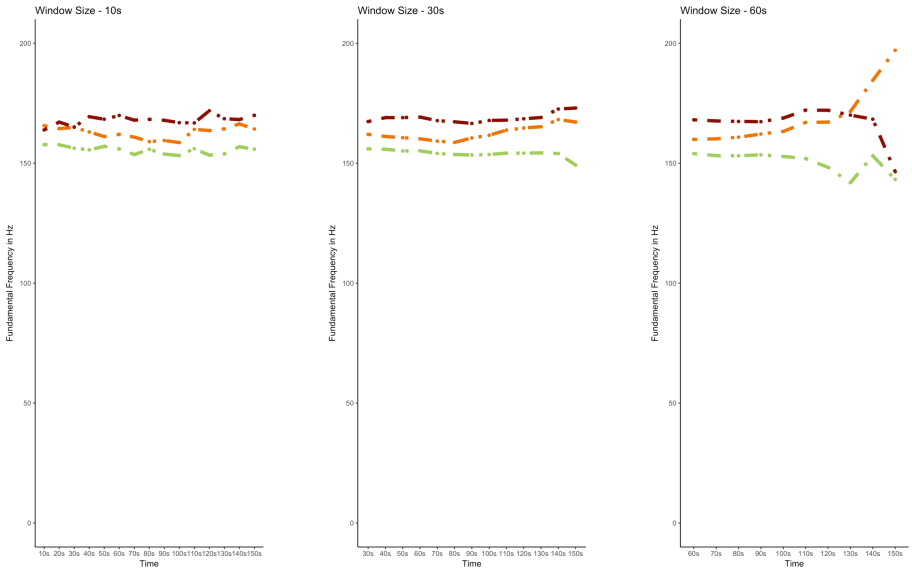
3. Horizontal fixation dispersion



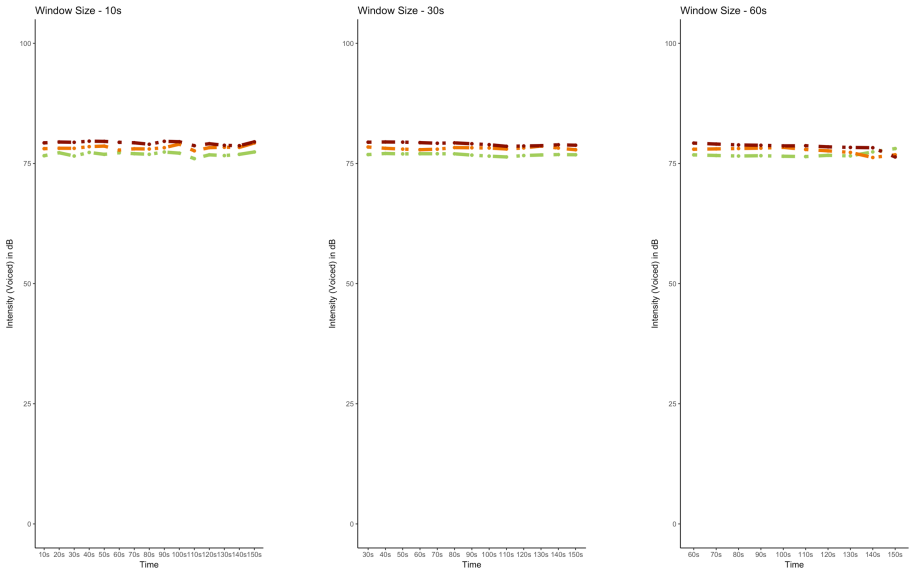
4. Pupil size


Window Size - 10s
Window Size - 30s
Window Size - 60s
Average Pupil Size Difference to BLand10s in mm
Average Pupil Size Difference to BLand30s in mm
Average Pupil Size Difference to BLand60s in mm
0.0
0.1
0.2
0.3
0.4
0.5
10s 20s 30s 40s 50s 60s 70s 80s 90s 100s 110s 120s 130s 140s 150s
30s 40s 50s 60s 70s 80s 90s 100s 110s 120s 130s 140s 150s
60s 70s 80s 90s 100s 110s 120s 130s 140s 150s
Time

5. Fundamental frequency of the voice



6. Voice intensity



Appendix C

Workload classification

C.1 Python software packages used for the classification

- sci-kit learn 0.18.1
- seaborn 0.7.1
- matplotlib 1.5.3
- scipy 0.18.1
- pandas 0.18.1

C.2 Classification

In [1]:
```
# Dissertation: Classification of task difficulty levels by physiological data
# General model

#import packages needed
import os
import glob
import pandas as pd
import numpy as np
import seaborn as sns
sns.set_style("whitegrid")
import matplotlib.pyplot as plt

# import necessary estimators
from sklearn import neighbors
from scipy import stats
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn import model_selection
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn.metrics import roc_auc_score
from sklearn.metrics import classification_report
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.datasets import load_iris

from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
from sklearn.cross_validation import cross_val_score
```

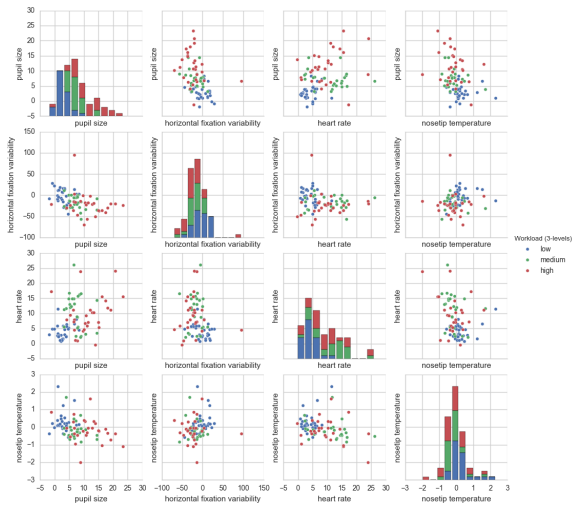
```
/Users/stefanruff/anaconda/lib/python3.5/site-packages/sklearn/cross_validatio
n.py:44: DeprecationWarning: This module was deprecated in version 0.18 in fav
or of the model_selection module into which all the refactored classes and fun
ctions are moved. Also note that the interface of the new CV iterators are dif
ferent from that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
```

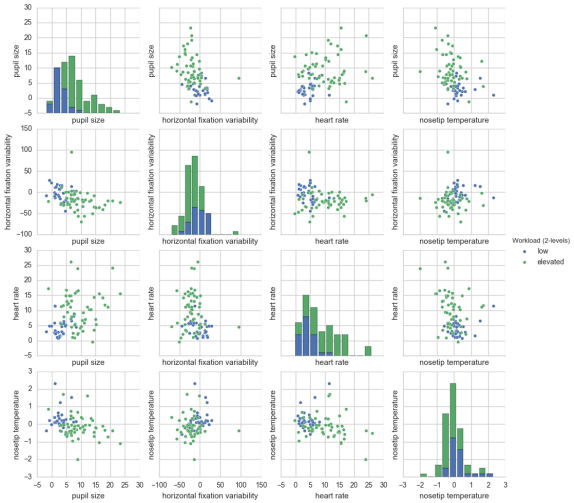
In [2]:
```
# Read in data set

path = r'/Users/stefanruff/HiDrive/users/securifeye/03_Technik/01_MachineLearni
ng/01_Kurs/Stefan'
picpath = r'/Users/stefanruff/HiDrive/users/securifeye/03_Technik/01_MachineLea
rning/01_Kurs/Stefan/Pictures/'

# read in file to pandas data frame

df = pd.read_csv(os.path.join(path, 'study2.csv'), index_col='Subject')
```

In [3]:
```
#Rename Condition for the plots
def transconum(morph):
    if (morph == '0-back'):
        return 'low'
    elif(morph == '1-back'):
        return 'medium'
    else:
        return 'high'

df['Workload (3-levels)'] = df['Condition'].apply(transconum)

# idea make new target by 0-back vs. 1-back/2-back !!sample sizes different bet
ween groups
df['Workload (2-levels)'] =np.where(df.Condition == '0-back', 'low','elevated')
```

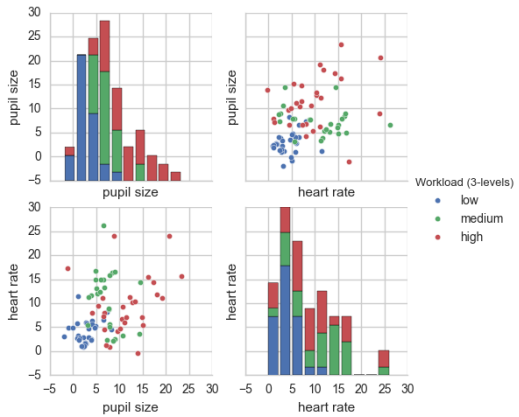

In [4]:
```
# plot scatter of physiological measures
# First all features for binary and trinary outcome
# Second only pupil size and heart rate for binary and trinary outcome

# setup graphics
a4_dims = ( 11.7,8.27)
a4_dims_inv = (8.27, 11.7)
suffix = '.pdf'

# Define features
phy_names=['diffpupilsSAtoBLall', 'difffixSDposxSAtoBLall', 'diffHRsSAtoBLall',
'diffNttSAtoBLall']

print('scatterplot matrix:')

fig, ax = plt.subplots(figsize=a4_dims_inv)
p=sns.pairplot(data=df, vars=phy_names, hue='Workload (3-levels)')
xlabels,ylabels = [],[]

xlabels=['pupil size', 'horizontal fixation variability', 'heart rate', 'noseti
p temperature']
ylabels=['pupil size', 'horizontal fixation variability', 'heart rate', 'noseti
p temperature']


for i in range(len(xlabels)):
    for j in range(len(ylabels)):
        p.axes[j,i].xaxis.set_label_text(xlabels[i])
        p.axes[j,i].yaxis.set_label_text(ylabels[j])


p.fig.savefig(os.path.join(picpath, 'Scatter_Means_physio_Condition' + suffix))
plt.close(fig)


fig, ax = plt.subplots(figsize=a4_dims_inv)
p=sns.pairplot(data=df, vars=phy_names, hue='Workload (2-levels)')
xlabels,ylabels = [],[]

xlabels=['pupil size', 'horizontal fixation variability', 'heart rate', 'noseti
p temperature']
ylabels=['pupil size', 'horizontal fixation variability', 'heart rate', 'noseti
p temperature']


for i in range(len(xlabels)):
    for j in range(len(ylabels)):
        p.axes[j,i].xaxis.set_label_text(xlabels[i])
        p.axes[j,i].yaxis.set_label_text(ylabels[j])

p.fig.savefig(os.path.join(picpath, 'Scatter_Means_physio_ConditionBinary' + su
ffix))
plt.close(fig)

plt.show()


# Subset of features
phy_names=['diffpupilsSAtoBLall', 'diffHRsSAtoBLall']

print('scatterplot matrix:')

fig, ax = plt.subplots(figsize=a4_dims_inv)
p=sns.pairplot(data=df, vars=phy_names, hue='Workload (3-levels)')
xlabels,ylabels = [],[]


xlabels=['pupil size','heart rate']
ylabels=['pupil size','heart rate']
```

