Acquisition of cognitive load under time pressure: combining multimodal measurements and a theory-based approach as a pathway to online adaptation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Natalia Sevcenko, M.Sc.

aus Riga / Lettland

Tübingen

2021

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	09.03.2022
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Peter Gerjets
2. Berichterstatter:	Prof. Dr. Stephan Schwan

Table of Contents

Tables & Figures

Tables

Figures

Equations

Abstract

Human-machine interactions (HMIs) are constantly evolving in complexity, while human cognitive resources remain limited. Ample evidence demonstrates an association between cognitive load and human performance, suggesting that the number of human errors might be reduced by using intelligent interfaces capable of keeping operators' cognitive load at medium degrees. The development of such systems primarily requires an appropriate measurement method for cognitive load, which should be (1) easily generalizability to similar situations, (2) capable of tracking cognitive load over time without impairing performance on the primary task, and (3) as discrete and unobtrusive as possible to be suitable for real-world and commercial developments. This dissertation attempts to address these challenges by presenting a novel theoretically-grounded approach for measuring cognitive load in situations involving managing resources under time pressure. In a series of three studies, the *initial temporal action density decay* metric (TADD) was derived relying on the time-based resource-sharing (TBRS) model (Barrouillet, Bernardin, & Camos, 2004) and validated using behavioral data, cortical hemodynamics, and eye-tracking features. In summary, it is consistently demonstrated that narrow theoretically-determined time periods can be used for predicting cognitive load and user performance and such a targeted approach may have an advantage over data-driven bottom-up methods. By using different measurement methods, a consistent picture of the method is elaborated and conclusions could be drawn about the cognitive processes occurring during the investigated time periods. Taken together, the presented dissertation provides evidence for the applicability of the TBRS model as a theoretical basis for measuring cognitive load in realistic time pressure situations and suggests that time pressure and attentions play a role in the development of cognitive load. Furthermore, it also provides an indication of which features are more likely to be useful for assessing cognitive load. Although further research is needed to refine the method, it seems promising for the development of HMI systems capable of adapting to the cognitive load of their operators.

Zusammenfassung

Die Komplexität der Mensch-Maschine-Interaktion (HMI) nimmt ständig zu, während die kognitiven Ressourcen des Menschen begrenzt bleiben. Es gibt zahlreiche Belege für einen Zusammenhang zwischen kognitiver Beanspruchung und menschlicher Leistung, was darauf hindeutet, dass die Anzahl menschlicher Fehler dadurch verringert werden könnte, dass man intelligente Schnittstellen, die die kognitive Beanspruchung der Bediener auf einem mittleren Niveau halten können, einsetzt. Die Entwicklung solcher Systeme erfordert in erster Linie eine geeignete Messmethode für die kognitive Beanspruchung, die (1) leicht auf ähnliche Situationen verallgemeinert werden kann, (2) in der Lage ist, die kognitive Beanspruchung über die Zeit zu verfolgen, ohne die Leistung bei der primären Aufgabe zu beeinträchtigen, und (3) so diskret und unauffällig wie möglich ist, um für kommerzielle Entwicklungen geeignet zu sein. Diese Dissertation versucht, diese Herausforderungen zu adressieren, indem sie einen neuartigen, theoretisch fundierten Ansatz zur Messung der kognitiven Beanspruchung in Situationen vorstellt, in denen es um das Management von Ressourcen unter Zeitdruck geht. In einer Reihe von drei Studien wurde die initial temporal action density decay Metrik (TADD) auf der Grundlage des time-based resource-sharing Modells (TBRS) von Barrouillet et al. (2004) entwickelt und anhand von Verhaltensdaten, kortikaler Hämodynamik und Eye-Tracking Korrelaten validiert. Zusammenfassend, konnte es gezeigt werden, dass spezifische Zeiträume für die Vorhersage der kognitiven Beanspruchung und der Benutzerleistung verwendet werden können und dass ein theoretisch fundierter Ansatz einen Vorteil gegenüber datengesteuerten Bottom-up Methoden haben kann. Durch den Einsatz verschiedener Messmethoden wurde ein konsistentes Bild der Methode erarbeitet und es konnten Rückschlüsse auf die kognitiven Prozesse gezogen werden, die in den untersuchten Zeiträumen ablaufen. Insgesamt liefert die vorliegende Dissertation erste Belege für die Einsetzbarkeit des TBRS-Modells als theoretische Grundlage für Messungen der kognitiven Beanspruchung in realistischen Situationen unter Zeitdruck und untermauert die Rolle der Aufmerksamkeit als entscheidenden Aspekt bei der Einschränkung des Arbeitsgedächtnisses (WM). Obwohl weitere Forschung zur Verfeinerung der Methode erforderlich ist, scheint sie vielversprechend für die Entwicklung von computerbasierten Systemen, die sich an die kognitive Beanspruchung ihrer Benutzer anpassen können.

List of publications

Accepted / published manuscripts

Sevcenko, N., Ninaus, M., Wortha, F., Moeller, K., & Gerjets, P. (2021). Measuring cognitive load using in-game metrics of a serious simulation game. *Frontiers in Psychology*, 12, 906. https://doi.org/10.3389/fpsyg.2021.572437

Sevcenko, N., Shopp, B., Dresler, T., Ehlis, A-C., Ninaus, M. Korbinian, M., & Gerjets, P. (2021). Neural Correlates Of Cognitive Load While Playing Emergency Simulation Game: A Functional Near-Infrared Spectroscopy (fNIRS) Study. 10.1109/TG.2022.3142954

Submitted manuscripts

Sevcenko, N., Appel, T., Ninaus, M., Moeller, K., & Gerjets, P. (2021). Theory-Based Approach For Assessing Cognitive Load During Time-Critical Resource-Managing Human-Computer Interactions: An Eye-Tracking Study

Contributions

This dissertation contains material from three submitted/published author manuscripts (Chapters 2-4). In addition, passages from the manuscripts are included in the Introduction (Chapter 1) and General Discussion (Chapter 5). All co-authors and their proportional contributions to these manuscripts are presented in the following tables.

Manuscript 1

Author	Author position	Scientific ideas	Data generation	Analysis & interpretation	Writing
Natalia Sevcenko	1	65%	100%	75%	60%
Manuel Ninaus	2	10%	0 %	0 %	15%
Franz Wortha	3	5%	0 %	15%	5%
Korbinian Moeller	4	10%	0 %	0 %	10%
Peter Gerjets	5	10%	0 %	10 %	10%
Publication title		Measuring Cognitive Load Using In-Game Metrics Of A Serious Simulation Game			
Status in publication process		published			

Manuscript 2

Author	Author position	Scientific ideas	Data generation	Analysis & interpretation	Writing
Natalia Sevcenko	1	65%	100%	60%	70%
Betti Schopp	2	0 %	0 %	10 %	2%
Thomas Dresler	3	5%	0 %	5 %	4%
Ann-Christine Ehlis	4	10%	0 %	10%	4%
Manuel Ninaus	5	5%	0 %	5%	5%
Korbinian Moeller	6	5%	0 %	0%	10%
Peter Gerjets	7	10%	0 %	10%	5%
Publication title		Neural Correlates Of Cognitive Load While Playing Emergency Simulation Game: A Functional Near-Infrared Spectroscopy (fNIRS) Study			
Status in publication process		submitted			

Manuscript 3

Author	Author position	Scientific ideas	Data generation	Analysis & interpretation	Writing
Natalia Sevcenko	1	80%	100%	70%	75%
Tobias Appel	2	0%	0%	20%	5%
Manuel Ninaus	3	5%	0%	0%	5%
Korbinian Moeller	4	5%	0%	0%	10%
Peter Gerjets	5	10%	0%	10%	5%
Publication title		Theory-Based Approach For Assessing Cognitive Load During Time-Critical Resource-Managing Human-Computer Interactions: An Eye-Tracking Study			
Status in publication process		submitted			

1 Introduction and Theoretical Framework

Human-machine interactions (HMIs) are constantly evolving in complexity. For example, whereas in 1983 the driver of an average passenger car operated about seven functions, in 2014 the number of functions already exceeded 60 (Ruck & Stottan, 2014). At the same time, computer-based systems such as autopilots, speed and lane assistants, etc. are taking on increasingly complex tasks, while monitoring these systems is becoming more sophisticated (Stapel, Mullakkal-Babu, & Happee, 2019).

However, human cognitive resources are limited (Cowan, 2010; G. A. Miller, 1956). According to the definition of Paas and Van Merriënboer (1994), cognitive load is a multidimensional construct and represents demands that a particular task imposes on the cognitive system. Despite the long history of research (e.g., Barrouillet et al., 2004; Eggemeier, Shingledecker, & Crabtree, 1985; Linton, Jahns, & Chatelier, 1978; Meshkati, 1988; Sheridan & Simpson, 1979; Sweller, Van Merrienboer, & Paas, 1998; Welford, 1978), studying cognitive load still represents a vibrant area of interest with critical relevance to real-life environments. Evidence indicates that different tasks can require different cognitive resources to varying degrees, depending on task difficulty, whereas at the same time different individuals may experience different levels of cognitive load when conducting a task even when achieving the same performance on it (cf. Babiloni, 2019).

Ample evidence has demonstrated an association between cognitive load and human performance in a variety of realistic settings such as transportation (Fan & Smith, 2017; G. Hancock, Hancock, & Janelle, 2012; P. Hancock, 1989), e-learning (Oviatt, 2006; Walter, Rosenstiel, Bogdan, Gerjets, & Spüler, 2017), office work (Aasted et al., 2015; Smith-Jackson & Klein, 2009) and medicine (Yurko, Scerbo, Prabhu, Acker, & Stefanidis, 2010). This relation seems to be shaped like an “inverted-U” (Yerkes & Dodson, 1908) with the highest performance under medium cognitive load. Importantly, this observation corresponds to Csikszentmihalyis’ concept of “flow” (Csikszentmihalyi, 1975; Kiili, Lindstedt, & Ninaus, 2018), which characterizes a state of total concentration on the task at hand and assumes that performance usually declines when cognitive demands are boring or overstraining, as well as to the “zone of proximal development” proposed by Vygotsky

(1980) in his theory of cognitive development, assuming the strongest training effects at moderately challenging conditions.

Thus, to optimize operators' performance and prevent errors during HMIs, cognitive load should be kept at a medium level, while avoiding states of over- and underload, which are likely to lead to reduced performance and increased errors (Orru & Longo, 2019). Specifically in critical emergencies, human errors might lead to dramatic consequences and should therefore be addressed. For instance, in Germany in 2016, 57% of the accidents involving goods vehicles with personal injury were due to human errors (Statistisches Bundesamt, 2017). One possible way to reduce the number of such errors might be to develop intelligent human-computer interfaces capable of capturing operators' cognitive load and adapting their appearance to it (e.g., by reducing displayed information when the operator is cognitively overloaded).

Indeed, empirical evidence indicates that cognitive load may be influenced by interface design. For instance, Oviatt (2006) found that students demonstrated significantly better performance in solving mathematical problems when using a digital pen and paper interface compared to graphical tablet interfaces. Another example was provided by the study of Charabati, Bracco, Mathieu, and Hemmerling (2009), who compared differently-designed interfaces for monitoring anesthesia parameters during a surgery and reported that participants perceived significantly lower cognitive load when using a mixed numerical-graphical interface compared to the numerical and advanced-graphical interfaces. At the same time, online adaptation of task difficulties to the cognitive load has been shown to significantly improve performance in a variety of realistic settings (see Section 1.1). Taken together, these examples indicate that the online adaptation of HMI to cognitive load is practicable and can indeed improve operators' performance. This raises the question of which measurement methods are best suited for assessing cognitive load in such systems.

This dissertation addresses precisely this point by presenting a theory-based method for measuring cognitive load in time-critical situations, with the aim of contributing to the development of realistic adaptive HMIs. Accordingly, the remainder of this dissertation is structured as follows. I first provide an overview of the modern research landscape on adaptation to cognitive load (Section 1.1) and describe four classes of measurement methods, particularly concentrating on the methods used in this dissertation (Section 1.2).

Subsequently, I present the theoretical framework for cognitive load (Section 1.3) and describe in detail the TBRS model (Barrouillet et al., 2004), which was used as a foundation for the temporal action density decay (TADD) metric (Section 1.4) developed and investigated in the course of the project. Closing the first chapter, objectives and research questions are specified (Section 1.5). In the next three chapters, I provide author manuscripts, describing three studies conducted as part of this dissertation (Chapters 2-4). Finally, the results of the studies are summarized (Section 5.1), scientific contribution of this dissertation is discussed (Section 5.2), strengths and limitations are outlined (Section 5.3), and an outlook for future research is provided (Section 5.4).