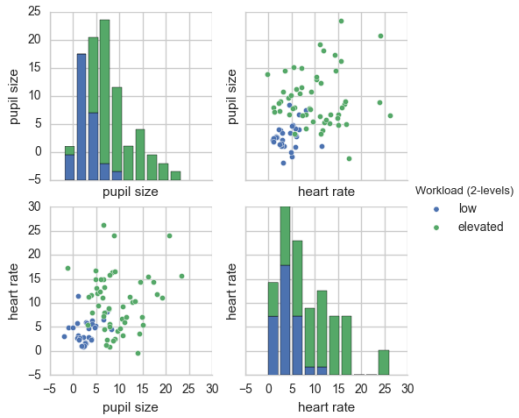
scatterplot matrix:





scatterplot matrix:






```
Correlation matrix:
                        diffpupilsSAtoBLall  difffixSDposxSAtoBLall  \
diffpupilsSAtoBLall                1.000000               -0.405034
difffixSDposxSAtoBLall            -0.405034                1.000000
diffHRsSAtoBLall                   0.353373               -0.205704
diffNttSAtoBLall                  -0.337740                0.152989

                        diffHRsSAtoBLall  diffNttSAtoBLall
diffpupilsSAtoBLall             0.353373         -0.337740
difffixSDposxSAtoBLall         -0.205704          0.152989
diffHRsSAtoBLall                1.000000         -0.283597
diffNttSAtoBLall               -0.283597          1.000000
```


In [4]:

```
# Function definition for exhaustive GridSearch

def extgridsearch(X_train, X_test, y_train, y_test):

    # Set parameters for gridsearchCV

    #SVM Parameters
    tuned_parameters_svm = [{'kernel': ['rbf'], 'gamma': ['auto'],\
                             'C': [1, 10, 100, 1000], 'class_weight': ['balance
d']},\
                            {'kernel': ['linear'], 'C': [1, 10, 100, 1000],'cla
ss_weight': ['balanced']},\
                            {'kernel': ['poly'], 'degree': [2,3,4], 'class_weig
ht': ['balanced']}]


    #KNN parameters
    k=range(1,30)
    weight_options=['uniform', 'distance']
    tuned_parameters_knn=dict(n_neighbors=k, weights=weight_options)

    #Random Forest parameters
    n_trees=range(1,30)
    min_sample_leaf=range(1,20)
    tuned_parameters_rf=dict(n_estimators=n_trees, min_samples_leaf=min_sample_
leaf, class_weight=['balanced'])


    # Define score for best model selection
    scores='accuracy'

    #create a list of models to use
    models = []
    models.append(('KNN', neighbors.KNeighborsClassifier(), tuned_parameters_kn
n))
    models.append(('SVM', SVC(probability=False ,decision_function_shape ='ovr'
), tuned_parameters_svm))
    models.append(('RF', RandomForestClassifier() , tuned_parameters_rf))

    #create empty dict to save best parameters for models
    keys = []
    for i in range(0,len(models)): keys.append(models[i][0])
    param={key: None for key in keys}
    classreport={key: None for key in keys}
    confmatrix={key: None for key in keys}

    for name, model, par in models:
        print(name)
        print(X_train.shape)
        clf=GridSearchCV(model, par, cv=5, scoring=scores)
        # non_nested score
        #clf.fit(X, y)
        clf.fit(X_train, y_train)
        print('NON NESTED SCORES of GRIDSEARCHCV')
        print("Best parameters set found on development set:")
        print()
        print(clf.best_params_)
        param[name]=clf.best_params_
        print()
        print("Detailed classification report:")
        print()
        print("The model is trained on the full trainings development set.")
        print("The scores are computed on the full test evaluation set.")
        print()
        y_true, y_pred = y_test, clf.predict(X_test)
        print(classification_report(y_true, y_pred))
        classreport[name]=classification_report(y_true, y_pred)
        print(confusion_matrix(y_true, y_pred))
        confmatrix[name]=confusion_matrix(y_true, y_pred)
```

In [5]:
```
#Gridsearch for best hyper parameters
#All features - three workload levels

#Feature definition
X=df[['diffpupilsSAtoBLall', 'difffixSDposxSAtoBLall', 'diffHRsSAtoBLall','diff
NttSAtoBLall']]

#define targets
y = df['Workload (3-levels)']

##########################################
# Split data in training (80%) and test (20%)#
# for extensive grid search              #
##########################################
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random
_state=1234,stratify=y)

classreport,confmatrix,param = extgridsearch(X_train, X_test, y_train, y_test)
```



```
KNN
(65, 4)
NON NESTED SCORES of GRIDSEARCHCV
Best parameters set found on development set:

{'n_neighbors': 4, 'weights': 'distance'}

Detailed classification report:

The model is trained on the full trainings development set.
The scores are computed on the full test evaluation set.

             precision    recall  f1-score   support

       high       0.50      0.17      0.25         6
        low       0.83      0.83      0.83         6
     medium       0.44      0.80      0.57         5

avg / total       0.60      0.59      0.55        17

[[1 1 4]
 [0 5 1]
 [1 0 4]]
SVM
(65, 4)
NON NESTED SCORES of GRIDSEARCHCV
Best parameters set found on development set:

{'C': 1000, 'class_weight': 'balanced', 'kernel': 'linear'}

Detailed classification report:

The model is trained on the full trainings development set.
The scores are computed on the full test evaluation set.

             precision    recall  f1-score   support

       high       0.60      0.50      0.55         6
        low       1.00      0.83      0.91         6
     medium       0.43      0.60      0.50         5

avg / total       0.69      0.65      0.66        17

[[3 0 3]
 [0 5 1]
 [2 0 3]]
RF
(65, 4)
NON NESTED SCORES of GRIDSEARCHCV
Best parameters set found on development set:

{'n_estimators': 7, 'class_weight': 'balanced', 'min_samples_leaf': 1}

Detailed classification report:

The model is trained on the full trainings development set.
The scores are computed on the full test evaluation set.

             precision    recall  f1-score   support

       high       0.75      0.50      0.60         6
        low       0.75      1.00      0.86         6
     medium       0.60      0.60      0.60         5

avg / total       0.71      0.71      0.69        17

[[3 1 2]
 [0 6 0]
 [1 1 3]]
```

In [6]:
```
# Use of the three classification approaches with
# their respective best tuning parameters on the whole data set
# Perform crossvalidation on full data set with optimized parameters
# Cross validation with optimized models for all features

# Three workload levels - all features

clf=neighbors.KNeighborsClassifier(n_neighbors=4, weights='distance')
res=cross_val_score(clf,X,y,cv=10)
print('KNN Accuracy','M:',res.mean(),'std:', res.std())


clf=SVC(kernel='linear', C=1000, class_weight='balanced')
res=cross_val_score(clf,X,y,cv=10)
print('SVM Accuracy','M:',res.mean(),'std:', res.std())

clf=RandomForestClassifier(n_estimators=7, min_samples_leaf=1, class_weight='ba
lanced')
res=cross_val_score(clf,X,y,cv=10)
print('RF Accuracy','M:', res.mean(),'std:', res.std())
```

```
KNN Accuracy M: 0.683333333333 std: 0.225137132422
SVM Accuracy M: 0.687103174603 std: 0.138292512122
RF Accuracy M: 0.669047619048 std: 0.173126420923
```


In [7]:
```
#Gridsearch for best hyper parameters
#All features - two workload levels

#Feature definition
X=df[['diffpupilsSAtoBLall', 'difffixSDposxSAtoBLall', 'diffHRsSAtoBLall','diff
NttSAtoBLall']]

#define targets
y = df['Workload (2-levels)']

##########################################
# Split data in training (80%) and test (20%)#
# for extensive grid search              #
##########################################
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random
_state=1235,stratify=y)

classreport,confmatrix,param = extgridsearch(X_train, X_test, y_train, y_test)
```

```
KNN
(65, 4)
NON NESTED SCORES of GRIDSEARCHCV
Best parameters set found on development set:

{'n_neighbors': 3, 'weights': 'uniform'}

Detailed classification report:

The model is trained on the full trainings development set.
The scores are computed on the full test evaluation set.

             precision    recall  f1-score   support

   elevated       1.00      0.91      0.95        11
        low       0.86      1.00      0.92         6

avg / total       0.95      0.94      0.94        17

[[10  1]
 [ 0  6]]
SVM
(65, 4)
NON NESTED SCORES of GRIDSEARCHCV
Best parameters set found on development set:

{'C': 1, 'class_weight': 'balanced', 'kernel': 'linear'}

Detailed classification report:

The model is trained on the full trainings development set.
The scores are computed on the full test evaluation set.

             precision    recall  f1-score   support

   elevated       1.00      1.00      1.00        11
        low       1.00      1.00      1.00         6

avg / total       1.00      1.00      1.00        17

[[11  0]
 [ 0  6]]
RF
(65, 4)
NON NESTED SCORES of GRIDSEARCHCV
Best parameters set found on development set:

{'n_estimators': 10, 'class_weight': 'balanced', 'min_samples_leaf': 3}

Detailed classification report:

The model is trained on the full trainings development set.
The scores are computed on the full test evaluation set.

             precision    recall  f1-score   support

   elevated       1.00      1.00      1.00        11
        low       1.00      1.00      1.00         6

avg / total       1.00      1.00      1.00        17

[[11  0]
 [ 0  6]]

Best tuning parameters for the tested classifiers
{'RF': {'n_estimators': 10, 'class_weight': 'balanced', 'min_samples_leaf': 3}
, 'KNN': {'n_neighbors': 3, 'weights': 'uniform'}, 'SVM': {'C': 1, 'class_weig
ht': 'balanced', 'kernel': 'linear'}}
```


In [8]:
```
# Use of the three classification approaches with
# their respective best tuning parameters on the whole data set
# Perform crossvalidation on full data set with optimized parameters
# Cross validation with optimized models for all features

# Two workload levels - all features

clf=neighbors.KNeighborsClassifier(n_neighbors=3, weights='uniform')
res=cross_val_score(clf,X,y,cv=10)
print('KNN Accuracy','M:',res.mean(),'std:', res.std())


clf=SVC(kernel='linear',C=1, class_weight='balanced')
res=cross_val_score(clf,X,y,cv=10)
print('SVM Accuracy','M:',res.mean(),'std:', res.std())

clf=RandomForestClassifier(n_estimators=10, min_samples_leaf=3,class_weight='ba
lanced')
res=cross_val_score(clf,X,y,cv=10)
print('RF Accuracy','M:', res.mean(),'std:', res.std())
```

```
KNN Accuracy M: 0.877380952381 std: 0.0753539978614
SVM Accuracy M: 0.842658730159 std: 0.0747204941256
RF Accuracy M: 0.880555555556 std: 0.0936238863686
```

In [9]:
```
#Gridsearch for best hyper parameters
#All features - three workload levels

#Feature definition
X=df[['diffpupilsSAtoBLall', 'diffHRsSAtoBLall']]

#define targets
y = df['Workload (3-levels)']

##########################################
# Split data in training (80%) and test (20%)#
# for extensive grid search              #
##########################################
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random
_state=1236,stratify=y)

classreport,confmatrix,param = extgridsearch(X_train, X_test, y_train, y_test)
```



```
KNN
(65, 2)
NON NESTED SCORES of GRIDSEARCHCV
Best parameters set found on development set:

{'n_neighbors': 2, 'weights': 'uniform'}

Detailed classification report:

The model is trained on the full trainings development set.
The scores are computed on the full test evaluation set.

             precision    recall  f1-score   support

       high       0.86      1.00      0.92         6
        low       0.83      0.83      0.83         6
     medium       0.75      0.60      0.67         5

avg / total       0.82      0.82      0.82        17

[[6 0 0]
 [0 5 1]
 [1 1 3]]
SVM
(65, 2)
NON NESTED SCORES of GRIDSEARCHCV
Best parameters set found on development set:

{'C': 1, 'gamma': 'auto', 'class_weight': 'balanced', 'kernel': 'rbf'}

Detailed classification report:

The model is trained on the full trainings development set.
The scores are computed on the full test evaluation set.

             precision    recall  f1-score   support

       high       0.86      1.00      0.92         6
        low       1.00      0.83      0.91         6
     medium       0.80      0.80      0.80         5

avg / total       0.89      0.88      0.88        17

[[6 0 0]
 [0 5 1]
 [1 0 4]]
RF
(65, 2)
NON NESTED SCORES of GRIDSEARCHCV
Best parameters set found on development set:

{'n_estimators': 18, 'class_weight': 'balanced', 'min_samples_leaf': 2}

Detailed classification report:

The model is trained on the full trainings development set.
The scores are computed on the full test evaluation set.

             precision    recall  f1-score   support

       high       1.00      0.67      0.80         6
        low       0.83      0.83      0.83         6
     medium       0.71      1.00      0.83         5

avg / total       0.86      0.82      0.82        17

[[4 1 1]
 [0 5 1]
 [0 0 5]]
```

In [10]:
```
# Use of the three classification approaches with
# their respective best tuning parameters on the whole data set
# Perform crossvalidation on full data set with optimized parameters
# Cross validation with optimized models for all features

# Three workload levels - subset of features

clf=neighbors.KNeighborsClassifier(n_neighbors=2, weights='uniform')
res=cross_val_score(clf,X,y,cv=10)
print('KNN Accuracy','M:',res.mean(),'std:', res.std())


clf=SVC(kernel='rbf',gamma='auto', C=1, class_weight='balanced')
res=cross_val_score(clf,X,y,cv=10)
print('SVM Accuracy','M:',res.mean(),'std:', res.std())

clf=RandomForestClassifier(n_estimators=18, min_samples_leaf=2)
res=cross_val_score(clf,X,y,cv=10)
print('RF Accuracy','M:', res.mean(),'std:', res.std())
```

```
KNN Accuracy M: 0.755158730159 std: 0.091722480088
SVM Accuracy M: 0.759325396825 std: 0.111284763678
RF Accuracy M: 0.7375 std: 0.179854407854
```


In [16]:
```
#Gridsearch for best hyper parameters
#All features - two workload levels

#Feature definition
X=df[['diffpupilsSAtoBLall', 'diffHRsSAtoBLall']]

#define targets
y = df['Workload (2-levels)']

##########################################
# Split data in training (80%) and test (20%)#
# for extensive grid search                  #
##########################################
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random
_state=1236,stratify=y)

classreport,confmatrix,param = extgridsearch(X_train, X_test, y_train, y_test)
```

```
KNN
(65, 2)
NON NESTED SCORES of GRIDSEARCHCV
Best parameters set found on development set:

{'n_neighbors': 21, 'weights': 'uniform'}

Detailed classification report:

The model is trained on the full trainings development set.
The scores are computed on the full test evaluation set.

             precision    recall  f1-score   support

   elevated       1.00      0.91      0.95        11
        low       0.86      1.00      0.92         6

avg / total       0.95      0.94      0.94        17

[[10  1]
 [ 0  6]]
SVM
(65, 2)
NON NESTED SCORES of GRIDSEARCHCV
Best parameters set found on development set:

{'C': 100, 'gamma': 'auto', 'class_weight': 'balanced', 'kernel': 'rbf'}

Detailed classification report:

The model is trained on the full trainings development set.
The scores are computed on the full test evaluation set.

             precision    recall  f1-score   support

   elevated       0.85      1.00      0.92        11
        low       1.00      0.67      0.80         6

avg / total       0.90      0.88      0.88        17

[[11  0]
 [ 2  4]]
RF
(65, 2)
NON NESTED SCORES of GRIDSEARCHCV
Best parameters set found on development set:

{'n_estimators': 14, 'class_weight': 'balanced', 'min_samples_leaf': 14}

Detailed classification report:

The model is trained on the full trainings development set.
The scores are computed on the full test evaluation set.

             precision    recall  f1-score   support

   elevated       1.00      1.00      1.00        11
        low       1.00      1.00      1.00         6

avg / total       1.00      1.00      1.00        17

[[11  0]
 [ 0  6]]

Best tuning parameters for the tested classifiers
{'RF': {'n_estimators': 14, 'class_weight': 'balanced', 'min_samples_leaf': 14
}, 'KNN': {'n_neighbors': 21, 'weights': 'uniform'}, 'SVM': {'C': 100, 'gamma'
: 'auto', 'class_weight': 'balanced', 'kernel': 'rbf'}}
```


In [17]:
```
# Use of the three classification approaches with
# their respective best tuning parameters on the whole data set
# Perform crossvalidation on full data set with optimized parameters
# Cross validation with optimized models for all features

# Two workload levels - subset of features

clf=neighbors.KNeighborsClassifier(n_neighbors=21, weights='uniform')
res=cross_val_score(clf,X,y,cv=10)
print('KNN Accuracy','M:',res.mean(),'std:', res.std())


clf=SVC(kernel='rbf', gamma='auto',C=100, class_weight='balanced')
res=cross_val_score(clf,X,y,cv=10)
print('SVM Accuracy','M:',res.mean(),'std:', res.std())

clf=RandomForestClassifier(n_estimators=14, min_samples_leaf=14, class_weight='
balanced')
res=cross_val_score(clf,X,y,cv=10)
print('RF Accuracy','M:', res.mean(),'std:', res.std())
```

```
KNN Accuracy M: 0.904166666667 std: 0.0738727008165
SVM Accuracy M: 0.827777777778 std: 0.140298403324
RF Accuracy M: 0.918055555556 std: 0.0733485884806
```

In [29]:
```
import sklearn as sk
import scipy as sci
import matplotlib as plt
import pandas as pd

print('sci-kit learn',sk.__version__)
print('seaborn',sns.__version__)
print('matplotlib',plt.__version__)
print('scipy',sci.__version__)
print('pandas',pd.__version__)
```

```
sci-kit learn 0.18.1
seaborn 0.7.1
matplotlib 1.5.3
scipy 0.18.1
pandas 0.18.1
```

In [ ]:

Bibliography

Aasman, J., Mulder, G., & Mulder, L.J.M. (1987). Operator effort and the measurement of heartrate variability. *Human Factors, 29(2), 161-170.*.

Allanson, J. (2002). Electrophysiologically interactive computer systems. *Computer*, *35*(3), 60–65.

Allanson, J., & Fairclough, S. H. (2004). A research agenda for physiological computing. *Interacting with computers*, *16*(5), 857–878.

Alm, H., & Nilsson, L. (1994). Changes in driver behaviour as a function of handsfree mobile phonesa simulator study. *Accident Analysis & Prevention*, *26*(4), 441–451.

Alm, H., & Nilsson, L. (1995). The effects of a mobile telephone task on driver behaviour in a car following situation. *Accident Analysis & Prevention*, *27*(5), 707–715.

Alter, K. (2002). Suprasegmentale merkmale und prosodie. In *Arbeitsbuch linguistik* (pp. 148–169). UTB-Ferdinand Schöningh.

AmericanHeartAssociation. (2015). *Target heart rates.* `http://www.heart.org/HEARTORG/HealthyLiving/PhysicalActivity/FitnessBasics/Target-Heart-Rates_UCM_434341_Article.jsp\#.WO-ZDhJ94yk`. ((Accessed: 2017-02-17))

Arnold, H. S., MacPherson, M. K., & Smith, A. (2014). Autonomic correlates of speech versus nonspeech tasks in children and adults. *Journal of Speech, Language, and Hearing Research*, *57*(4), 1296–1307.

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *Neuroimage*, *59*(1), 36–47.

Backs, R. W. (1995). Going beyond heart rate: autonomic space and cardiovascular assessment of mental workload. *The international journal of aviation psychology*, *5*(1), 25–48.

Baddeley, A. D. (1992). Working memory. *Science*, *255*(5044), 556.

Baddeley, A. D. (2003). Working memory: looking back and looking forward. *Nature reviews. Neuroscience*, *4*(10), 829–839. doi: 10.1038/nrn1201

Baddeley, A. D., Bressi, S., Della Sala, S., Logie, R., & Spinnler, H. (1991). The decline of working memory in alzheimer's disease. *Brain*, *114*(6), 2521–2542.

Baddeley, A. D., Chincotta, D., & Adlam, A. (2001). Working memory and the control of action: Evidence from task switching. *Journal of Experimental Psychology: General*, *130*(4), 641–657. doi: 10.1037/0096-3445.130.4.641

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of learning and motivation*, *8*, 47–89.

Baken, R. J., & Orlikoff, R. F. (2000). *Clinical measurement of speech and voice* - (Revised. ed.). Clifton Park, NY: Cengage Learning.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276–292. doi: 10.1037/0033-2909.91.2.276

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), (chap. 6). Cambridge University Press.

Beda, A., Jandre, F. C., Phillips, D. I., Giannella-Neto, A., & Simpson, D. M. (2007). Heart-rate and blood-pressure variability during psychophysiological tasks involving speech: Influence of respiration. *Psychophysiology*, *44*(5), 767–778.

Bednarik, R., Kinnunen, T., Mihaila, A., & Fränti, P. (2005). Eye-movements as a biometric. *Image analysis*, 16–26.