1.1 Adaptation to cognitive load

Adaptive interaction systems, which can change their appearance and/or behavior depending on the operators' cognitive load to improve performance and avoid errors, are increasingly becoming the focus of modern research and development. Evidence indicates that such systems might be used to optimize performance and motivation in a variety of realistic contexts, such as education, gaming, emergency management, automotive and aviation.

In the educational context, efforts are underway to develop adaptive systems that can continuously monitor a learners' cognitive load in realistic settings (Grissmann et al., 2017; Spüler et al., 2016). In educational video games (for review see: Nebel & Ninaus, 2019; Ninaus & Nebel, 2021), adaptation to cognitive load has been found to be associated with increasing learning outcomes. As one example, Yuksel et al. (2016) used near-infrared spectroscopy (NIRS) to identify states of cognitive underload in pianists during a musical learning task and responded by increasing the difficulty of the respective lessons. In this way, a significantly better learning performance compared to a control group could be achieved. Walter et al. (2017) developed a closed-loop EEG-based environment to detect learners' cognitive load and adopt the complexity of the learning content accordingly. It is worth noting that the optimal load was defined based on small pilot sample data and was not individually calibrated. Nevertheless, the use of this system resulted in performance improvements comparable to those obtained when using conventional error-based

adaptation. Thus, adaptive learning environments represent a very promising research field, whereas, as Kalyuga (2007) highlighted, the design of such systems must simultaneously take into account the notion that techniques used in adaptive learning environments can also create additional load on learners.

Medicine, and especially emergency medicine in the context of surgical procedures, recognizes the importance of online assessment of cognitive load to avoid intraoperative errors in clinical settings and achieve better training effects in medical education (for review see: Dias, Ngo-Howard, Boskovski, Zenati, & Yule, 2018). As one example of an attempt to develop adaptive simulations for training purposes to improve the learning outcomes of emergency physicians, Sarkar et al. (2019) designed a multitasking deep neural network to classify high and low cognitive load in experts and novices while performing a trauma simulation based on electrocardiogram (ECG) signals from both novices and experts. At the same time, efforts are undertaken to make existing measurement methods fit for realistic use, e.g., by developing novel artifact filtering algorithms for EEG data (e. g. Rosanne et al., 2021).

Again in the context of emergency management, it is increasingly recognized that many critical errors are caused by cognitive overload and are avoidable through the use of intelligent methods (Croskerry & Sinclair, 2001; Vella, Hall, van Merrienboer, Hopman, & Szulewski). At the same time, suitable ways to determine the cognitive load of managers in emergencies are being sought, which might form the basis for intelligent adaptive systems in the future. Whereas such practical solutions as the analysis of linguistic features (Khawaja, Chen, Owen, & Hickey, 2009), or eye-tracking features (Appel et al., 2019) are proposed, at the same time attempts are being made to develop adaptive computerized systems to reduce the cognitive load of operation managers. For example, Mirbabaie and Fromm (2019) aimed to support the emergency management decision-making process during realistic emergency situations by developing an augmented reality support system.

In the automotive field, the vision of an adaptive HMI for optimal driver support is not new and was already expressed by Michon (1993). Since then, several attempts have been made to develop an in-car system adaptive to cognitive load. One example is phones' adaptive system developed by BMW (Piechulla, Mayser, Gehrke, & König, 2003), which redirects incoming calls to the mailbox in cases of detected cognitive overload, based on

traffic situation and driving dynamics. Toyotas' adaptive messaging system adopts the same approach, postponing voice messages when the driver is overloaded. The estimation of cognitive load was developed based on accelerator pedal operation. Comparatively, researchers from Daimler (Riener & Noldi, 2015) investigated the relationship between the drivers' cognitive load and his or her motion dynamics in the car seat. Another way to reduce drivers' cognitive load comprises optimizing intervals between messages coming from in-vehicle systems (Wu, Tsimhoni, & Liu, 2008). If messages appear too often or even simultaneously, it can increase drivers' cognitive load and impair performance, whereas adaptive increasing the intervals between messages, based on driving conditions and task properties, was found to reduce drivers' cognitive load and enhance performance (Lin & Wu, 2010). As another example, Kohlmorgen et al. (2007) could significantly improve driving performance during a real driving task on a highway that included executing secondary tasks unrelated to driving. Thereby, driver overload states were detected using electroencephalography (EEG) and the difficulty of the secondary task was reduced accordingly. These examples represent interesting but punctual application areas for a system adaptive to a drivers' cognitive load and demonstrate the variability of the approaches. One can extend the vision and think about systems that can e.g., completely take over the driving task in the future if a cognitive overload of the driver is detected. However, there is still a long way to go and there remains much research to be conducted to develop comprehensive generic intelligent driver support, as envisioned by Michon (1993), and accurate detection of cognitive load in the vehicle context remains a vibrant area of research.

Another area in which research of cognitive load has a long tradition is aviation. The first studies aiming to detect cognitive states in pilots date back to the 1980s and 1990s. For example, Wilson, Purvis, Skelly, Fullenkamp, and Davis (1987) investigated associations between heart rate, blink count and duration as well as EEG data and cognitive load in pilots during simulated and real flights. Ten years later, J. Veltman and Gaillard (1996) investigated respiratory activity, blood pressure and blink count during simulated flight. More such studies followed in subsequent years, with the idea of developing intelligent assistive systems adapted to cognitive load coming to the foreground. As one example, adaptive cognitive agent (Roth, Schulte, Schmitt, & Brand, 2019; Strenzke et al., 2011) was developed to support helicopter crew members individually. Based on task load and

behavioral data, a cognitive load estimator detects states of cognitive overload and decides whether the particular crew member needs help and which of two available levels of help to offer. Another cognitive agent was developed to help with air traffic management. In a simulator study using an EEG-based brain-computer interface, Aricò et al. (2016) showed that the adaptive condition (BCI system activates help in detecting cognitive overload) led to significant improvement in operator performance. As another example, Wilson and Russell (2007) found that adaptive aiding may significantly enhance pilots' performance. Researchers used neural network to detect states of high cognitive load during a simulated uninhabited aerial vehicle task based on (neuro-) psychophysiological data (EEG, electrocardiogram, electrooculography, and eye-tracking) and adapted the task difficulty through reducing the velocity of the vehicle and displaying helpful status messages.

Summarizing this brief literature overview on the topic of adaptation to cognitive load, one can draw two main conclusions. First, adaptation to cognitive load remains a very nascent research field (for review, see: Fontaine et al., 2019; Ninaus & Nebel, 2021), with no common theoretical framework yet established and most attempts in this regard often based on data-driven probabilistic performance evaluations (Magerko, Stensrud, & Holt, 2006; Spronck, Ponsen, Sprinkhuizen-Kuyper, & Postma, 2006; Zook & Riedl, 2012). Second, the presented results appear very promising for a variety of realistic contexts and the online adaptation of HMI to cognitive load is practicable and can indeed improve operators' performance. This raises the question of which measurement methods are best suited for assessing cognitive load in such systems.

1.2 Measuring cognitive load in adaptive systems

Evidence suggests that cognitive load can rapidly change during the processing of a task, which means that a measurement procedure suitable for online adaptation should be able to respond sensitively to these variations without causing external disturbances to performance on the primary task (Orru & Longo, 2019). In the literature, there is no overall consensus on the classification of measurement methods of cognitive load. Usually, three or four main categories are distinguished, with some methods, such as self-reported

questionnaires, being consistently categorized as subjective measures, while other methods (e.g.,

eye-tracking, analysis of speech patterns) can be found in different classes depending on the aspects on which the authors focus. Based on the literature research, in the following I introduce four categories of cognitive load measures (Brünken, Seufert, & Paas, 2010; F. Chen et al., 2016; Eggemeier, Wilson, Kramer, & Damos, 1991; Johannsen, 1979; Scerbo, 1996) and discuss their strengths and limitations with respect to the use in realistic adaptive systems.

1.2.1 Subjective measures

Subjective measures are grounded in the assumption that people are able to accurately interpret and adequately describe the cognitive load they have experienced while performing a task (Gopher & Braune, 1984). These self-reports are collected using predefined scales in questionnaires such as SWAT (Reid & Nygren, 1988) and NASA-TLX (Hart & Staveland, 1988). Whereas, for best recall and, thus, the most accurate ratings completion of the questionnaire should occur immediately after finishing a task.

Subjective measures do not require expensive equipment, they are easy to collect, and are usually well validated, hence widely accepted (O'Donnell & Eggemeier, 1986). Unfortunately, filling in questionnaires would interrupt task processing and therefore can only be done after the task has been completed, which leads to several limitations. First, fading memory may distort the rating. Second, memories of perceived cognitive load may be biased by experienced successes or failures, that is, experienced failure may lead participants to rate their cognitive load higher compared to experienced success (P. Hancock, 1989). Third, ratings obtained in this way represent only a rough summary of experience, which is particularly unfortunate when evaluating complex tasks that take a considerable amount of time. Finally and most importantly, subjective measures are unable to capture variations in cognitive load over time, which makes these measurements hardly practicable for online adaptation systems.

1.2.2 Performance-based measures

Performance-based measurement methods evaluate variations in human performance. Empirical evidence indicates that performance declines in response to cognitive under- or overload (Babiloni, 2019; J. Veltman & Jansen, 2005; Yerkes & Dodson, 1908). Accordingly, a drop of performance may help to identify these undesirable cognitive load states.

As a main objective of assessing cognitive load in adaptive systems is prediction and subsequent improvement of operators' performance, this class seems most obvious and direct to apply. Performance-based measurements are unobtrusive and relatively simple to implement. However, these measures yield no independent assessments of cognitive load, because it is not possible to determine whether the observed fluctuations in performance are actually due to changes in cognitive load or to other relevant factors such as arousal, motivation (Brünken et al., 2010), or emotions (Grissmann, Faller, Scharinger, Spüler, & Gerjets, 2017). In addition, it is usually not possible to obtain performance data online during actual task execution, so performance-based measurements are often computed and analyzed retrospectively, while "dual-task" measurements (Kerr, 1973) can be performed online but burden the operator with an additional task. Thus, also this class of measures appears impractical for widespread use in online adaptation systems.

1.2.3 Behavioral measures

Behavioral measurements attempt to estimate cognitive load based on variations in the users' interaction behavior during task processing, such as mouse and keyboard usage, or speech and voice patterns, etc. (e. g. Berthold & Jameson, 1999; Ikehara & Crosby, 2005; Lim, Ayesh, & Stacey, 2015; Magnusdottir, Borsky, Meier, Johannsdottir, & Gudnason, 2017; Ruiz, Liu, Yin, Farrow, & Chen, 2010). Because digital systems allow for recording of individual behavior during HMI, behavioral measures can be implemented in a cost-effective and unobtrusive manner that does not distract participants from the task at hand. Importantly, they potentially enable continuous online measurement of cognitive load during task execution, making them potentially promising for use in realistic online adaptive systems. However, it must be taken into account that behavioral patterns can also be

influenced by factors other than cognitive load and therefore do not provide independent assessment.

1.2.4 Physiological measures

Physiological measures rely on the evidence that changes in cognitive states are associated with physiological changes (Ahmad, Malik, Kamel, & Reza, 2016; Buchwald et al., 2019; Fowler, Nesbitt, & Canossa, 2019; Johannsen, 1979; Liang, Liang, Qu, & Yang, 2018; McDuff, Gontarek, & Picard, 2014). They allow for continuous recording of data and thus might be used for online adaptation to cognitive load. Depending on the type of signal and assessment technique, they differ considerably in their cost and obtrusiveness. For instance, while heart rate variability (HRV) and electrodermal activity (EDA) sensors tend to be discreet, functional magnetic resonance imaging (fMRI) appears impractical in real-world situations due to the immobility required of the participant (for an overview see Ninaus et al., 2013).

According to Brunken, Plass, and Leutner (2003), based on the relationship between cognitive load and observed variables, physiological measures can be further categorized into direct and indirect methods. From this perspective, such methods as eye-tracking and assessment of HRV are indirectly related to cognitive load, because they can be co-influenced by other factors such as stress or emotions, whereas brain imaging techniques such as fMRI, electroencephalography (EEG), positron emission tomography (PET), and functional near-infrared spectroscopy (fNIRS) can be considered direct, because they "directly" assess cortical activation in the predefined areas of interest. Although these methods also derive cognitive activity from blood flow or electrical activation, I retain this terminology because there is no more direct method of detecting cognitive load available today. In the following I describe in more detail two (neuro-) physiological methods that were used in this dissertation: fNIRS and eye-tracking.

1.2.4.1. Functional near-infrared spectroscopy (fNIRS)

FNIRS represents a non-invasive neuroimaging method for measuring cerebral hemodynamics, first presented in the fundamental work of Jobsis (1977). Following the terminology of Brunken et al. (2003), it can be considered a direct method for assessing cognitive load, because it can directly assess cortical activation in predefined areas of interest and thus separate different components of the response. Besides, it offers several advantages over other neuroimaging methods in terms of mobility, cost, and robustness, making it potentially promising for use in ecologically valid real-life settings.

FNIRS relies on the mechanism of neurovascular coupling and optical spectroscopy (for review see: Fallgatter, Ehlis, Wagener, Michel, & Herrmann, 2004; Herold, Wiegel, Scholkmann, & Müller, 2018; G. Strangman, Boas, & Sutton, 2002). During this procedure, near-infrared light is emitted through the participants' scalp. It penetrates up to approx. 2 cm (G. E. Strangman, Li, & Zhang, 2013) deep into the tissue, reaching outer cortical gray matter. Here it becomes partially absorbed by oxygenated (HbO2) and deoxygenated (HbR) hemoglobin molecules, which differ significantly in their absorption spectra. After that the residual reflection of the respective wavelengths is received by a detector (Herold et al., 2018; G. E. Strangman et al., 2013), and the relative concentration of HbO2 and HbR in the respective brain tissue is calculated (see Cope et al., 1988). In this way, inferences about local changes in blood flow and further conclusions about which brain regions are active and thus which processes take place in the brain at a certain point of time are drawn (Hoge et al., 1999; Kim, Rostrup, Larsson, Ogawa, & Paulson, 1999). When attempting to measure fluctuations in cognitive load, prefrontal cortex (PFC) is usually regarded (Herold et al., 2018). This choice is based on a solid body of empirical evidence indicating that PFC is involved in decision-making processes, which are often summarized as "executive functions" or "cognitive control" along with at least partial influence of working memory (E. K. Miller & Cohen, 2001; Thier, 2006) and that fNIRS assessments of PFC are sensitive to fluctuations in cognitive load (Ayaz, Izzetoglu, Bunce, Heiman-Patterson, & Onaral, 2007; Ehlis, Herrmann, Wagener, & Fallgatter, 2005; Fishburn, Norr, Medvedev, & Vaidya, 2014; Herff et al., 2014; Herrmann et al., 2007; Li, Gong, Gan, & Luo, 2005; Smith & Jonides, 1997; Xu et al., 2017).

When comparing the temporal and spatial resolution of neuroimaging methods, fNIRS provides solid results in both domains (Parasuraman & Rizzo, 2006; G. Strangman, Goldstein, Rauch, & Stein, 2006). The great strength of the method is that, depending on the fNIRS device used, experimental tasks can be performed while sitting, standing, or even in motion, which makes the obtained results ecologically far more valid than results from experiments in which participants have to lie down, as during fMRI measurements. Furthermore, compared to EEG, this method is less susceptible to artifacts caused by participants' movements, electro-oculographic and facial electromyographic activity, as well as electrical environmental noise - which might be particularly useful when measuring neuronal activation during human-machine interactions (see Derosière, Mandrick, Dray, Ward, & Perrey, 2013).

1.2.4.2 Eye-Tracking

Eye-tracking is another physiological method which is becoming increasingly popular for assessing cognitive load. Although it can be considered an indirect method, the subtle realization of modern video-based eye-tracking systems that do not require direct physical contact with user (for an overview see: Hutton, 2019) makes this technique potentially promising for real-world developments and well-suitable for commercial approaches. Based on empirical evidence, described below, fixations, blinks, saccades, microsaccades, and pupil diameter were used as indices for participants' cognitive load in this dissertation.

***Fixations*.** Voluntarily controlled stable gazes lasting from 200 - 300 milliseconds to up to several seconds are called fixations. During these periods eyes stay relatively still, while the person processes information from the fixation area (Pouget, 2019). The relation between cognitive load and fixation duration seems to depend on the task at hand. Empirical evidence indicates that increased task complexity is associated with fewer but longer fixations (for reviews see: Clifton Jr et al., 2016; Rayner, 1998). For instance, S. Chen, Epps, Ruiz, and Chen (2011) concluded that increased fixation duration as well as decreased fixation rate may indicate increased attentional effort on a more demanding task. Similarly, De Rivecourt, Kuperus, Post, and Mulder (2008) found that increased task complexity was associated with longer fixations on the control instruments during simulated

flight. Contrarily, Van Orden, Limbert, Makeig, and Jung (2001) observed fixation frequency to systematically increase with the visual complexity of a target classification task. Thus, it seems that for visual tasks the fixation rate increases with task difficulty whereas non-visual task demands lead to a decrease in fixation rate.

Saccades. Eye movements between two fixations that allow for exploration of the surroundings and attention control are called saccades. Empirical evidence indicated that that in non-visual tasks saccadic rate and decrease with task difficulty (Nakayama, Takahashi, & Shimizu, 2002), whereas visual complexity seems to increase saccadic rates (Benedetto, Pedrotti, & Bridgeman, 2011; He, Wang, Gao, & Chen, 2012).

Microsaccades. If our eyes would stay completely still during fixation, the visual image would gradually fade because neural response weakens with constant stimulation (Pouget, 2019). Microsaccades are small unintentional eye movements, which cover less than 1° of visual angle and prevent currently viewed visual information from fading. Evidence suggests that microsaccadic frequency increases with increasing visual complexity of the task at hand (Benedetto et al., 2011), whereas in non-visual tasks microsaccadic rate seems to decrease and microsaccadic magnitude to increase with task difficulty (Gao, Yan, & Sun, 2015; Siegenthaler et al., 2014).

Blinks. A commonly known function of blinking consists of keeping the eyeball moist and protecting it from physical damage. Besides that, in addition to microsaccades, blinking is also needed to prevent perceptual fading (Alexander & Martinez-Conde, 2019). Moreover, bursts of blinks seem to occur before and after periods of intense information processing (Siegle, Ichikawa, & Steinhauer, 2008), whereas, high blink rates were found to be associated with high cognitive load (Nakayama et al., 2002).

Pupil dilation. This metric is most commonly considered in cognitive load research (for a general overview see: Andreassi, 2013). In states of high cognitive load, pupil diameter was repeatedly observed to increase proportionally both in visual and non-visual tasks (Fukuda, Stern, Brown, & Russo, 2005; He et al., 2012; Klingner, Tversky, & Hanrahan, 2011).

1.2.5 Multimodal approach

The brief overview provided above shows that a variety of measurement methods can be used to assess cognitive load. However, a perfect single measurement method capable of capturing all facets of cognitive workload, preferably in real time, simply does not exist. In recent years, the trend is moving towards the development of complex multimodal measurement systems to capture cognitive workload (Herff et al., 2014; Ikehara & Crosby, 2005; Zhou et al., 2020). For instance, Zhou et al. (2020) examined cognitive workload during a simulated robotic surgery, whereby the researchers collected different physiological data from wireless sensors and compared the predictive power of different classifiers. The best classification accuracy was achieved for the multimodal approach, which accounted for 83% of variance in cognitive workload. A combination of eye-tracking and EEG was found to be advantageously for studying cognitive load during learning with multimedia materials, because it allows for fixation specific analysis of brain activation, while eye-tracking features such as pupil dilation allow for cognitive load estimation and can be used as an additional measure (Scharinger, 2018; Scharinger, Kammerer, & Gerjets, 2015; Scharinger, Schüler, & Gerjets, 2020).

Because subjective and performance-based measurement can be hardly applied online during task processing, a combination of (neuro-) physiological and behavioral measures appears to be promising in the context of adaptation to cognitive load. However, depending on the techniques used, the resulting multimodal measurement methods might become complex and expensive and therefore poorly suited for practical use. Moreover, data-driven approaches such as machine learning solutions as described by Gerjets, Walter, Rosenstiel, Bogdan, and Zander (2014) mostly require specific calibration and thus cannot be easily generalized to different subjects and situations.

This raises the question of whether this limitation can be solved by selecting suitable measurement methods that are as simple as possible to collect and process, and analyzing them using a suitable theoretical framework that exploits knowledge about the nature of the task being performed as well as the cognitive structure of users. Accordingly, when trying to apply the proposed measurement system to similar settings, it might be possible to avoid additional calibration and thus achieve better generalizability of the method. As mentioned above, the concept of cognitive load is founded in the recognition of the limited nature of

human cognitive system and it is therefore closely relation to the concept of WM. In the following, the most influential contemporary theoretical models of WM are presented and discussed in terms of finding an appropriate theoretical foundation for a top-down measurement of cognitive load.

1.3 Cognitive load in working memory models

The idea of working memory (WM) is based on the view that all human reasoning is based on some common system, regardless of the problem to be solved. There are many different models of WM, and although they differ in details and their backgrounds, most of them are very consistent in their main propositions (cf. Baddeley, 2012; Malmberg, Raaijmakers, & Shiffrin, 2019). Depending on how the known limitations of WM and thus cognitive load are explained, all models can be roughly divided into two main classes: theories that emphasize memory vs. attention as a limited resource.

Baddeleys' multicomponent model (Baddeley & Hitch, 1974) is an example of a theoretical framework that emphasizes the role of memory in constraining human cognition. It is one of the most influential models of WM, which updated the concept of short-term memory from the modal model of Atkinson and Shiffrin (1968) and was based on more recent neurophysiological findings from patient studies. Extending the concept of short-term memory, the model suggests that WM combines both functions of storage and processing of information. The initial version of the multicomponent model assumed three components of WM. The central executive was considered to act as a central master system that controls two sensory slave systems with limited capacity: the phonological loop and the visuospatial sketchpad, which serves to temporally store information from different modalities. It is important to notice that both slave systems possess their own rehearsal mechanisms to keep information active for some time (e.g., inner speech production is the rehearsal mechanism used by the phonological loop). Later the model was enhanced with an episodic buffer (Baddeley, 2000), which served as a passive storage for integrated information (e.g., processing of words bounded in sentences or combinations of colors and shapes to colored objects). The model assumes that each component is limited in capacity, namely in the amount and duration of information processing (Baddeley & Logie, 1999),

whereas individual differences result from differences in underlying brain structures (e.g., due genetic factors of brain damages) and expertise (e.g., fluency in inner speech production, which serves as the rehearsal mechanism in the phonological loop) by both integrating information (e.g., process a word as a single chunk of information instead of processing each character separately) and increasing processing speed. Thus, according to this model, cognitive load imposed by a task at hand can be seen as the extent to which all components are filled with processed information. Under heavy load, resources intended for storage of elements are completely occupied, whereby there is no capacity left and the performance deteriorates.

Similarly, cognitive load theory (Sweller et al., 1998), which is based on Baddeleys' multicomponent model and focuses on instructional design, describes cognitive load as the extent to which the learners' limited cognitive resources (i.e. their working memory storage component) are occupied. The theory distinguishes between three different aspects of cognitive load. Intrinsic cognitive load refers to the difficulty of the task, i.e. the number of elements that must be processed simultaneously. Extraneous cognitive load describes the unnecessary load caused by irrelevant or poorly-designed instructional materials. Germane cognitive load describes the effort that the learner invests in building long-term knowledge. It relies on the remaining cognitive resource after intrinsic and extraneous load is imposed onto learners. For example, if the task itself is difficult and the instructional design is poor, all cognitive resources may already be occupied, leaving no resource for deeper knowledge construction, resulting in a decline in learning performance.

By contrast to these models focusing on the limited capacity of storage structures, other approaches emphasize the role of limited attentional resources as a basis of WM load. In the context of differential psychology, Engle and Kane (2004) consider WM as a part of a single long-term memory in which a limited amount of information can get activated for processing by means of directing controlled (or executive) attention to these information. According to these researchers, individual differences in WM capacity can be explained by means of differences in the ability to control attention, which therefore is an important predictor for a wide range of real-life tasks. Similarly, the embedded process model of WM by Cowan (1998) does not distinguish between long- and short-term memories but assumes that there is only one store for information, namely long-term memory, while short-term

memory is based on an activated part of long-term memory. Moreover, an attentional focus can be used to select and process a limited number of elements from the activated part of long-term memory, thereby constituting WM. In this model, WM is rather limited by the capacity and temporal limitation of this attentional focus, because the activation of information units decay within approx. 10-20 second without reactivation (Cowan, 1999).