Berntson, G. G., Bigger Jr, J., Eckberg, D., Grossman, P., Kaufmann, P., Malik, M., ... others (1997). Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology*, *34*(6), 623–648.

Berntson, G. G., Cacioppo, J. T., Binkley, P. F., Uchino, B. N., Quigley, K. S., & Fieldstone, A. (1994). Autonomic cardiac control. iii. psychological stress and cardiac response in autonomic space as revealed by pharmacological blockades. *Psychophysiology*, *31*(6), 599–608.

Berntson, G. G., Cacioppo, J. T., & Quigley, K. S. (1991). Autonomic determinism: the modes of autonomic control, the doctrine of autonomic space, and the laws of autonomic constraint. *Psychological review*, *98*(4), 459.

Berntson, G. G., Cacioppo, J. T., & Quigley, K. S. (1993). Cardiac psychophysiology and autonomic space in humans: empirical perspectives and conceptual implications. *Psychological bulletin*, *114*(2), 296.

Biel, L., Pettersson, O., Philipson, L., & Wide, P. (2001). Ecg analysis: a new approach in human identification. *IEEE Transactions on Instrumentation and Measurement*, *50*(3), 808–812.

Billman, G. E. (2011). Heart rate variability–a historical perspective. *Frontiers in Physiology*, *2*, 86.

Billman, G. E. (2013). The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance. *Frontiers in Physiology*, *4*. Retrieved from `http://dx.doi.org/10.3389/fphys.2013.00026` doi: 10.3389/fphys.2013.00026

Boersma, P. P. G. (2002). Praat, a system for doing phonetics by computer. *Glot international*, *5*(9/10), 341–345.

Böhle, K., Coenen, C., Decker, M., & Rader, M. (2013). Biocybernetic adaptation and privacy. *Innovation: The European Journal of Social Science Research*, *26*(1-2), 71–80.

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, *44*, 58–75.

Bořil, H., Kleinschmidt, T., Boyraz, P., & Hansen, J. H. (2010). Impact of cognitive load and frustration on drivers speech. *The Journal of the Acoustical Society of America*, *127*(3), 1996–1996.

Bořil, H., Omid Sadjadi, S., Kleinschmidt, T., & Hansen, J. H. (2010). Analysis and detection of cognitive load and frustration in drivers' speech. *Proceedings of INTERSPEECH 2010*, 502–505.

Bortz, J. (2005). Formulierung und überprüfung von hypothesen. In *Statistik: für human- und sozialwissenschaftler* (pp. 107–133). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/3-540-26430-2_4

Bortz, J., & Döring, N. (2007). *Forschungsmethoden und evaluation für human-und sozialwissenschaftler: Limitierte sonderausgabe*. Springer-Verlag.

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607.

Brajkovic, D., & Ducharme, M. B. (2006). Facial cold-induced vasodilation and skin temperature during exposure to cold wind. *European journal of applied physiology*, *96*(6), 711–721.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Briggs, G. F., Hole, G. J., & Land, M. F. (2016). Imagery-inducing distraction leads to cognitive tunnelling and deteriorated driving performance. *Transportation research part F: traffic psychology and behaviour*, *38*, 106–117.

Brookhuis, K. A. (2004, aug). Psychophysiological methods. In *Handbook of human factors and ergonomics methods* (pp. 17–1–17–5). Informa UK Limited. Retrieved from `http://dx.doi.org/10.1201/9780203489925.ch17` doi: 10.1201/9780203489925.ch17

Brookhuis, K. A., de Waard, D., & Samyn, N. (2004). Effects of mdma (ecstasy), and multiple drugs use on (simulated) driving performance and traffic safety. *Psychopharmacology*, *173*(3-4), 440–445.

Brookhuis, K. A., Vries, G. d., & Waard, D. d. (1991). The effects of mobile telephoning on driving performance. *Accident Analysis & Prevention*, *23*(4), 309–316. doi: 10.1016/0001-4575(91)90008-S

Brookhuis, K. A., & Waard, D. d. (2010). Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident Analysis & Prevention*, *42*(3), 898–903. doi: 10.1016/j.aap.2009.06.001

Brosschot, J. F., & Thayer, J. F. (2003). Heart rate response is longer after negative emotions than after positive emotions. *International Journal of Psychophysiology*, *50*(3), 181–187.

Cain, B. (2007). *A review of the mental workload literature* (Tech. Rep.). New York: DTIC Document.

Caird, J. K., Johnston, K. A., Willness, C. R., Asbridge, M., & Steel, P. (2014). A meta-analysis of the effects of texting on driving. *Accident Analysis & Prevention*, *71*, 311–318.

Caird, J. K., Willness, C. R., Steel, P., & Scialfa, C. (2008). A meta-analysis of the effects of cell phones on driver performance. *Accident Analysis & Prevention*,

40(4), 1282–1293. doi: 10.1016/j.aap.2008.01.009

Cantin, V., Lavallière, M., Simoneau, M., & Teasdale, N. (2009). Mental workload when driving in a simulator: Effects of age and driving complexity. *Accident Analysis & Prevention*, *41*(4), 763–771.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning* (pp. 161–168).

Chen, L.-l., Zhao, Y., Zhang, J., & Zou, J.-z. (2015). Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning. *Expert Systems with Applications*, *42*(21), 7344–7355.

Collet, C., Clarion, A., Morel, M., Chapon, A., & Petit, C. (2009). Physiological and behavioural changes associated to the management of secondary tasks while driving. *Applied ergonomics*, *40*(6), 1041–1046.

Consortium, A. (2004-2008). *Aide - adaptive integrated driver-vehicle interface, fraunhofer iao.* `http://www.aide-eu.org/`. ((Accessed on 04/14/2017))

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (Vol. 351). Houghton Mifflin Boston.

Cools, R. (2010). N-back test. In *Encyclopedia of psychopharmacology* (pp. 822–822). Springer.

Cooper, J. M., Medeiros-Ward, N., & Strayer, D. L. (2013). The impact of eye movements and cognitive workload on lateral position variability in driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *55*(5), 1001–1014.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21–27.

Dasarathy, B. V. (1990). *Nearest neighbor (NN) norms: NN pattern classification techniques*. Los Alamitos, CA: IEEE Computer Society Press.

De Waard, D., & Brookhuis, K. (1997). On the measurement of driver mental workload. In J. Rothengatter & E. Carbonell Vaya (Eds.), *Traffic and transport psychology* (pp. 161 – 171). Pergamon Press.

Destatis. (2017, 3). *Fachserie verkehr, verkehrsunfälle* (Tech. Rep. Nos. Fachserie 8, Reihe 7). Wiesbaden: Statistisches Bundesamt. (An optional note)

Deubel, H. (2008). The time course of presaccadic attention shifts. *Psychological research*, *72*(6), 630–640.

De Waard, D. (1996). *The measurement of drivers' mental workload*. Groningen University, Traffic Research Center Netherlands.

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, 201513271.

Di Nocera, F., Camilli, M., & Terenzi, M. (2007). A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload.

Journal of Cognitive Engineering and Decision Making, *1*(3), 271–285.

Dong, Y., Hu, Z., Uchimura, K., & Murayama, N. (2011). Driver inattention monitoring system for intelligent vehicles: A review. *IEEE transactions on intelligent transportation systems*, *12*(2), 596–614.

Drews, F. A., Pasupathi, M., & Strayer, D. L. (2008). Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied*, *14*(4), 392.

Drews, F. A., Yazdani, H., Godfrey, C. N., Cooper, J. M., & Strayer, D. L. (2009). Text messaging during simulated driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*.

Dromey, C., & Bates, E. (2005). Speech interactions with linguistic, cognitive, and visuomotor tasks. *Journal of Speech, Language, and Hearing Research*, *48*(2), 295–305.

Drummond, P. D. (1994). Sweating and vascular responses in the face: Normal regulation and dysfunction in migraine, cluster headache and harlequin syndrome. *Clinical Autonomic Research*, *4*(5), 273–285. doi: 10.1007/BF01827433

Duchowski, A. T., Cournia, N., & Murphy, H. (2004). Gaze-contingent displays: A review. *CyberPsychology & Behavior*, *7*(6), 621–634.

Duffy, V. G. (2007). *Digital human modeling: First international conference on digital human modeling, icdhm 2007, held as part of hci international 2007, beijing, china, july 22-27, 2007 : proceedings*. Berlin and and New York: Springer.

Eggemeier, F. T., Wilson, G. F., Kramer, A. F., & Damos, D. L. (1991). Workload assessment in multi-task environments. *Multiple-task performance*, 207–216.

Eilers, K., Nachreiner, F., & Hänecke, K. (1986). Entwicklung und überprüfung einer skala zur erfassung subjektiv erlebter anstrengung. *Zeitschrift für Arbeitswissenschaft*(4), 214–224.

Engström, J., Aust, M. L., & Viström, M. (2010). Effects of working memory load and repeated scenario exposure on emergency braking performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(5), 551–559.

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*(2), 97–120.

Ephrath, A., Tole, J., Stephens, A. T., & Young, L. (1980). Instrument scanis it an indicator of the pilot's workload? In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 24, pp. 257–258).

Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting with computers*, *21*(1), 133–145.

Fairclough, S. H. (2010). Physiological computing: interfacing with the human nervous system. In *Sensing emotions* (pp. 1–20). Springer.

Fairclough, S. H. (2014). Physiological data must remain confidential. *Nature*, *505*(7483), 263.

Fairclough, S. H., & Houston, K. (2004). A metabolic measure of mental effort. *Biological psychology*, *66*(2), 177–190.

Fairclough, S. H., & Mulder, L. (2011). Psychophysiological processes of mental effort investment. *How motivation affects cardiovascular response: Mechanisms and applications*, 61–76.

Ferdinand, A. O., & Menachemi, N. (2014). Associations between driving performance and engaging in secondary tasks: a systematic review. *American journal of public health*, *104*(3), e39–e48.

Fernandez, R., & Picard, R. W. (2003). Modeling drivers speech under stress. *Speech communication*, *40*(1), 145–159.

Flohr, H. (2002). Grundbegriffe der phonetik. In *Arbeitsbuch linguistik* (pp. 47–75). UTB-Ferdinand Schöningh.

Fodor, I. K. (2002). A survey of dimension reduction techniques. *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory*, *9*, 1–18.

Fuller, R. (2005). Towards a general theory of driver behaviour. *Accident Analysis & Prevention*, *37*(3), 461–472.

Gavhed, D., Mäkinen, T., Holmér, I., & Rintamäki, H. (2000). Face temperature and cardiorespiratory responses to wind in thermoneutral and cool subjects exposed to - 10°c. *European journal of applied physiology*, *83*(4), 449–456.