In sum, temporal limitations are mentioned in most models of WM as an important aspect of limited WM capacity. However, these aspects are mostly related to the assumption that temporary memory traces decay without refreshing or reactivation - but in these models time restriction are not considered to directly affect WM functioning. At the same time, evidence from experiments in realistic HMI settings indicates that time pressure by itself may also play an important role in inducing cognitive load. To mention one example, Andreessen, Gerjets, Meurers, and Zander (2021) trained a passive brain-computer interface based on EEG to distinguish between high and low cognitive load and used it to validly predict the complexity of the presented text, which was read at different speeds. The authors found that increased reading speed was associated with higher cognitive load due to increased processing demands per time unit. Processing demands per time unit are focused on as an important source of WM load in the time-based resource-sharing model (TBRS), which is based on the assumption that time and time pressure plays a crucial role in WM functioning (Barrouillet et al., 2004).

1.3.1 Time-based resource-sharing model

As other models of WM, the TBRS is based on experiments with working memory tasks such as span tasks, developed by Daneman and Carpenter (1980). In these tasks, memorizing items (e.g., last words of sentences) is interleaved with processing items (e.g., verifying sentences). The number of items that can reliably be remembered by participants after a sequence is used as a measure of their working-memory span. This measure was found to be a very good predictor for performance on a wide range of tasks in realistic settings (for reviews see: Daneman & Merikle, 1996; Engle, Kane, & Tuholski, 1999). In order to study the role of time pressure for working-memory limitations, Barrouillet introduced paced span tasks that require participants to faster memorize or process items.

In contrast to preceding models, Barrouillet assumed that “the dual function of a working memory devoted to the active maintenance of information while concurrent processing is performed demands a subtle interplay between activation in which time plays the crucial role” (Barrouillet et al., 2004, p. 60). TBRS (for a comprehensive overview of the model and its development history, see Barrouillet et al., 2004; Barrouillet & Camos, 2015) describes WM as the core system of cognition dedicated to the processing and storage of information, whereby both storage and processing components of WM are required for executing a cognitive task. According to the model, both components require attention to switch between subtasks, which results in complex and time-critical interactions between them and eventually causes interruptions in their processing. In this way, researchers agree with the suggestion of (Kahneman, 1973) that complex activities impose high load not directly due to their complexity, but rather because time pressure is part of their nature. Therefore, even the simplest tasks can induce a high cognitive load if they are presented at a high pace and continuously demand the focus of attention.

This idea can be illustrated by a simple example. Considering an arithmetic task such as two-digit multiplication: the processing component would be occupied with arithmetic operations, while the storage component would be needed to memorize intermediate results. Similarly, in a reading task, one needs to remember the context of what is currently being read as well as decoding a sequence of words to understand the meaning of a new sentence. From these examples, it seems intuitively clear that processing components of WM requires attention, but at the same time one must also somehow “refresh” the intermediate results of processing by means of intentionally thinking about them. This means that attention must be shared between both components of WM. This idea is responsible for the second part of the models’ name as TBRS assumes that attention is a limited resource that must be shared in such a way that only one central process such as storage or processing of information can be performed at a time. As soon as attention is directed to the processing component (e.g., to an arithmetic operation), the stored information (e.g., an intermediate result from a previous calculation step) begins to fade from memory. This so-called decay of memory traces progresses during the time during which attention is captured due to the ongoing calculation process. However, the TBRS postulates that simultaneous task execution can be mimicked by rapid switches of attention between to-be-performed subtasks, potentially interrupting the processing component of the

current task (i.e. one may briefly interrupt a simple arithmetic operation to remind oneself of the intermediate result). This leads to a complex and time-critical interplay between executed processing and storage activities, yielding the notion that attention sharing happens in a time-based manner, which explains the full name of the TBRS model.

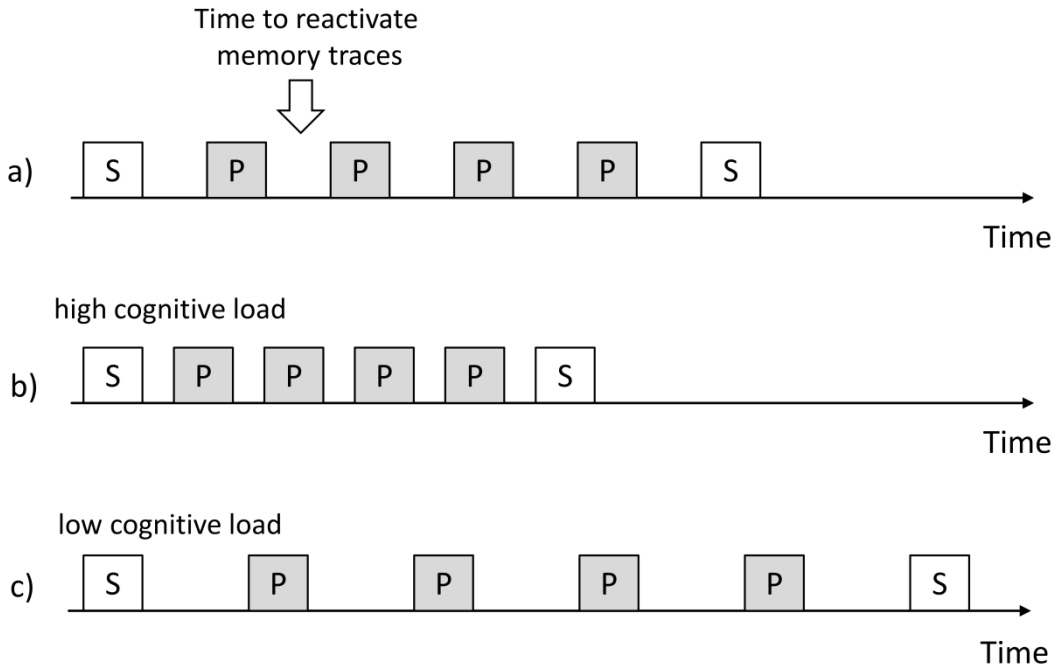


Figure 1. *Schematic representation of the emergence of cognitive load from varying speeds in a working memory span task. S: presentation of a character that has to be stored. P: processing components, e.g., reading letters (cf. Barrouillet & Camos, 2004).*

Returning to the example of a two-digit multiplication, what if a subject is perhaps very young and not very skilled in this type of task so that the arithmetic operation captures all of his attention without providing the storage process with any chance of refreshing intermediate results? After some time these results probably can no longer be retrieved from memory, so that further calculations would be rendered impossible, yielding a decline in performance. By contrast, for a very skilled subject, the arithmetic operations might be carried out in a more automated way requiring less attention, so that it would not be difficult to “refresh” intermediate results and show optimal performance.

Similarly, when executing simple processing activities such as reading letters in a paced WM span task (see Figure 1a), if the letters are presented at a comfortable pace, participants have sufficient time to reactivate memory traces and recall memorized items between their presentation. However, if the presentation pace becomes too fast (Figure 1 b), the time left for reactivation of memory traces might become insufficient, resulting in increased cognitive load and reduce performance. By contrast, slowing of the presentation pace (Figure 1 c) would result in lower cognitive load and correspondingly better performance. Taking these considerations into account, TBRS predicts that cognitive load and thus performance depends on the proportion of time during which attention is captured in such a way that the storage of information is disturbed, whereas individual differences in cognitive load and performance depend on the duration of processing units as well as reactivation speed.

1.4 Temporal action density decay metric

Unfortunately, it is not trivial to determine the exact time during which attention is captured by processing demands. For this reason, the TBRS model was evaluated using span tasks (Case, Kurland, & Goldberg, 1982; Daneman & Carpenter, 1980) along with a specifically designed paced span tasks, allowing for definition of certain retention and storage intervals at a predefined pace (Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007; Camos, Portrat, Vergauwe, & Barrouillet, 2007; Lépine, Bernardin, & Barrouillet, 2005; Liefooghe, Barrouillet, Vandierendonck, & Camos, 2008; Portrat, 2008). Although hard paced tasks can be found in HMI settings (e.g., playing Tetris) they do not represent a typical setting. More frequently, users find themselves in environments dominated by time pressure, for example, as a result of predefined time limits in the processing of intelligent test items or in situations such as critical emergency, in which although no time limit is usually known, but the manager is highly motivated to act as fast as possible to save lives. In such settings, the pace at which sub-tasks are processed is internally determined by the desire to complete the entire task within the prescribed (or assumed) time limit. Because the TBRS model specifically emphasizes the role of time pressure in WM functioning, the model seems appropriate for modeling such situations. At the same time, because the method has been validated using hard paced tasks in a controlled laboratory environment, it is unknown

whether it can be applied in realistic situations with self-defined pace induced by time pressure, as described above. So, how might cognitive load be quantified based on the TBRS model applied in such a situation?

In different persons, the same task can capture attention to varying degrees, depending on their cognitive resources, which may differ, e.g., through experience or training (Babiloni, 2019; Case et al., 1982). This means that under time pressure, a person experiencing lower demands on her/his attentional resources for the task-processing component may deliberately increase her/his processing speed (task density) without affecting his/her memory component, whereas a person experiencing higher attentional demands would not be able to do so. Accordingly, when presenting these two hypothetical persons with a block of certain tasks under time pressure (*action block*), one would observe two activity phases. In a first phase (*burst*), one would see both persons performing the presented tasks at a maximum speed. In a second phase (*idle*), they would have to wait until the end of the current *action block*, in other words until the subsequent *action block* begins.

$$temporal\ action\ density\ decay\ (TADD) = \frac{burst}{burst + idle} \qquad (Equation\ 1)$$

Assuming that both persons have operated at their limits, their cognitive load in the *burst* phase would be equivalent, namely at a maximum. By contrast, the duration of the *burst* phase would depend on their cognitive resources. Therefore, cognitive load of the entire *action block* could be estimated by the relation of the duration of the *burst* phase to the total duration of the *action block* (see Equation 1). In terms of the TBRS model, this implies that the person experiencing lower demands has more temporal processing resources left and might therefore also be able to solve more difficult tasks, whereas the other person has fewer resources left for the time-based sharing.

One possibility to apply this approach to realistic HMI using a serious game as an example is to define *burst* and *idle* periods based on in-game activities captured in log files for each participant. During the *burst* period, participants manage their emergency personnel, and after the last available personnel have been assigned to a task the *idle* interval starts and lasts until the first personnel finish their tasks and are available again.

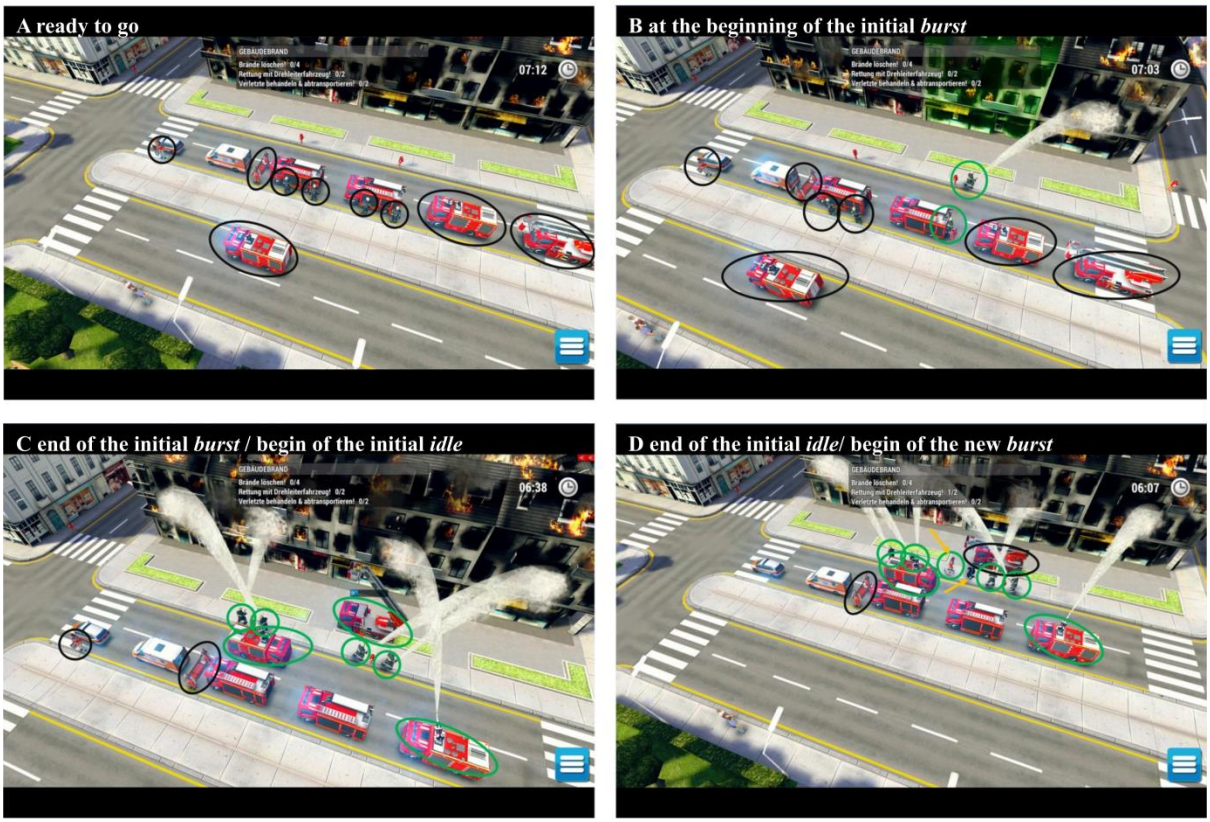


Figure 2. *Illustration of burst and idle time periods based on the example of an Emergency serious game. Inactive task forces are marked with black circles, and active ones with green.*

At the beginning of the game level, all emergency personnel are ready and inactive (Figure 2 A). Subsequently, the initial burst begins with the first task that the player assigns, which lasts until all available personnel are actively engaged. In this example, the emergency doctor and paramedics are not available for assignment because there are no injured people to be treated (Figure 2 B). The initial burst ends as soon as the last available personnel are assigned, and the emergency doctor and paramedics cannot yet be assigned. This time also marks the beginning of the initial idle, in which the player must wait until some personnel become available again or until new tasks occur. In this example, all available personnel are already active and the player must wait until the person is rescued from the burning building, only after which can he/she be treated by the doctor (Figure 2 C). The initial idle ends when the first personnel are available again. In this example, the initial

idle phase ends as soon as the rescued person appears lying on the road. At this moment, an emergency doctor becomes free and can be assigned to treat the patient. At the same time, the ladder truck also becomes free again and can be assigned to rescue the next person (Figure 2 D).

1.5. Research questions

1.5.1 Objectives of this dissertation

Adequate adaptation to cognitive load might help to reduce human errors and improve performance (Kohlmorgen et al., 2007; Walter et al., 2017; Yuksel et al., 2016). Specifically, in time-critical situations this might potentially save lives (Statistisches Bundesamt, 2017) and therefore holds crucial importance. Data-driven methods commonly used for such adaptation seem very promising, but at the same time they bring certain limitations: machine learning systems need to be trained on possibly large data sets and therefore cannot be easily transferred to a different situation and different users. This raises the question of whether this limitation can be solved by using an appropriate theoretical approach that would take advantage of knowledge about the nature of the task being performed and its structure.

The TBRS model of Barrouillet et al. (2004) emphasizes the crucial role of time on WM functioning and therefore appears to present a promising theoretical framework for this purpose. However, the model was developed and evaluated mainly for WM span and specific hard-paced experimental paradigms, and it remains unclear how it can be applied to a realistic situation. One example of a real-life situation that comes close to a hard-paced WM span task may be a time-critical management situation in which the pace is determined indirectly due to the reaction and execution time of available resources whenever an emergency manager is highly motivated to show maximal performance. However, in these situations separation into subtasks would be more difficult than in a WM span task.

In this dissertation, I propose a method for the application of TBRS on a time pressure situation as described in Section 1.4. My objective is to investigate whether predicted time periods (*burst*, *idle*, *action block*) can be found in the course of a serious

game simulating a time-critical emergency and, if so, whether the analysis of these periods can predict cognitive load and player performance in an expected way. By doing so, I hope to contribute to the development of adaptive HMI systems for real-world applications. Therefore, the measurement methods used were chosen to make this method potentially suitable for commercial applications that require unobtrusive realization and are typically run on rather small devices with low computing power.

1.5.2 Overview of conducted studies

In a series of three studies, the proposed approach was introduced and validated using behavioral, fNIRS and eye-tracking methods. To induce different levels of cognitive load in participants, a specifically customized version of the serious game Emergency (Promotion Software GmbH, 1999), which simulated two typical emergency scenarios at three levels of difficulty each, was used throughout the studies. Table 1 provides an overview of the entire project.

In the first study, the TBRS model was identified as a potentially suitable theoretical framework when modeling the management of time-critical emergencies. Accordingly, it was attempted to apply its theoretical predictions to a realistic HMI during a serious game by defining and calculating the *TADD* metric (see Section 1.3). Thereby, the *TADD* metric was presented in three different variations: (1) *initial TADD,* as the first *TADD*, calculated at the beginning of the level; (2) *mean TADD,* defined as mean value of all sequential *TADDs* calculated per level; and (3) as *normalized gaming time* defined as the relation of actual gaming time to the predefined time limit. In addition, the difficulty of the levels within the customized Emergency serious game was validated using subjective self-reports assessed using the NASA-TLX questionnaire.

Table 1. *Overview of conducted studies.*

Is it possible to measure cognitive load in real-time in a top-down manner?		
	Aim	Measurement domain
Study 1	Search for an appropriate theoretical framework and propose a practical measurement method.	behavioral data
Study 2	Does cortical activation correspond to the theoretical assumptions underlying the proposed method?	cortical activation (fNIRS)
Study 3	Do eye-tracking features correspond to the theoretical assumptions underlying the proposed method?	eye-tracking features (eye-tracking)

Based on the results obtained in the first study, the *initial TADD* metric was selected for further investigations. The further important step was to determine whether theoretical assumptions made when developing the metric were correct. This becomes an important issue when intending to generalize the method to other scenarios and settings. Because behavioral data is not sufficient to accurately determine the users' cognitive states, whereas subjective ratings do not reflect fluctuations in cognitive load over time (see Section 1.1), two additional studies involving (neuro-) physiological measurement methods were conducted. Thereby, two main questions were addressed. First, the important theoretical assumption used in defining the *initial TADD* metric was that due to the time pressure imposed, all participants should be working at their cognitive limit during the *initial burst* phase. Thus, no associations between (neuro-) physiological activation and difficulty level or performance were expected during this time period. Second, while participants were executing predefined tasks during the *initial burst* phase, which was captured in game logs,

no recordable actions were performed during *the initial idle* phase due to the nature of this time period. For this reason, it was intended to investigate this time period in further depth using (neuro-) physiological methods with the aim of better understanding what cognitive activities take place during this time and how possible differences in cognitive behavior among participants during the *initial idle* period are related to their cognitive load/outcome of the level.

Following the aforementioned considerations, in the second study a direct (according to the classification of Brunken et al., 2003) neuroimaging method fNIRS was used to measure cortical hemodynamics during the *initial burst* and *initial idle* periods. Building on empirical evidence suggesting that the prefrontal cortex is involved in decision-making processes (Thier, 2006) and that hemodynamic activation in this region is significantly associated with cognitive load (Ayaz et al., 2007; Ehlis et al., 2005; Fishburn et al., 2014), this area was specifically under focused on by analyzing activation in the dorsolateral prefrontal cortex (DLPFC) and inferior frontal gyrus (IFG), the part of the PFC involved in language processing.

To obtain a comprehensive understanding of the proposed method and evaluate it from different perspectives, the third study examined the aforementioned research questions once again, this time using the eye-tracking method. According to the same classification (Brunken et al., 2003), eye-tracking is considered an indirect method for measuring cognitive load, but at the same time it brings enormous benefits regarding its usability for commercial developments due to its discreet implementation. Based on empirical evidence showing that these features respond to variations in cognitive load, fixations, blinks, saccades, microsaccades, and pupil diameter were analyzed during the *initial burst* and *initial* idle periods in the third study.

Taken together, the three studies presented and evaluated a novel theory-based approach to measuring cognitive load in time-critical situations while providing a consistent comprehensive picture of the cognitive activities underlying the proposed metrics. In the process, the feasibility was demonstrated using various measurement domains, while at the same time new research perspectives were opened up to be explored. The authors' manuscripts of these studies are presented in the following sections.

2 Study 1

Please cite as:

Sevcenko, N., Ninaus, M., Wortha, F., Moeller, K., & Gerjets, P. (2021). Measuring cognitive load using in-game metrics of a serious simulation game. *Frontiers in Psychology, 12*, 906. doi:https://doi.org/10.3389/fpsyg.2021.572437

Measuring Cognitive Load Using In-Game Metrics Of A Serious Simulation Game

Authors:

Natalia Sevcenko (1, 2), Manuel Ninaus(3, 4), Franz Wortha (3, 4), Korbinian Moeller (5, 3), Peter Gerjets (3, 4)

Affiliations:

1 - Daimler Trucks AG, Stuttgart, Germany

2 - Psychology, Faculty of Science, Eberhard Karls University, Tübingen, Germany

3 - Leibniz-Institut für Wissensmedien, Tübingen, Germany

4 - LEAD Graduate School & Research Network, Eberhard Karls University, Tübingen, Germany

5 - Centre for Mathematical Cognition, School of Science, Loughborough University, Loughborough, UK

Keywords: cognitive load, in-game metric, adaptivity, serious games, simulation

Abstract

Serious games have become an important tool to train individuals in a range of different skills. Importantly, serious games or gamified scenarios allow for simulating realistic time-critical situations to train and also assess individual performance. In this context, determining the users' cognitive load during (game-based) training seems crucial for predicting performance and potential adaptation of the training environment to improve training effectiveness. Therefore, it is important to identify in-game metrics sensitive to users' cognitive load. According to Barrouillets' time-based resource-sharing model, particularly relevant for measuring cognitive load in time-critical situations, cognitive load doesn't depend solely on the complexity of actions but also on temporal aspects of a given task. In this study, we applied this idea to the context of a serious game by proposing in-game metrics for workload prediction that reflect a relation between the time during which participants' attention is captured and the total time available for the task at hand. We used an emergency simulation serious game requiring management of time-critical situations. 47 participants completed the emergency simulation and rated their workload using the NASA-TLX questionnaire. Results indicated that the proposed in-game metrics yielded significant associations both with subjective workload measures as well as with gaming performance. Moreover, we observed that a prediction model based solely on data from the first minutes of the gameplay predicted overall gaming performance with a classification accuracy significantly above chance level and not significantly different from a model based on subjective workload ratings. These results imply that in-game metrics may qualify for a real-time adaptation of a game-based learning environment.

1 Introduction

Serious games have become an important tool for educating and training people in a variety of different skills, ranging from military purposes to education and health care (for an overview see: Boyle et al., 2016; Susi, Johannesson, & Backlund, 2007); Unlike traditional analog learning, which cannot be automatically adapted to individual needs, serious games and simulations can be programmed to create targeted learning programs. While digital training in areas such as maths, language learning, exercise or healthy eating can easily be replaced by analogue setups, a range of situations such as aircraft crashes, surgical operations, or - more generally - time-critical emergency situations, can hardly be trained in real-life situations, it may benefit considerably from simulations and/or serious games. The most pronounced advantage of such digital training consists not only of the potential to simulate dangerous and time-critical situations, hard to recreate in analogue surroundings, but also of the fact that any digital training system also allows for the collection of individual in-game metrics (e.g., performance progression or computer mouse/keyboard usage) upon which learning analytics can be applied (Freire et al., 2016). Measures such as memory and learning outcomes may directly be used for an adjustment of difficulty levels of the learning environment. However, these outcome measures are usually only available after a particular task has been completed. In contrast, estimations of players' cognitive or emotional states based on in-game metrics (Nebel & Ninaus, 2019), might be used to adapt systems to increase training effectiveness, performance, and motivation. Among different affective and cognitive components, cognitive load seems to be particularly interesting as it is considered to reflect the degree to which available cognitive resources are engaged in the task at hand (Babiloni, 2019). As P. Gerjets et al. (2014) pointed out, the actual level of cognitive load is relevant in a variety of realistic settings, such as adaptive learning environments, where optimal learning content is characterized by an intermediate level of cognitive load. The researcher showed that the learners' cognitive load while solving complex realistic tasks can be classified by analyzing electroencephalography (EEG) data using machine learning algorithms. Moreover, previous results indicated that adaptations based on measured cognitive load can lead to significant learning improvements comparable to effects of failure-based adaptations, even when a generalized prediction model without user-specific calibration is used (Walter et al., 2017).

In the current study, we used a serious simulation game for training emergency personnel with the aim to assess participants' cognitive load by in-game metrics using a theory-driven approach. Below we provide a brief overview of cognitive load and its measurement methods. This is followed by a more detailed description of the time-based resource-shared model of Barrouillet et al. (2004), which provides the theoretical foundation for our approach on in-game metrics measuring cognitive load before we describe the details of the current study and hypotheses.