Genno, H., Ishikawa, K., Kanbara, O., Kikumoto, M., Fujiwara, Y., Suzuki, R., & Osumi, M. (1997). Using facial skin temperature to objectively evaluate sensations. *International Journal of Industrial Ergonomics*, *19*(2), 161–171.

Giddens, C. L., Barron, K. W., Byrd-Craven, J., Clark, K. F., & Winter, A. S. (2013). Vocal indices of stress: a review. *Journal of Voice*, *27*(3), 390–e21.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of educational research*, *42*(3), 237–288.

Gopher, D. (1994). Analysis and measurement of mental load. *International Perspectives on Psychological Science: The state of the art: state of the art lectures*, *2*, 265.

Gramann, K., & Schandry, R. (2009). *Psychophysiologie: Körperliche indikatoren psychischen geschehens* (4., vollst. überarb. Aufl ed.). Weinheim and Basel: Beltz, PVU.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature neuroscience*, *6*(3), 316–322.

Gronwall, D. M. (1977). Paced auditory serial-addition task: a measure of recovery from concussion. *Perceptual and motor skills*, *44*(2), 367–373.

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2008). *Feature extraction: foundations and applications* (Vol. 207). Springer.

Haigney, D., Taylor, R., & Westerman, S. (2000). Concurrent mobile (cellular) phone use and driving performance: task demand characteristics and compensatory processes. *Transportation Research Part F: Traffic Psychology and Behaviour*, *3*(3), 113–121.

Hammerschmidt, O. (2013). *Eignung der Nasentemperatur als physiologischer Parameter für die Erfassung unterschiedlicher Beanspruchungsniveaus beim Führen eines Fahrzeuges* (B.S. Thesis). Technische Universität Berlin, Germany.

Hancock, P., & Verwey, W. B. (1997). Fatigue, workload and adaptive driver systems. *Accident Analysis & Prevention*, *29*(4), 495–506.

Hansen, D. W., & Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on pattern analysis and machine intelligence*, *32*(3), 478–500.

Harbluk, J. L., Burns, P. C., Lochner, M., & Trbovich, P. L. (2007). Using the lane-change test (lct) to assess distraction: Tests of visual-manual and speech-based operation of navigation system interfaces. In *Proceedings of the 4th international driving symposium on human factors in driver assessment, training, and vehicle design* (pp. 16–22).

Harbluk, J. L., & Lalande, S. (2005). Performing e-mail tasks while driving: The impact of speech-based tasks on visual detection. In *Proceedings of the 3rd international driving symposium on human factors in driving assessment, training and vehicle design* (pp. 304–310).

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (pp. 139–183). Elsevier. doi: 10.1016/S0166-4115(08)62386-9

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing monte carlo results in methodological research: The one-and two-factor fixed effects anova cases. *Journal of educational statistics*, *17*(4), 315–339.

Hayter, A. J. (1984). A proof of the conjecture that the tukey-kramer multiple comparisons procedure is conservative. *The Annals of Statistics*, 61–75.

He, J., Chaparro, A., Nguyen, B., Burge, R. J., Crandall, J., Chaparro, B., ... Cao, S. (2014). Texting while driving: Is speech-based text entry less risky than handheld text entry? *Accident Analysis & Prevention*, *72*, 287–295.

He, J., McCarley, J. S., & Kramer, A. F. (2013). Lane keeping under cognitive load performance changes and mechanisms. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 0018720813485978.

Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, *6*(2), 156–166.

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132*(3423), 349–350.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190–1192. doi: 10.1126/science.143.3611.1190

Hettinger, L. J., Branco, P., Encarnacao, L. M., & Bonato, P. (2003). Neuroadaptive technologies: applying neuroergonomics to the design of advanced interfaces. *Theoretical Issues in Ergonomics Science*, *4*(1-2), 220–237.

Ho, T. K. (1995). Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on* (Vol. 1, pp. 278–282).

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.

Hockey, G. R. J., Gaillard, A. W. K., & Coles, M. G. H. (1986). *Energetics and human information processing*. Dordrecht and Boston: Nijhoff.

Hollnagel, E., Nåbo, A., & Lau, I. (2003). A systemic model for driver-in-control. In *second international driving symposium on human factors in driver assessment, training and vehicle design* (pp. 86–91).

Holloway, G. A., & Watkins, D. W. (1977). Laser doppler measurement of cutaneous blood flow. *Journal of Investigative Dermatology*, *69*(3), 306–309.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Horberry, T., Anderson, J., Regan, M. A., Triggs, T. J., & Brown, J. (2006). Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. *Accident Analysis & Prevention*, *38*(1), 185–191.

Horrey, W. J. (2011). *Assessing the effects of in-vehicle tasks on driving performance* (No. 4). Retrieved 31.01.2012, from `http://erg.sagepub.com/content/19/4/4.full.pdf` doi: 10.1177/1064804611419961

Horrey, W. J., Lesch, M. F., & Garabet, A. (2008). Assessing the awareness of performance decrements in distracted drivers. *Accident Analysis & Prevention*, *40*(2), 675–682.

Horrey, W. J., Lesch, M. F., & Garabet, A. (2009). Dissociation between driving performance and drivers' subjective estimates of performance and workload in dual-task conditions. *Journal of safety research*, *40*(1), 7–12. doi: 10.1016/j.jsr.2008.10.011

Horrey, W. J., & Wickens, C. D. (2006). Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(1), 196–205.

Horrey, W. J., Wickens, C. D., & Consalus, K. P. (2006). Modeling drivers' visual attention allocation while interacting with in-vehicle technologies. *Journal of Experimental Psychology: Applied*, *12*(2), 67.

Hosking, S. G., Young, K. L., & Regan, M. A. (2009). The effects of text messaging on young drivers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*(4), 582–592.

Hu, S., & Zheng, G. (2009). Driver drowsiness detection with eyelid related parameters by support vector machine. *Expert Systems with Applications*, *36*(4), 7651–7658.

Hubber, P. J., Gilmore, C., & Cragg, L. (2014). The roles of the central executive and visuospatial storage in mental arithmetic: A comparison across strategies. *The Quarterly Journal of Experimental Psychology*, *67*(5), 936–954.

Huemer, A. K., & Vollrath, M. (2011). Driver secondary tasks in germany: Using interviews to estimate prevalence. *Accident Analysis & Prevention*, *43*(5), 1703–1712.

Hurts, K., Angell, L. S., & Perez, M. A. (2011). The distracted driver mechanisms, models, and measurement. *Reviews of human factors and ergonomics*, *7*(1), 3–57.

Huttunen, K., Keränen, H., Väyrynen, E., Pääkkönen, R., & Leino, T. (2011). Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights. *Applied Ergonomics*, *42*(2), 348–357. doi: 10.1016/j.apergo .2010.08.005

Iani, C., Gopher, D., & Lavie, P. (2004). Effects of task difficulty and invested mental effort on peripheral vasoconstriction. *Psychophysiology*, *41*(5), 789–798. doi: 10.1111/j.1469-8986.2004.00200.x

Itoh, M. (2009). Individual differences in effects of secondary cognitive activity during driving on temperature at the nose tip. In *Mechatronics and automation, 2009. icma 2009. international conference on* (pp. 7–11). Retrieved from `http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5246188`

Itoh, M., & Inagaki, T. (2008). Effects of non-driving cognitive activity on driver's eye movement and their individual difference. *Journal of Mechanical Systems for Transportation and Logistics*, *1*(2), 203–212. doi: 10.1299/jmtl.1.203

Jacucci, G., Fairclough, S., & Solovey, E. T. (2015). Physiological computing. *Computer*, *48*(10), 12–16.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the n-back task as a working memory measure. *Memory*, *18*(4), 394–412.

Jamson, A. H., Westerman, S. J., Hockey, G. R. J., & Carsten, O. M. (2004). Speech-based e-mail and driver behavior: Effects of an in-vehicle message system interface. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(4), 625–639.

Ji, Q., Zhu, Z., & Lan, P. (2004). Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE transactions on vehicular technology*, *53*(4), 1052–1068.

Jin, L., Niu, Q., Hou, H., Xian, H., Wang, Y., & Shi, D. (2012). Driver cognitive distraction detection using driving performance measures. *Discrete Dynamics in Nature and Society*, *2012*.

Jorna, P. G. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological psychology*, *34*(2), 237–257.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, *87*(4), 329.

Just, M. A., Carpenter, P. A., Keller, T. A., Emery, L., Zajac, H., & Thulborn, K. R. (2001). Interdependence of nonoverlapping cortical systems in dual cognitive tasks. *NeuroImage*, *14*(2), 417–426.

Just, M. A., Keller, T. A., & Cynkar, J. (2008). A decrease in brain activation associated with driving when listening to someone speak. *Brain research*, *1205*, 70–80.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs and N.J: Prentice-Hall.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583–1585. doi: 10.1126/science.154.3756.1583

Kahneman, D., Tursky, B., Shapiro, D., & Crider, A. (1969). Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of experimental psychology*, *79*(1p1), 164.

Kaku, M. (2012). *Physics of the future: How science will shape human destiny and our daily lives by the year 2100.* Anchor.

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the n-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615.

Kang, J., & Babski-Reeves, K. (2008). Detecting mental workload fluctuation during learning of a novel task using thermography. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 52, pp. 1527–1531).

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, *3*(3), 203–220.

Kim, K. H., Bang, S. W., & Kim, S. R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing*, *42*(3), 419–427.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, *55*(4), 352.

Klauer, S. G., Guo, F., Simons-Morton, B. G., Ouimet, M. C., Lee, S. E., & Dingus, T. A. (2014). Distracted driving and risk of road crashes among novice and experienced drivers. *New England journal of medicine*, *370*(1), 54–59.

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, *48*(3), 323–332. doi: 10.1111/j.1469-8986.2010.01069.x

Klinke, R., & Bauer, C. (1996). *Lehrbuch der physiologie: 52 tabellen.* Thieme.

Knoll, O. (2013). *Eignung der Herzrate und Herzratenvariabilität als Indikator für die mentale Beanspruchung beim Autofahren* (B.S. Thesis). Technische Universität Berlin, Germany.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137–1145).

Kotsiantis, S. (2007). Supervised machine learning: a review of classification techniques. *Informatica*, *31*(3), 249–269.

Kramer, A. E. (1991). Physiological metrics of mental workload: a review of recent progress. *In D. L. Damos (Ed.), Multiple-task performance (pp. 279-328). London: Taylor & Francis.*

Kutila, M., Jokela, M., Mäkinen, T., Viitanen, J., Markkula, G., & Victor, T. (2007). Driver cognitive distraction detection: Feature estimation and implementation. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, *221*(9), 1027–1040.