1.1 Cognitive load and adaptation to cognitive load

The concept of cognitive load goes back to the finding that working memory capacity is limited to approx. seven chunks of information (G. A. Miller, 1956), and thus cognitive resources, in general, are limited. According to the definition of Paas and Van Merriënboer (1994), cognitive load is a multidimensional construct and represents demands that a particular task imposes on the cognitive system. While this definition offers a good initial idea of the construct, the theoretical details of how cognitive load should be precisely conceptualized are still under discussion. Thus, even though the research on cognitive load has a long history (Barrouillet et al., 2004; Eggemeier et al., 1985; Linton et al., 1978; Meshkati, 1988; Sheridan & Simpson, 1979; Sweller et al., 1998; Welford, 1978) it's still a scientifically vibrant field of interest given its crucial importance for everyday life. As noted by Babiloni (2019) in his recent review on the topic, cognitive load can be characterized by a complex interplay between different task demands and a variety of mental processes such as alertness, vigilance, fatigue, etc., and thus represents a result of a complex interaction of different aspects. That is, cognitive load is a dynamic variable that may change rapidly during task processing. Nevertheless, three general assumptions regarding the construct of cognitive load can be derived from the literature (cf. Babiloni, 2019). First, human cognitive and attentional resources are limited. Second, different tasks can require different cognitive resources to varying degrees. And third, different individuals may experience different levels of cognitive load when conducting a task even when achieving the same performance level on it.

Ample evidence emphasizes the importance of cognitive load in our everyday life. For instance, cognitive load plays a crucial role in performing everyday activities such as learning/education (Ruiz et al., 2010), car driving (G. Hancock et al., 2012; Kohlmorgen et al., 2007), rail industry (Fan & Smith, 2017b), air force (P. Hancock, 1989), office work (Smith-Jackson & Klein, 2009), and medicine (Yurko et al., 2010). Thus, accurately measuring cognitive load seems of considerable importance for a better understanding of the fluctuations in human performance.

According to an influential theoretical account, the relationship between cognitive load and performance is non-linear and can be described following an "inverted-U" shaped function (Babiloni, 2019; Veltman & Jansen, 2005), see also Yerkes and Dodson (1908). Importantly, the general idea of this "inverted-U" shaped relationship is also closely related to the concept of "flow" proposed by Csikszentmihalyi (1975). Flow is described as a positive emotional and cognitive state (Kiili et al., 2018) of optimal concentration and absorption. The state of flow is achieved when there is a good balance between the demands of a given task and the perceived skills and resources of an individual to solve the task. That is, a given task should not be too difficult (i.e. cognitive overload) or too easy (i.e. cognitive underload and boredom) to elicit a flow state allowing for optimal performance. Consequently, optimal learning content should be moderately challenging but should neither induce cognitive over- nor underload. The very same consideration is also reflected in classical theories of instructional design. So moderately challenging optimal training state corresponds to the "zone of proximal development" (cf. Vygotsky (1980) and "amount of invested mental effort" cf. Salomon (1984).

Empirical evidence substantiated this theorized relationship between cognitive load and performance. For instance, Cummings and Nehme (2009) evaluated the relationship between cognitive load and performance of operators supervising multiple unmanned vehicles during a simulation of a military mission. In a series of two experiments, they showed that the addition of non-linear parabolic components into their performance prediction model improved its predictive power significantly. This demonstrated the non-linear character of this relationship and indicates that individuals perform best at medium levels of cognitive load (e.g., Anderson, 1994; Montani, Vandenberghe, Khedhaouria, & Courcy, 2020; Watters, Martin, & Schreter, 1997). These results generalize to educational

environments as well as serious games and reflecting that to achieve best learning outcomes learners should be kept in an intermediate range of cognitive load where they are not bored (Pekrun, Goetz, Daniels, Stupnisky, & Perry, 2010) but also not overstrained (Chang, Warden, Liang, & Lin, 2018; Geng & Yamada, 2020; Niederhauser, Reynolds, Salmen, & Skolmoski, 2000).

In this way, it becomes clear that an ideal learning environment should not only be tailored to specific needs of the learners (P. H. Gerjets & Hesse, 2004; Richards et al., 2007), for instance distinguishing between different expertise levels (cf. "expertise reversal effect" by Kalyuga, 2007). But also needs to consider that cognitive load is a dynamic variable that depends on different cognitive processes and may change during task accomplishment. Therefore, in order to keep learners within an optimal intermediate range of cognitive load, such systems should be able to identify undesirable states of under- and overload in real-time and adapt an ongoing task accordingly. In this way, performance and learning outcomes might be optimized.

Empirical evidence suggests that such online adaptation is indeed practicable (Appel et al., 2019; P. Gerjets et al., 2014) and can improve performance. For instance, Kohlmorgen et al. (2007) examined whether an adaptive reduction of cognitive load would lead to improved performance in a real-world driving task. Using electroencephalography (EEG) they were able to detect drivers' cognitive overload and to adapt to it accordingly by making the task easier. In turn, this led to improved driving performance. Similarly, Yuksel et al. (2016) reported better performance as a result of adapting task difficulty to cognitive load. They used near-infrared spectroscopy (NIRS) to detect states of cognitive underload in pianists during a musical learning task and increased difficulty of the respective lessons accordingly. Moreover, Walter et al. (2017) developed a learning environment that adapted task difficulty based on EEG recordings reflecting the cognitive load of learners. Optimal cognitive load was deduced from EEG data and was not individually calibrated. Nonetheless, this system led to learning outcomes similar to that observed for error-based adaptation.

These examples indicate the growing popularity of this approach and its importance for future studies. However, they also point to the diversity of measurement techniques in

this field. The following section introduces and classifies different ways of measuring cognitive load.

1.1.1 Measurement of cognitive load

Cognitive load assessment techniques that might be used to guide adaptations to cognitive load should be able to respond sensitively to variations in cognitive demands of the task at hand or interaction with learning systems without causing external disturbances to performance on the primary task (Orru & Longo, 2019). The literature distinguishes between four main categories of cognitive load measurement techniques: subjective measures, performance measures, behavioral measures, and physiological measures (Brünken et al., 2010; Eggemeier et al., 1991; Johannsen, 1979; Scerbo, 1996).

1.1.1.1 Subjective measures. Subjective measurements are based on the observation that people are able to interpret and adequately describe their experienced cognitive load during a particular task (Gopher & Braune, 1984). These self-reported descriptions are collected using questionnaires such as SWAT (Reid & Nygren, 1988) and NASA-TLX (Hart & Staveland, 1988), which require participants to rate their experiences using predefined scales immediately after completing a specific task. Subjective measures are easy to collect, they are inexpensive and they usually provide consistent results (O'Donnell & Eggemeier, 1986). Therefore, these measures are widely accepted and have been thoroughly evaluated. Despite their advantages, subjective measurements have also a number of limitations. The main issue is that responding to a questionnaire interrupts task execution and thus can only be carried out after the task has already been completed, which has some potentially confounding consequences. Firstly, a retrospective view of an experienced cognitive load may be distorted by fading memory. Secondly, experienced failures (or successes) can bias the post-hoc perception of cognitive load (P. Hancock, 1989). Thirdly, only a rough summary of the experience can be grasped in this way, which is not capable of tracking fine variations of cognitive load over time. And finally, self-reported measurements are only able to reflect conscious aspects of the cognitive load experienced during task accomplishment.

1.1.1.2 Performance measures. Performance-based approaches evaluate variations in human performance. Based on empirical evidence, performance should decrease in case of cognitive overload (Babiloni, 2019; Veltman & Jansen, 2005; Yerkes & Dodson, 1908). Accordingly, a drop of performance may help to detect cognitive overload. As a main objective of cognitive load measurement is the prediction of task performance, this cluster of measurement techniques appears intuitively to be the most obvious and direct to apply. Unfortunately, it cannot be determined whether observed variations in performance have actually occurred due to changes in cognitive load or due to other relevant factors such as arousal or motivation (Brünken et al., 2010). Therefore, these measures yield no independent assessments of cognitive load for performance prediction. Moreover, in many cases it is not possible to obtain performance data during actual task completion, so that performance-based measurements can very often only be calculated and analyzed post-factum, rendering them useless for prediction or adaptation.

1.1.1.3 Behavioral measures. Behavioral measures rely on the analysis of differences in interaction behavior during task processing, such as speech and voice patterns (Berthold & Jameson, 1999; Magnusdottir et al., 2017; Ruiz et al., 2010) or differences in the usage of input modalities such as keyboard or mouse (Ikehara & Crosby, 2005; Lim et al., 2015).These measures are usually unobtrusive and do not distract participants from the task at hand. Moreover, they do not require additional equipment and are usually inexpensive. Behavioral measures potentially allow for a continuous online measurement of cognitive states during task execution. However, identifying in the data related to cognitive load behavioral patterns is by no means a trivial endeavor, as these behavioral patterns might also be influenced by other factors such as emotions or stress.

1.1.1.4 Physiological measures. Physiological measures of cognitive load rely on detecting physiological changes associated with cognitive states (Johannsen, 1979). Depending on the type of signal to be recorded, they can be more or less obtrusive. While sensors for electrodermal activity (EDA) or heart rate variability (HRV) can be rather discreet, electroencephalography (EEG) or functional magnetic resonance imaging (fMRI)

are less practical or even impracticable in real-life situations because of their complexity, immobility, and obtrusiveness (for an overview see Ninaus et al., 2013). One major advantage of physiological measures is that they allow for continuous online recording. However, physiological measures require special equipment, cause additional costs, and the detection of cognitive states based on physiological signals is also not a trivial task (Appel et al., 2019; P. Gerjets et al., 2014). Because physiological processes are not only driven by cognitive states but can also be influenced by a variety of other factors, such as motor actions or emotions, it is not always unambiguously clear whether a change in a physiological signal was actually caused by the targeted cognitive state (Kramer, 1991). Moreover, physiological signals often require user-specific calibrations due to the signals' high inter-subject variability.

1.1.1.5 Conclusion. While there seem to be numerous methods for measuring cognitive load, a perfect single assessment approach capable of capturing all relevant facets of cognitive load, preferably in real-time, simply does not exist. In recent years, a trend towards the development of complex multimodal measurement systems to capture cognitive load can be observed (Herff et al., 2014; Ikehara & Crosby, 2005; Zhou et al., 2020). However, due to their inherent complexity, multimodal approaches seem to be primarily useful for extensive online data acquisition in the laboratory. In real-world scenarios outside the laboratory, such as gameplay, it seems reasonable to focus on metrics that on the one hand reflect users' behavior and performance and on the other hand can be easily collected during gameplay without requiring additional equipment. In view of future developments, such simple but reliable metrics might also become part of more complex monitoring systems. However, as argued above, changes in users' behavior and performance do not necessarily directly reflect changes in cognitive load, so that a solid theoretical framework for the development of such metrics will be needed. In this paper, we will rely on the time-based resource-sharing (TBRS) model described below to provide a suitable theoretical basis for assessing cognitive load based on behavioral and performance measures in time-critical multitasking environments requiring simultaneous execution of several tasks under severe time constraints.

1.2 The time-based resource-sharing model

TBRS (for a comprehensive overview of the model and its development history see Barrouillet et al., 2004; Barrouillet & Camos, 2015) describes working memory as the core system of cognition dedicated to the processing and storage of information, whereby both storage and processing components of working memory are required for the execution of a cognitive task. This idea can be illustrated by a simple example. Considering an arithmetic task, such as two-digit multiplication, the processing component would be occupied with arithmetic operations, while the storage component would be needed to memorize intermediate results. Similarly, in a reading task, one needs to remember the context of what is currently being read as well as to decode a sequence of words to understand the meaning of a new sentence.

From these examples, it seems intuitively clear that processing components of working memory requires attention, but at the same time one must also somehow "refresh" the intermediate results of processing by means of intentionally thinking about them. That means that attention must be shared between both components of working memory. This idea is responsible for the second part of the models' name as TBRS assumes that attention is a limited *resource* that must be *shared* in a way that *only one* central process such as storage or processing of information can be performed at a time. As soon as attention is directed to the processing component (e.g., to an arithmetic operation), the stored information (e.g., an intermediate result from a previous calculation step) will begin to fade from memory. This so-called decay of memory traces is progressing in the time during which attention is captured due to the ongoing calculation process. However, TBRS postulates that simultaneous task execution can be mimicked by rapid switches of attention between to-be-performed subtasks - potentially interrupting the processing component of the current task (i.e., one may briefly interrupt a simple arithmetic operation to remind oneself of the intermediate result). This leads to a complex and time-critical interplay between executed processing and storage activities yielding that attention sharing happens in a *time-based* manner, which explains the full name of the TBRS model: time-based resource-sharing model.

Coming back to the example of a two-digit multiplication, what if a subject is perhaps very young and not very skilled in this type of task so that the arithmetic operation captures

all of his attention without providing the storage process with any chance of refreshing intermediate results? Probably, after some time these results cannot be retrieved from memory anymore so that further calculations would be rendered impossible, yielding a drop in performance. In the contrary, for a very skilled subject, the arithmetic operations might be carried out in a more automated way requiring less attention, so that it would not be difficult to “refresh” intermediate results and show optimal performance. Taking these considerations into account, TBRS predicts that cognitive load and thus performance will depend on the proportion of time during which attention is captured in such a way that the storage of information is disturbed.