Lacey, J. I. (1959). Psychophysiological approaches to the evaluation of psychotherapeutic process and outcome. In *Research in psychotherapy.*

Lacey, J. I., Kagan, J., Lacey, B. C., & Moss, H. (1963). The visceral level: Situational determinants and behavioral correlates of autonomic response patterns. *Expression of the emotions in man*, *9*.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry a window to the preconscious? *Perspectives on psychological science*, *7*(1), 18–27.

Lau, M. K. (2009). Dtk: Dunnett–tukey–kramer pairwise multiple comparison test adjusted for unequal variances and unequal sample sizes. *R package version*, *2*.

Lawrence, M. A. (2015). ez: Easy analysis and visualization of factorial experiments [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=ez` (R package version 4.3)

Lee, J. D. (2014). Dynamics of driver distraction: the process of engaging and disengaging. *Annals of advances in automotive medicine*, *58*, 24.

Lee, J. D., Caven, B., Haake, S., & Brown, T. L. (2001). Speech-based interaction with in-vehicle computers: The effect of speech-based e-mail on drivers' attention to the roadway. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(4), 631–640. doi: 10.1518/001872001775870340

Lee, J. D., Young, K. L., & Regan, M. A. (2008). Defining driver distraction. *Driver distraction: Theory, effects, and mitigation*, *13*(4), 31–40.

Lei, S., Toriizuka, T., & Roetting, M. (2017). Driver adaptive task allocation: A field driving study. *Le travail humain*, *80*(1), 93–112.

Leibowitz, H. W., & Dichgans, J. (1980). The ambient visual system and spatial orientation. *Spatial Disorientation in Flight: Current Problems, Perdriel, G., and Benson, AJ,(editors), AGARD CP-287*.

Leibowitz, H. W., & Post, R. B. (1982). The two modes of processing concept and some implications. *Organization and representation in perception*, 343–363.

Lenneman, J. K., & Backs, R. (2009). Cardiac autonomic control during simulated driving with a concurrent verbal working memory task. *Human Factors, 53(3), 404-418.*.

Lenneman, J. K., Shelley, J. R., & Backs, R. W. (2005). Deciphering psychological-physiological mappings during a simulated driving task. *In Proceedings of the Third International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design (pp. 493–498). Iowa City: University of Iowa.*.

Lesch, M. F., & Hancock, P. A. (2004). Driving performance during concurrent cell-phone use: are drivers aware of their performance decrements? *Accident Analysis & Prevention*, *36*(3), 471–480.

Liang, Y., Lee, J., & Reyes, M. (2007). Nonintrusive detection of driver cognitive distraction in real time using bayesian networks. *Transportation Research Record: Journal of the Transportation Research Board*(2018), 1–8.

Liang, Y., & Lee, J. D. (2010). Combining cognitive and visual distraction: Less than the sum of its parts. *Accident Analysis & Prevention*, *42*(3), 881–890.

Liang, Y., & Lee, J. D. (2014). A hybrid bayesian network approach to detect driver cognitive distraction. *Transportation research part C: emerging technologies*, *38*, 146–155.

Liang, Y., Reyes, M. L., & Lee, J. D. (2007). Real-time detection of driver cognitive distraction using support vector machines. *IEEE transactions on intelligent transportation systems*, *8*(2), 340–350.

Liao, Y., Li, S. E., Wang, W., Wang, Y., Li, G., & Cheng, B. (2016). Detection of driver cognitive distraction: A comparison study of stop-controlled intersection

and speed-limited highway. *IEEE Transactions on Intelligent Transportation Systems*, *17*(6), 1628–1637.

Linden, W. (1987). A microanalysis of autonomic activity during human speech. *Psychosomatic Medicine*, *49*(6), 562–578.

Linguatech. (2016). *Text-to-speech demo voice reader 15.* Retrieved 2016-06-03, from `http://www.linguatec.de/text-to-speech/demo/`

Liu, B.-S., & Lee, Y.-H. (2006). In-vehicle workload assessment: effects of traffic situations and cellular telephone use. *Journal of safety research*, *37*(1), 99–105.

Lively, S. E. (1993). Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *The Journal of the Acoustical Society of America*, *93*(5), 2962. doi: 10.1121/1.405815

Lix, L. M., Keselman, J. C., & Keselman, H. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance f test. *Review of educational research*, *66*(4), 579–619.

Loewenfeld, I. E., & Lowenstein, O. (1993). *The pupil: Anatomy, physiology, and clinical applications* (Vol. 2). Wiley-Blackwell.

Maciej, J., & Vollrath, M. (2009). Comparison of manual vs. speech-based interaction with in-vehicle information systems. *Accident Analysis & Prevention*, *41*(5), 924–930. doi: 10.1016/j.aap.2009.05.007

MacPherson, M. K., Abur, D., & Stepp, C. E. (2016). Acoustic measures of voice and physiologic measures of autonomic arousal during speech as a function of cognitive load. *Journal of Voice*.

Manzey, D. (1988). *Determinanten der aufgabeninterferenz bei doppeltátigkeiten und ressourcentheoretische modellvorstellungen in der kognitiven psychologie.* (Tech. Rep. No. DFVLR-FB 88-14). Kóln: Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt DFVLR.

Marshall, S. P. (2002). The index of cognitive activity: Measuring cognitive workload. In *Human factors and power plants, 2002. proceedings of the 2002 ieee 7th conference on* (pp. 7–5).

Mattes, S. (2003). The lane-change-task as a tool for driver distraction evaluation. *Quality of Work and Products in Enterprises of the Future*, *2003*, 57.

Mattes, S., & Hallén, A. (2008, oct). Surrogate distraction measurement techniques. In *Driver distraction: Theory, effects, and mitigation* (pp. 107–122). Informa UK Limited.

Matthews, G., & Campbell, S. E. (2009). Sustained performance under overload: personality and individual differences in stress and coping. *Theoretical Issues in Ergonomics Science*, *10*(5), 417–442.

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(1), 125–143. doi: 10.1177/0018720814539505

Matthews, R., Legg, S., & Charlton, S. (2003). The effect of cell phone type on drivers subjective workload during concurrent driving and conversing. *Accident Analysis & Prevention*, *35*(4), 451–457.

May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye movement indices of mental workload. *Acta psychologica*, *75*(1), 75–89.

McEvoy, S. P. (2015). The role of mobile phones in real world motor vehicle crashes. In *Encyclopedia of mobile phone behavior* (pp. 1366–1375). IGI Global. Retrieved from `https://doi.org/10.4018/978-1-4666-8239-9.ch108` doi: 10.4018/978-1-4666-8239-9.ch108

McKee, S. P., & Nakayama, K. (1984). The detection of motion in the peripheral visual field. *Vision research*, *24*(1), 25–32.

Mehler, B., Reimer, B., Coughlin, J., & Dusek, J. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board*(2138), 6–12.

Mehler, B., Reimer, B., & Coughlin, J. F. (2012). Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task an on-road study across three age groups. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(3), 396–412.

Mehler, B., Reimer, B., & Dusek, J. (2011). Mit agelab delayed digit recall task (n-back). *Cambridge, MA: Massachusetts Institute of Technology*.

Mendoza, E., & Carballo, G. (1998). Acoustic analysis of induced vocal stressby means of cognitive workload tasks. *Journal of Voice*, *12*(3), 263–273.

Meyer, D., & Wien, F. T. (2015). Support vector machines. *The Interface to libsvm in package e1071*.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

Michon, J. A. (1979). *Dealing with danger* (Tech. Rep.). Rijksuniversiteit Groningen Haren, The Netherlands: Verkeerskundig Studiecentrum.

Michon, J. A. (1985). A critical view of driver behavior models: what do we know, what should we do? In *Human behavior and traffic safety* (pp. 485–524). Springer.

Min, B.-C., Chung, S.-C., Min, Y.-K., & Sakamoto, K. (2004). Psychophysiological evaluation of simulator sickness evoked by a graphic simulator. *Applied ergonomics*, *35*(6), 549–556.

Mitchell, T. M. (2006). *The discipline of machine learning* (Vol. 9). Carnegie Mellon University, School of Computer Science, Machine Learning Department.

Miyaji, M., Danno, M., Kawanaka, H., & Oguri, K. (2008). Drivers cognitive distraction detection using adaboost on pattern recognition basis. In *Vehicular electronics and safety, 2008. icves 2008. ieee international conference on* (pp. 51–56).

Miyake, S., Yamada, S., Shoji, T., Takae, Y., Kuge, N., & Yamamura, T. (2009). Physiological responses to workload change. a test/retest examination. *Applied ergonomics*, *40*(6), 987–996.

Montag, C., Błaszkiewicz, K., Sariyska, R., Lachmann, B., Andone, I., Trendafilov, B., ... Markowetz, A. (2015). Smartphone usage in the 21st century: who is active on whatsapp? *BMC research notes*, *8*(1), 331.

Moray, N. (1967). Where is capacity limited? a survey and a model. *Acta psychologica*, *27*, 84–92.

Mulder, G. (1986). The concept and measurement of mental effort. In *Energetics and human information processing* (pp. 175–198). Springer.

Mulder, L. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological psychology*, *34*(2), 205–236.

Mulder, L. B., de Waard, D., & Brookhuis, K. A. (2004, aug). Estimating mental effort using heart rate and heart rate variability. In *Handbook of human factors and ergonomics methods* (pp. 20–1–20–8). CRC Press. Retrieved from `https://doi.org/10.1201/9780203489925.ch20` doi: 10.1201/9780203489925.ch20

Murai, K., Hayashi, Y., & Inokuchi, S. (2004). A basic study on teammates' mental workload among ship's bridge team. *IEICE TRANSACTIONS on Information and Systems*, *87*(6), 1477–1483.

Murai, K., Hayashi, Y., Okazaki, T., Stone, L. C., & Mitomo, N. (2008). Evaluation of ship navigator's mental workload using nasal temperature and heart rate variability. In *Systems, man and cybernetics, 2008. smc 2008. ieee international conference on* (pp. 1528–1533).

Murai, K., Okazaki, T., Stone, L. C., & Hayashi, Y. (2007). A characteristic of a navigators mental workload based on nasal temperature. In *2007 ieee international conference on systems, man and cybernetics* (pp. 3639–3643).

Myrtek, M., Deutschmann-Janicke, E., Strohmaier, H., Zimmermann, W., Lawerenz, S., Brügner, G., & Müller, W. (1994). Physical, mental, emotional, and subjective workload components in train drivers. *Ergonomics*, *37*(7), 1195–1203.

Nakamura, Y., Yamamoto, Y., & Muraoka, I. (1993). Autonomic control of heart rate during physical exercise and fractal dimension of heart rate variability. *Journal of Applied Physiology*, *74*(2), 875–881.

Nakayama, K., Goto, S., Kuraoka, K., & Nakamura, K. (2005). Decrease in nasal temperature of rhesus monkeys (macaca mulatta) in negative emotional state. *Physiology & behavior*, *84*(5), 783–790.

National Highway Traffic Safety Administration, D. o. T. D. (2012). Visual-manual nhtsa driver distraction guidelines for in-vehicle electronic devices. *Washington, DC: National Highway Traffic Safety Administration (NHTSA), Department of Transportation (DOT)*.

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological review*, *86*(3), 214.

Newman, S. D., Keller, T. A., & Just, M. A. (2007). Volitional control of attention and brain activation in dual task performance. *Human brain mapping*, *28*(2), 109–117.

Niezgoda, M., Tarnowski, A., Kruszewski, M., & Kamiński, T. (2015). Towards testing auditory–vocal interfaces and detecting distraction while driving: A comparison of eye-movement measures in the assessment of cognitive workload. *Transportation research part F: traffic psychology and behaviour*, *32*, 23–34.

Nishimura, N., Murai, K., & Hayashi, Y. (2011). Basic study of a mental workload for student's simulator training using heart rate variability, salivary amylase activity and facial temperature. In *System of systems engineering (sose), 2011*

6th international conference on (pp. 67–71).

Novak, D., Mihelj, M., & Munih, M. (2012). A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing. *Interacting with Computers*, *24*(3), 154–172. doi: 10.1016/j.intcom .2012.04.003

Obrist, P. A. (1963). Cardiovascular differentiation of sensory stimuli. *Psychosomatic Medicine*, *25*(5), 450–459.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff (Ed.), *Handbook of perception and human performance.* New York: Wiley.

Or, C. K., & Duffy, V. G. (2007). Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. *Occupational Ergonomics*, *7*(2), 83–94.

Orlikoff, R. F., & Baken, R. (1989). The effect of the heartbeat on vocal fundamental frequency perturbation. *Journal of Speech, Language, and Hearing Research*, *32*(3), 576–582.

Östlund, J., Peters, B., Thorslund, B., Engström, J., Markkula, G., Keinath, A., ... Foehl, U. (2005). *Driving performance assessment - methods and metrics* (Tech. Rep.). AIDE Deliverable 2.2.5. Retrieved from `http://www.aide-eu.org/res_sp2.html`

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping*, *25*(1), 46–59.

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 141–144).

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International journal of human-computer studies*, *59*(1), 185–198.

Patten, C. J., Kircher, A., Östlund, J., & Nilsson, L. (2004). Using mobile telephones: cognitive workload and attention resource allocation. *Accident analysis & prevention*, *36*(3), 341–350.

Pavlidis, I., Dcosta, M., Taamneh, S., Manser, M., Ferris, T., Wunderlich, R., ... Tsiamyrtzis, P. (2016). Dissecting driver behaviors under cognitive, emotional, sensorimotor, and mixed stressors. *Scientific reports*, *6*.

Peck, E. M., Afergan, D., Yuksel, B. F., Lalooses, F., & Jacob, R. J. K. (2014). Using fnirs to measure mental workload in the real world. In S. H. Fairclough & K. Gilleade (Eds.), *Advances in physiological computing* (pp. 117–139). London: Springer London. Retrieved from `http://dx.doi.org/10.1007/978-1-4471-6392-3_6` doi: 10.1007/978-1-4471-6392-3_6

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pedrotti, M., Lei, S., Dzaack, J., & Rötting, M. (2011). A data-driven algorithm for offline pupil signal preprocessing and eyeblink detection in low-speed eye-tracking protocols. *Behavior Research Methods*, *43*(2), 372–383.

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual review of neuroscience*, *35*, 73.

Picard, R. W., & Picard, R. (1997). *Affective computing* (Vol. 252). MIT press Cambridge.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, *23*(10), 1175–1191.

Piechulla, W., Mayser, C., Gehrke, H., & König, W. (2003). Reducing drivers mental workload by means of an adaptive man–machine interface. *Transportation Research Part F: Traffic Psychology and Behaviour*, *6*(4), 233–248.

Pinel, J. P. J. (2000). *Biopsychology* - (4th ed.). Boston: Allyn and Bacon.

Poh, M.-Z., McDuff, D. J., & Picard, R. W. (2010). Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, *18*(10), 10762–10774.

Pomplun, M., Reingold, E. M., & Shen, J. (2001). Investigating the visual span in comparative search: The effects of task difficulty and divided attention. *Cognition*, *81*(2), B57–B67.

Posner, M. I., & Fan, J. (2008). Attention as an organ system. *Topics in integrative neuroscience*, 31–61.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual review of neuroscience*, *13*(1), 25–42.

Pribram, K. H., & McGuinness, D. (1975). Arousal, activation, and effort in the control of attention. *Psychological review*, *82*(2), 116.

Putze, F., Jarvis, J.-P., & Schultz, T. (2010). Multimodal recognition of cognitive workload for multitasking in the car. In *Pattern recognition (icpr), 2010 20th international conference on* (pp. 3748–3751).

R Core Team. (2015a). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

R Core Team. (2015b). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Radulescu, E., Nagai, Y., & Critchley, H. (2015). Mental effort: Brain and autonomic correlates in health and disease. In *Handbook of biobehavioral approaches to self-regulation* (pp. 237–253). Springer.

Rakauskas, M. E., Gugerty, L. J., & Ward, N. J. (2004). Effects of naturalistic cell phone conversations on driving performance. *Journal of Safety Research*, *35*(4), 453–464. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0022437504000830` doi: 10.1016/j.jsr.2004.06.003

Recarte, M. A., & Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of experimental psychology: Applied*, *6*(1), 31.

Recarte, M. A., & Nunes, L. M. (2002). Mental load and loss of control over speed in real driving.towards a theory of attentional speed control. *Transportation*

Research Part F: Traffic Psychology and Behaviour, *5*(2), 111–122. doi: 10 .1016/S1369-8478(02)00010-4

Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of experimental psychology: Applied*, *9*(2), 119.

Regan, M. A., Hallett, C., & Gordon, C. P. (2011). Driver distraction and driver inattention: Definition, relationship and taxonomy. *Accident Analysis & Prevention*, *43*(5), 1771–1781.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in psychology*, *52*, 185–218.

Reimer, B. (2009). Impact of cognitive task complexity on drivers' visual tunneling. *Transportation Research Record: Journal of the Transportation Research Board*(2138), 13–19.

Reimer, B., & Mehler, B. (2011). The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation. *Ergonomics*, *54*(10), 932–942.

Reimer, B., Mehler, B., Wang, Y., & Coughlin, J. F. (2012). A field study on the impact of variations in short-term memory demands on drivers visual attention and driving performance across three age groups. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(3), 454–468.

Reyes, M. L., & Lee, J. D. (2008). Effects of cognitive load presence and duration on driver eye movements and event detection performance. *Transportation research part F: traffic psychology and behaviour*, *11*(6), 391–402.

Reyes, M. L., Lee, J. D., Liang, Y., Hoffman, J. D., & Huang, R. W. (2009). Capturing driver response to in-vehicle human-machine interface technologies using facial thermography. In *Proceedings of the international driving symposium on human factors in driver assessment, training and vehicle design* (Vol. 5, pp. 536–542).

Rimm-Kaufman, S. E., & Kagan, J. (1996). The psychological significance of changes in skin temperature. *Motivation and Emotion*, *20*(1), 63–78. doi: 10.1007/ BF02251007

Ring, E., & Ammer, K. (2012). Infrared thermal imaging in medicine. *Physiological measurement*, *33*(3), R33.

Robinson, B. F., Epstein, S. E., Beiser, G. D., & Braunwald, E. (1966). Control of heart rate by the autonomic nervous system studies in man on the interrelation between baroreceptor mechanisms and exercise. *Circulation Research*, *19*(2), 400–411.

Ross, V., Jongen, E. M., Wang, W., Brijs, T., Brijs, K., Ruiter, R. A., & Wets, G. (2014). Investigating the influence of working memory capacity when driving behavior is combined with cognitive load: An {LCT} study of young novice drivers. *Accident Analysis & Prevention*, *62*, 377 - 387. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0001457513002601` doi: http://doi.org/10.1016/j.aap.2013.06.032

Rothkrantz, L. J. M., Wiggers, P., Wees, J.-W. A., & Vark, R. J. (2004). Voice stress analysis. In D. Hutchison et al. (Eds.), *Text, speech and dialogue* (Vol.

3206, pp. 449–456). Berlin and Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-540-30120-2_57

Rötting, M. (2001). *Parametersystematik der augen-und blickbewegungen für arbeitswissenschaftliche untersuchungen.* Shaker.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, *53*(1), 61–86.

Ruff, S., & Rötting, M. (2013). Impact of increasing workload on facial temperature in a simulated driving task. In *Proceedings of the 10th berlin workshop human-machine systems* (pp. 58–65).

Sahayadhas, A., Sundaraj, K., Murugappan, M., & Palaniappan, R. (2015). A physiological measures-based method for detecting inattention in drivers using machine learning approach. *Biocybernetics and Biomedical Engineering*, *35*(3), 198–205.

Sanders, A. (1983). Towards a model of stress and human performance. *Acta psychologica*, *53*(1), 61–97.

Schaap, T., Van der Horst, A., Van Arem, B., & Brookhuis, K. A. (2013). *The relationship between driver distraction and mental workload* (Vol. 1). Ashgate Farnham.

Scherer, K. R., Grandjean, D., Johnstone, T., Klasmeyer, G., & Bänziger, T. (2002). Acoustic correlates of task load and stress. In *Seventh international conference on spoken language processing.*

Schmidt, R. F., & Schaible, H.-G. (2006). *Neuro-und sinnesphysiologie.* Berlin: Springer.

Schneider, W., & Detweiler, M. (1988). The role of practice in dual-task performance: toward workload modeling a connectionist/control architecture. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *30*(5), 539–566.

Scholl, A. (2016). *Einfluss der Arbeitsgedächtniskapazität auf die Fahrerbeanspruchung bei Dual-Task Aufgaben* (Unpublished master's thesis). Technischen Universität Berlin, Germany.

Schwalm, M. (2009). *Pupillometrie als methode zur erfassung mentaler beanspruchungen im automotiven kontext* (Doctoral dissertation, Universität des Saarlandes, Postfach 151141, 66041 Saarbrücken). Retrieved from `http://scidok.sulb.uni-saarland.de/volltexte/2009/2082`

Seifert, K. (2002). *Evaluation multimodaler computer-systeme in frühen entwicklungsphasen* (Unpublished doctoral dissertation). Technische Universität Berlin.

Shastri, D., Merla, A., Tsiamyrtzis, P., & Pavlidis, I. (2009). Imaging facial signs of neurophysiological responses. *IEEE Transactions on Biomedical Engineering*, *56*(2), 477–484.