As Barrouillet et al. (2004) emphasizes, it is unfortunately not trivial to determine the exact time during which attention is captured by processing demands. Moreover, as the model was developed and evaluated mainly for working-memory span tasks (Case et al., 1982; Daneman & Carpenter, 1980), it’s further evaluation and extension to other executive functions required the design of specific experimental paradigms, allowing for defining certain retention and storage intervals at a predefined pace (Barrouillet et al., 2007; Camos et al., 2007; Lépine et al., 2005; Liefooghe et al., 2008; Portrat, 2008).

As such fine-tuned and hard-paced settings are hardly present in everyday life, two questions arise: Can the model also be applied to more realistic setups, and if so, how should such setups look like. A real-life situation that comes closest to a hard-paced working-memory span task may be a computer-based test with restricted execution time for particular subtasks. Another situation with inherited pacing could be a time-critical management situation in which the pace is determined indirectly due to the reaction and execution time of available resources. However, as in both described situations, separation into subtasks would be more difficult than in a working memory span task, it remains unclear how the model and its prediction of the resulting cognitive load can be used in more general situations, such as serious games, where the pace is only indirectly determined while time pressure is still relevant. In this study, we aimed to address this question by proposing an in-game metric for measuring cognitive load based on the theoretical framework of TBRS.

1.3 The present study

Determining cognitive load during a serious game might be crucial for performance predictions as well as for providing adaptations to improve learning outcomes. In this study, we aimed at evaluating a practicable and parsimonious solution for cognitive load detection in serious games based on TBRS with regard to its reliability and potential suitability for online assessments and evaluations. To validate our approach we used commonly applied subjective reports of cognitive load as assessed by the NASA-TLX (Hart & Staveland, 1988). We focused on the use of in-game metrics based on users' behavior and performance as sources of information because these measures can be easily collected during gaming without extra equipment and provide relevant empirical evidence in terms of the time-based resource-sharing model Barrouillet et al. (2004). The validity of the proposed metrics for predicting cognitive load was evaluated in terms of their relation to the cognitive load as reported in the NASA-TLX and to the overall gaming performance. To implement sufficient variance in cognitive load we used an adaptation of a complex serious game simulating an emergency situation with different scenarios and levels of difficulty.

In particular, we pursued the following hypotheses. First, proposed measures of cognitive load based on in-game metrics as well as subjective self-report should validly reflect differences between various scenarios and levels of difficulty as a manipulation check. We expected that cognitive load should be higher in more difficult scenarios and levels as indicated by both in-game metrics as well as subjective ratings. Second, on an individual level we expected that cognitive load as indicated by the in-game metrics used should be associated significantly with participants' subjective rating of their cognitive load as measured by the NASA-TLX as well as with their overall gaming performance. Third, we hypothesized that the in-game metrics developed should allow for the prediction of overall gaming performance comparably well as subjective ratings provided by the NASA-TLX.

2 Methods

This study focused on measuring cognitive load with behavioral in-game metrics. It was carried out as part of a larger project that included several other physiological measures such as functional near-infrared spectroscopy (fNIRS), cardiac measurements, galvanic skin response, and eye-tracking (cf. Appel et al., 2019). As the aim of the current study was the evaluation of a simple and practicable parsimonious solution for cognitive load detection in serious games, the current analyses solely focused on behavioral and performance measures.

2.1 Participants

47 volunteers (33 females, 14 males) aged between 15 and 49 years (*M* = 24.6; *SD* = 6.4) participated in the study with most of them being students (95.7%). Informed consent was obtained from all participants or their parents when under the age of 18 (1 participant). All participants were right-handed, fluent in German, recruited via an online database, and compensated with 8 EUR for completing the study. The study was approved by the local ethic committee and a written informed consent was obtained. Participants reported no neurological, psychiatric, cardiovascular disorders, and did not take any psychotropic medications.

2.2 Task

Participants played a customized version of the serious game *Emergency* (Promotion Software GmbH, 1999), which provides simulations of different emergency situations. The game comprised different scenarios with three levels of difficulty each. During gaming, participants' task was to coordinate six types of emergency personnel, such as paramedics, emergency doctors, firefighters, ambulances, as well as fire- and ladder trucks, to rescue victims and extinguish fires.

The game was played from an isometric view where the viewing angle is shifted, creating a three-dimensional effect and showing some details of the environment that are not visible when viewed directly from above or from the side (see Figure 1). Participants had to choose

an appropriate command from an action menu by clicking on an available emergency force, and then select a target of the requested action by clicking on the desired object. For instance, participants clicked on an emergency doctor who would then be ordered to serve a respective victim, or on a firefighter who would be ordered to put out a fire or to free a person trapped in a car. Interaction with the game was realized using a conventional computer mouse only.

Figure 1. An example scene from scenario *Fire*.

After getting familiar with the game by playing an introductory tutorial and a training scenario, participants completed two target scenarios: *Fire* and *Train Crash*. Each scenario had to be played with three levels of difficulty: *easy*, *medium*, and *hard*. The difficulty levels and the scenarios differed with regard to the number of tasks to be accomplished and the number of personnel to be coordinated within a given period of time. At the beginning of

each level the number of tasks was equal for all players. Whereas, the number of victims was held constant, which means that no new victims were added during a game, the number of fires depended very much on the performance of players and therefore could grow rapidly (i.e., by fires spreading to adjacent buildings or objects if not extinguished). As the increasing task-density across levels and scenarios required not only more actions but also better coordination and prioritization, we expected that cognitive load of participants would increase with increasing difficulty of levels and scenarios. Additionally, there was a time limit for each level and scenario to impose time pressure onto participants. A summary of all parameters describing the task difficulty of each scenario and each level can be found in Table 1.

Table 1
Overview over the initial game parameters

Scenario / Game Parameters	Difficulty		
	easy	medium	hard
Scenario: Fire			
Time limit (sec)	450	450	450
Tasks – total	*8+*	*13+*	*18+*
Victims	2	3	4
Fires	4+	7+	10+
Ladder Rescues	2	3	4
Resources –total	*9*	*12*	*15*
Doctors	1	2	2
Paramedics	1	2	2
Fire Fighters	4	4	6
Fire Trucks	2	3	4
Ladder Trucks	1	1	1
Scenario: Train Crash			
Time limit (sec)	600	600	600
Tasks – total	*20+*	*30+*	*40+*
Victims	10	15	20
Cars to cut	7	10	13
Fires	3+	5+	7+
Resources – total	*10*	*14*	*18*
Doctors	2	3	4
Paramedics	3	5	6
Fire Fighters	4	4	6
Fire Trucks	1	2	2

Note: The number of fires depended on players' performance and might grow. These cases are marked by the '+' sign.

2.2.1 Training scenario

The learning sequence involved a car accident at an intersection. The players' task was to free all persons trapped in the crashed vehicles, treat them for health issues and transport them to the hospital. The time limit for this scenario was set to 5 minutes.

2.2.2 Fire

In this scenario, participants had to fight a burning building block. In addition, some residents had to be freed from the burning house, treated for health issues, and transported to the hospital. The number of fires varied depending on the players' performance in extinguishing fires and could eventually increase rapidly. The time limit was set to 7.5 minutes.

2.2.3 Train Crash

This scenario depicted a train crashing into a building, causing a quick-spreading fire. The task was to free trapped passengers from the train, treat them for health issues, and then transport them to the hospital. At the same time, numerous fires had to be extinguished. In this scenario, the number of fires also varied depending on the players' extinguishing performance. An additional difficulty was to protect emergency doctors working near a fire. The time limit for this scenario was set to 10 minutes.

2.3 Measures

In the current study, cognitive load was measured by means of two methods. The objective estimation of cognitive load was performed using behavioral in-game metrics, which were defined in line with the time-based resource-sharing model by Barrouillet et al. (2004). For the validation of these metrics, we acquired a subset of the NASA-TLX

questionnaire as a widely accepted and thoroughly evaluated subjective instrument (i.e., *mental demand*, *time demand*, and *effort*). The details for both assessment strategies are provided below. Gaming performance was reflected by a binary indicator of whether the game was completed successfully within a given time limit or not. Additional personal information on participants such as age, gaming experience, and sex were collected prior to the experiment using a self-report questionnaire. To measure gaming experience, we asked participants to indicate how often they play (online) digital games on a 5-point Likert scale ("never", "several times a year", "several times a month", "several times a week", "every day").

2.3.1 Behavioral in-game metrics

According to the TBRS Model (Barrouillet et al., 2004), working memory represents the core system of cognition dedicated to the processing and storage of information, whereby both storage and processing components are normally required for the execution of a cognitive task. In situations with pre-defined pace, cognitive load can be estimated as a relation between the time during which participants' attention is captured by the processing of information and the total time available. This model was well evaluated on modified span tasks with a pre-defined pace (Barrouillet & Camos, 2015). In the current study, we applied this metric to a more general situation where the pace is only indirectly determined by the nature of the task and inherent time pressure.

The same task can capture attention to varying degrees in different persons, depending on their cognitive resources, which may differ, e.g., through experience or training (Babiloni, 2019; Case et al., 1982). This means that under time pressure, a person experiencing lower demands on her/his attentional resources for the task-processing component may deliberately increase her/his processing speed (task density) without affecting his/her memory component, whereas a person experiencing higher attentional demands would not be able to do so. Accordingly, when these two hypothetical persons were presented with a block comprising a certain number of tasks under time pressure (*action block*), one would observe two activity phases. In a first phase (*burst*) one would see both persons performing the presented tasks at a maximum speed. In a second phase

(*idle*), they would have to wait until the end of the current *action block* until the subsequent *action block* begins. During the idle phase both persons can only observe how their actions during the *burst* played out.

$$temporal\ action\ density\ decay\ (TADD) = \frac{burst}{burst + idle} \qquad (Equation\ 1)$$

Assuming that both persons have operated at their limits, their cognitive load in the *burst* phase would be equivalent, that is, at maximum. In contrast, the duration of the *burst* phase would be different. Therefore, cognitive load of the entire *action block* could be estimated by the relation of the duration of the *burst* phase to the total duration of the *action block* (see Equation 1). In terms of the TBRS model this implies that the person experiencing lower demands has more temporal processing resources left and might therefore also be able to solve more difficult tasks whereas the other person has lesser resources left for time-based sharing. We transferred these assumptions to the situation of the game or gameplay, respectively. As a result, the following three in-game metrics were derived.

2.3.1.1 Normalized gaming time. The most obvious, but also the most basic option is to work with time-limited levels and to consider the entire level as an *action block*, while the *burst* phase would correspond to the factual gaming time and the *idle* phase to the time remaining until the end of the level. Based on this consideration, the total cognitive load for the entire level could be estimated. As this metric equals 1 for persons who failed at a game level and has a potential range between 0 and 1 for those who complete the respective level, it directly represents success in the game or level, respectively. Therefore, it can be seen as a performance in-game metric, which however can only be calculated retrospectively once the level has been completed.

2.3.1.2 TADD. A more fine-grained option would be to take a closer look at the course of the game action and to try identifying smaller *action blocks* within each level. This can be done by means of the following rationale: In the game, participants have to coordinate a set of tasks to be accomplished by a set of emergency personnel by prioritizing tasks and resources as quickly as possible (*burst* phase). When no more resources are available (i.e., when all emergency personnel are distributed to existing tasks and busy), an inevitable break occurs (*idle* phase). This *idle* phase lasts until the first emergency personnel are ready to take up a new task (beginning of the *burst* phase of the new *action block*). This theoretical approach can be applied to a range of different learning scenarios that can be found in (game-based) simulations where tasks have to be prioritized and teams/resources to be managed, e.g., utilizing elements of (real-time) strategy games for training managerial skills (Simons, Wohlgenannt, Weinmann, & Fleischer, 2020), computer programming (Muratet, Torguet, & Jessel, 2009), or mathematics problem solving (Hernández-Sabaté, Albarracín, Calvo, & Gorgorió, 2016).

2.3.1.3 Initial TADD. For predictive (and adaptive) purposes, it would be ideal to base cognitive-load estimations on very early *action blocks* within each level of a game. Therefore, we defined the *TADD* calculated for the very first *action block* of each game level as the *initial TADD*. The initial TADD comprises the time from the first user action until the first assigned emergency personnel becomes free again. The advantage of this measure is that it can be calculated during the first minutes of the gameplay and thus be used for near-real-time predictions and adaptations.