Shastri, D., Papadakis, M., Tsiamyrtzis, P., Bass, B., & Pavlidis, I. (2012). Perinasal imaging of physiological stress and its affective potential. *IEEE Transactions on Affective Computing*, *3*(3), 366–378.

Shepherd, A. P., & Öberg, P. Å. (2013). *Laser-doppler blood flowmetry* (Vol. 107). Springer Science & Business Media.

Shue, Y.-L. (2010a). *Voicesauce - a program for voice analysis.* `http://www.seas.ucla.edu/spapl/voicesauce/`. (Accessed: 2016-07-26)

Shue, Y.-L. (2010b). *The voice source in speech production: Data, analysis and models* (Unpublished doctoral dissertation). University of California Los Angeles.

Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(6), 679–692.

Smiley, A., & Brookhuis, K. A. (1987). Alcohol, drugs and traffic safety. road users and traffic safety. *Publication of: VAN GORCUM & COMP BV*.

Solovey, E. T., Zec, M., Garcia Perez, E. A., Reimer, B., & Mehler, B. (2014). Classifying driver workload using physiological and driving performance data: two field studies. In *Proceedings of the 32nd annual acm conference on human factors in computing systems* (pp. 4057–4066).

Statista. (2015). *U.s. smartphone use while driving 2015 — survey.* `https://www.statista.com/statistics/537883/us-smartphone-activities-while-driving/`. ((Accessed on 04/14/2017))

Statista. (2016a). *Safety & driving assistance.* `https://www.statista.com/outlook/322/100/safety-driving-assistance/worldwide`. (Accessed: 2016-08-18)

Statista. (2016b). *Smartphones - statista dossier.* `https://www.statista.com/study/10490/smartphones-statista-dossier/`. (Accessed: 2016-08-18)

Steinhauer, S. R., Siegle, G. J., Condray, R., & Pless, M. (2004). Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *International journal of psychophysiology*, *52*(1), 77–86.

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of experimental psychology: Applied*, *9*(1), 23.

Strayer, D. L., Turrill, J., Cooper, J. M., Coleman, J. R., Medeiros-Ward, N., & Biondi, F. (2015). Assessing cognitive distraction in the automobile. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(8), 1300–1324.

Summala, H., Nieminen, T., & Punto, M. (1996). Maintaining lane position with peripheral vision during in-vehicle tasks. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *38*(3), 442–451.

Szmalec, A., Verbruggen, F., Vandierendonck, A., & Kemps, E. (2011). Control of interference during working memory updating. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 137.

Sztajzel, J. (2004). Heart rate variability: a noninvasive electrocardiographic method to measure the autonomic nervous system. *Swiss medical weekly*, *134*, 514–522.

Tarvainen, M. P., Niskanen, J.-P., Lipponen, J. A., Ranta-Aho, P. O., & Karjalainen, P. A. (2014). Kubios hrv–heart rate variability analysis software. *Computer methods and programs in biomedicine*, *113*(1), 210–220.

Tefft, B. C. (2014, 11). *Prevalence of motor vehicle crashes involving drowsy drivers, united states, 2009 2013* (Tech. Rep.). 607 14th Street, NW, Suite

201; Washington DC 20005: AAA Foundation for Traffic Safety. Retrieved from www.aaafoundation.org

Tijerina, L., Parmer, E., & Goodman, M. J. (1998). Driver workload assessment of route guidance system destination entry while driving: A test track study. In *Proceedings of the 5th its world congress* (pp. 12–16).

Tjolleng, A., Jung, K., Hong, W., Lee, W., Lee, B., You, H., ... Park, S. (2017). Classification of a driver's cognitive workload levels using artificial neural network on ecg signals. *Applied Ergonomics*, *59*, 326–332.

Tole, S. A. V. M. E. A. . Y. L., J.R. (1983, 7). *Visual scanning behavior and pilot workload* (Tech. Rep.). Hampton (VA), USA: NASA Langley Research Center.: Massachusetts Institute of Technology, Cambridge, Massachusetts. (NASA Contractor Report 3717)

Tsai, Y.-F., Viirre, E., Strychacz, C., Chase, B., & Jung, T.-P. (2007). Task performance and eye activity: predicting behavior relating to cognitive workload. *Aviation, space, and environmental medicine*, *78*(Supplement 1), B176–B185.

Vaish, A., & Kumari, P. (2014). A comparative study on machine learning algorithms in emotion state recognition using ecg. In *Proceedings of the second international conference on soft computing for problem solving (socpros 2012), december 28-30, 2012* (pp. 1467–1476).

Veltman, J. A., & Gaillard, A. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology*, *42*(3), 323–342. doi: 10.1016/0301-0511(95)05165-1

Veltman, J. A., & Vos, W. K. (2005). Facial temperature as a measure of operator state. In *11th international congres on human-computer interactions 2005.*

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech communication*, *48*(9), 1162–1181.

Victor, T. W., Harbluk, J. L., & Engström, J. A. (2005). Sensitivity of eye-movement measures to in-vehicle task difficulty. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*(2), 167–190.

Vidulich, M. A., & Tsang, P. S. (2012). Mental workload and situation awareness. In *Handbook of human factors and ergonomics* (pp. 243–273). John Wiley & Sons, Inc. Retrieved from http://dx.doi.org/10.1002/9781118131350.ch8 doi: 10.1002/9781118131350.ch8

Voßkühler. (2015). *open gaze and mouse analyzer.* http://www.ogama.net/. (Accessed: 2016-07-26)

Wagner, J., Kim, J., & André, E. (2005). From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *Multimedia and expo, 2005. icme 2005. ieee international conference on* (pp. 940–943).

Wang, M., Jeong, N., Kim, K., Choi, S., Yang, S., You, S., ... Suh, M. (2016). Drowsy behavior detection based on driving information. *International journal of automotive technology*, *17*(1), 165–173.

Wartzek, T., Eilebrecht, B., Lem, J., Lindner, H.-J., Leonhardt, S., & Walter, M. (2011). Ecg on the road: robust and unobtrusive estimation of heart rate. *IEEE Transactions on Biomedical Engineering*, *58*(11), 3112–3120.

Wesley, A., Shastri, D., & Pavlidis, I. (2010). A novel method to monitor driver's distractions. In *Chi'10 extended abstracts on human factors in computing systems* (pp. 4273–4278).

Wickens, C. D. (1980). The structure of attentional resources. *Attention and performance VIII*, *8*.

Wickens, C. D. (1991). Processing resources and attention. *Multiple-task performance*, *1991*, 3–34.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177. doi: 10.1080/14639220210123806

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 449–455. doi: 10.1518/001872008X288394

Wickens, C. D. (2014). Effort in human factors performance and decision making. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 0018720814558419.

Wickens, C. D., Goh, J., Helleberg, J., Horrey, W. J., & Talleur, D. A. (2003). Attentional models of multitask pilot performance using advanced display technology. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *45*(3), 360–380.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance* (3rd ed.). Upper Saddle River and NJ: Prentice Hall.

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *35*(2), 263–281.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, *52*(4B), 1238–1250.

Wilson, G. F. (1992). Applied use of cardiac and respiration measures: Practical considerations and precautions. *Biological Psychology*, *34*(2-3), 163–178.

Winn, B., Whitaker, D., Elliott, D. B., & Phillips, N. J. (1994). Factors affecting light-adapted pupil size in normal human subjects. *Investigative Ophthalmology & Visual Science*, *35*(3), 1132–1137.

WIVW. (2016). *Fahrsimulation und silab.* Retrieved from `https://wivw.de/de/silab` (Accessed: 2016-07-26)

Wyatt, H. J., & Musselman, J. F. (1981). Pupillary light reflex in humans: evidence for an unbalanced pathway from nasal retina, and for signal cancellation in brainstem. *Vision research*, *21*(4), 513–525.

Xie, B., & Salvendy, G. (2000a). Prediction of mental workload in single and multiple tasks environments. *International journal of cognitive ergonomics*, *4*(3), 213–242.

Xie, B., & Salvendy, G. (2000b). Review and reappraisal of modelling and predicting mental workload in single-and multi-task environments. *Work & stress*, *14*(1), 74–99.

Yager, C. (2013). *An evaluation of the effectiveness of voice-to-text programs at reducing incidences of distracted driving* (Tech. Rep.). College Station, Texas:

Southwest Region University Transportation Center (SWUTC).

Yamakoshi, T., Yamakoshi, K., Tanaka, S., Nogawa, M., Park, S.-B., Shibata, M., ... Hirose, Y. (2008). Feasibility study on driver's stress detection from differential skin temperature measurement. In *2008 30th annual international conference of the ieee engineering in medicine and biology society* (pp. 1076–1079).

Yarbus, A. L. (1967). *Eye movements during perception of complex objects*. Springer.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*(5), 459–482. doi: 10.1002/cne.920180503

Young, M., & Stanton, N. (2004, aug). Mental workload. In *Handbook of human factors and ergonomics methods* (pp. 39–1–39–9). CRC Press. Retrieved from `https://doi.org/10.1201/9780203489925.ch39` doi: 10.1201/9780203489925.ch39

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17.

Young, M. S., & Stanton, N. A. (2002a). Attention and automation: new perspectives on mental underload and performance. *Theoretical Issues in Ergonomics Science*, *3*(2), 178–194.

Young, M. S., & Stanton, N. A. (2002b). Malleable attentional resources theory: a new explanation for the effects of mental underload on performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(3), 365–375.

Young, M. S., & Stanton, N. A. (2004). Mental workload. In *Handbook of human factors and ergonomics methods* (p. 39-1-39-9). CRC Press. Retrieved from `http://dx.doi.org/10.1201/9780203489925.ch39` doi: 10.1201/9780203489925.ch39

Young, R. (2012). Cognitive distraction while driving: A critical review of definitions and prevalence in crashes. *SAE International journal of passenger cars-electronic and electrical systems*, *5*(2012-01-0967), 326–342.

Zhai, J., & Barreto, A. (2006). Stress detection in computer users based on digital signal processing of noninvasive physiological variables. In *Engineering in medicine and biology society, 2006. embs'06. 28th annual international conference of the ieee* (pp. 1355–1358).

Zhang, Y., Owechko, Y., & Zhang, J. (2004). Driver cognitive workload estimation: A data-driven perspective. In *Intelligent transportation systems, 2004. proceedings. the 7th international ieee conference on* (pp. 642–647).

Zhao, C., Zheng, C., Zhao, M., Tu, Y., & Liu, J. (2011). Multivariate autoregressive models and kernel learning algorithms for classifying driving mental fatigue based on electroencephalographic. *Expert Systems with Applications*, *38*(3), 1859–1865.

Zijlstra, F. R. H. (1993). *Efficiency in work behaviour: A design approach for modern tools*. TU Delft, Delft University of Technology.