2.3.1.4 Mean TADD. In addition to the *initial TADD* we also calculated a *mean TADD*, reflecting the average of *TADD* for all identified *action blocks* per level. This metric can, of course, also be calculated only retrospectively and was used mainly for an additional validation of *initial TADD*.

2.3.2 NASA-TLX

The NASA-TLX (Hart & Staveland, 1988) is a multidimensional instrument for the assessment of subjective workload, with good psychometric properties and a very high degree of acceptance in the research community (Hart, 2006). It consists of six items, estimating different aspects of subjective workload from 0 to 100 points with steps of 5 points, resulting in a 21-level scale. The dimensions of the NASA-TLX correspond to various theories that distinguish between physical, mental, and emotional demands imposed on an operator (Hart, 2006). For the current study, we relied on a subset of these items to specifically assess the mental facet of workload, i.e. *mental demand*, *temporal demand*, and *effort.* Using various subsets of items is quite common when investigating specific facets of workload (Haerle et al., 2013; Temple, Dember, Warm, Jones, & LaGrange, 1997). Moreover, focusing on specific items allows for a time-efficient assessment of participants' workload, which was particularly important for the current study as subjective workload was assessed after each level of difficulty for each scenario to be able to associate behavioral and subjective indices of cognitive load for the different scenarios and difficulty levels.

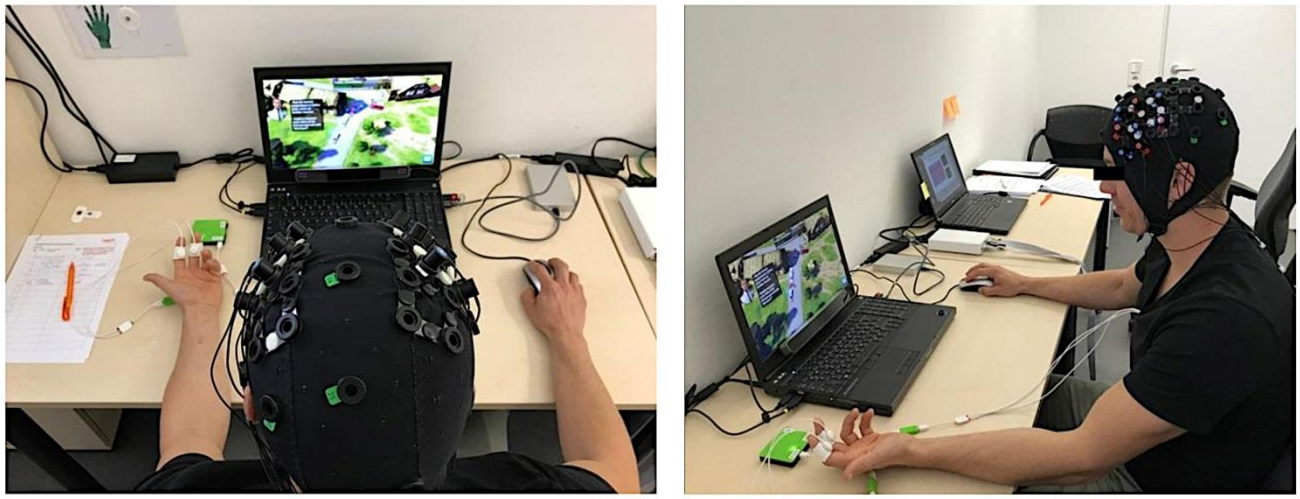


Figure 2. Experimental setup.

2.4 Experiment procedure

The study took place in a quiet laboratory under constant light conditions. The serious game was presented on a notebook with 16" screen providing a 1920 x 1080 resolution (see Figure 2). All instructions were presented in German. The study was implemented in a within-subject design, which means that each participant completed all scenarios and levels. The game started with an introductory tutorial and a training scenario directly after welcoming participants and collecting demographic data. Each game level was followed by a brief assessment of subjective cognitive state through an adapted NASA-TLX survey.

3 Results

3.1 Statistical Analyses

For statistical analyses we used R (R Core Team, 2020) with the *lme4* package (Bates et al., 2014) to perform generalized linear mixed-effects analyses as well as the *multcomp*, *emmeans* packages (Hothorn et al., 2008; Lenth et al., 2019) to conduct the posthoc comparisons described below. The *p*-values were obtained by likelihood ratio tests of the full model with the effect in question tested against a reduced model without the effect in question and were specified in further model analyses. Tukeys' adjustment method was used for multiple comparisons. We used the *report* package (Dominique Makowski & Lüdecke, 2019) to support the description of the results. Standardized parameters were obtained by fitting a model on a standardized version of the dataset. Effect sizes were labeled following the recommendations by Funder and Ozer (2019) and (H. Chen, Cohen, & Chen, 2010) for linear and generalized linear models respectively. No obvious deviations from homoscedasticity or normality were revealed using visual inspection of residual plots.

The final composition of tested models was determined by pairwise likelihood ratio tests. Thereby, the null model, which only contained test subjects as a random factor, was stepwise extended by fixed effects for scenario, difficulty, gaming experience, age, and gender. According to this procedure consideration of gaming experience, age and gender did not improve model fit significantly beyond the model only incorporating fixed effects of scenario and difficulty. For this reason gaming experience, age and gender were not considered in further analyses. In all models we considered the effect of the two scenarios as random[1], because we were primarily interested in relations between in-game metrics, subjective ratings, and difficulty levels within scenarios, regardless of the gaming scenario (for an overview of the composition and main outcomes of mixed-effect analyses see Table 2).

[1] Two models did not converge with participants and scenarios as random effects. Therefore, we decided to run models with participants as the only random effect, considering scenario as fixed factor instead (see Table 2).

Table 2
Overview of the mixed model analyses performed

	outcome	effects		*p*
		fixed	random intercepts	
Manipulation check				
NASA-TLX	mental demand	difficulty	participant, scenario	<.001
	time demand	difficulty	participant, scenario	<.001
	effort	difficulty	participant, scenario	<.001
Performance	failure/success	difficulty	participant, scenario	<.001
In-game metrics vs. subjective cognitive workload (NASA-TLX)				
normalized gaming time (NGT)	mental demand	NGT	participant, scenario	<.001
	time demand	NGT, scenario	participant	<.001
	effort	normed GT	participant, scenario	<.001
initial TADD	mental demand	initial TADD	participant, scenario	<.001
	time demand	initial TADD	participant, scenario	<.001
	effort	initial TADD	participant, scenario	<.001
mean TADD	mental demand	mean TADD	participant, scenario	.001
	time demand	mean TADD	participant, scenario	.003
	effort	mean TADD	participant, scenario	<.001
In-game metrics vs. gaming performance				
	failure/success	initial TADD	participant, scenario	<.001
	failure/success	mean TADD, scenario	participant	<.001

Note: Gaming performance was represented by the binary indicator of whether the game was completed successfully, i.e. all fires extinguish and all injured persons transported to the hospital (success) or not (failure). The *p*-values were obtained by likelihood ratio tests of the full model with the fixed effect in question against the reduced model without the fixed effect in question

For an overview of correlations among variables assessed in the current study, please see Table 3. Prediction of game performance was conducted using the Python (Van Rossum & Drake, 2009) module *scikit-learn* (Pedregosa et al., 2011). In particular, we used linear discriminant analysis with Leave-One-Subject-Out-Cross-Validation to train and test the models and permutation tests for model comparisons.

3.2 Manipulation check

3.2.1 Subjective ratings

To test whether the experimentally induced levels of task difficulty of the game are reflected in subjective workload measurements we ran linear mixed-effect models with the fixed factor difficulty and random intercepts for participants and scenarios on the relationship between selected items of NASA-TLX (*mental demand, time demand, effort*) and levels of difficulty (*easy, medium, hard*). For a general overview of all collected parameters see Appendix, Table 3.

3.2.1.1 Mental demand - difficulty

Linear mixed-effect analysis revealed a significant main effect of difficulty ($\chi^2(2) = 79.87$, $p < .001$) on the subjective rating of mental demand. The models' total explanatory power was substantial (conditional $R^2 = 0.82$, marginal $R^2 = 0.06$). Within this model perceived *mental demand* was significantly higher for *medium* difficulty levels compared to *low* difficulty levels, this effect can be considered as small (*beta* = 7.23, *SE* = 1.42, *std. beta* = 0.32, $p < .001$; see Figure 3); also, perceived *mental demand* was significantly higher for *high* difficulty levels compared to *low* difficulty levels. This effect can be considered as medium (*beta* = 13.83, *SE* = 1.42, *std. beta* = 0.61, $p < .001$; see Figure 3). Posthoc comparisons showed significant differences for all combinations of difficulty levels. Participants rated their mental demand higher during levels with higher experimentally induced task difficulty.

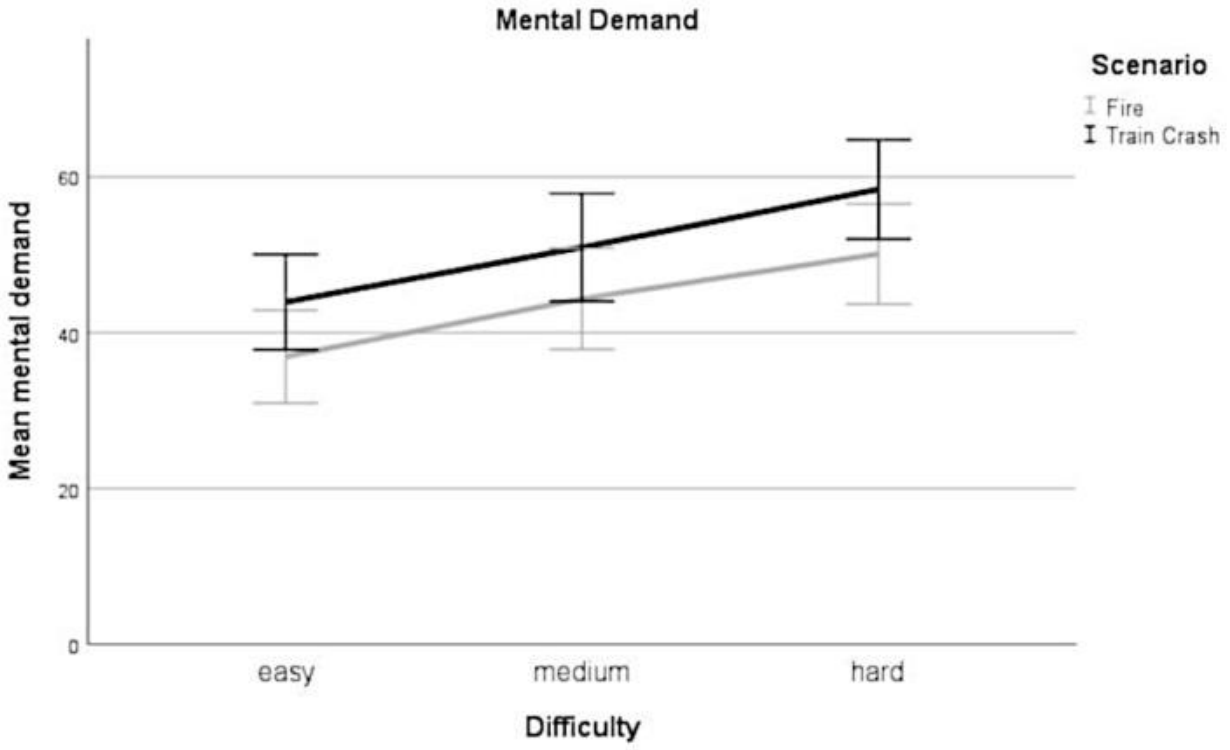


Figure 3. Mental Demand. Mean perceived mental demand for all levels of difficulty (*easy*, *medium*, *hard*) for each scenario (*Fire*, *Train Crash*). Error bars depict +/- 2 *SE*, which corresponds to 95% *CI*.

3.2.1.2 Time demand - difficulty

Linear mixed-effect analysis identified a significant main effect of difficulty on the subjective rating of time demand ($\chi^2(2) = 140.48$, $p < .001$). The models' total explanatory power was substantial (conditional $R^2 = 0.66$, marginal $R^2 = 0.23$). Perceived *time demand* was significantly higher for levels with *medium* difficulty compared to levels with *low* difficulty, this effect can be considered large (*beta* = 19.84, *SE* = 2.36, *std. beta* = 0.71, $p < .001$; see Figure 4); perceived *time demand* was significantly higher for *hard* difficulty levels compared to *low* difficulty levels, this effect can be considered very large (*beta* = 32.45, *SE* = 2.36, *std. beta* = 1.17, $p < .001$; see Figure 4). Post-hoc comparisons showed significant differences for all combinations of difficulty levels. Participants rated their time demand higher during levels with higher experimentally induced task difficulty.

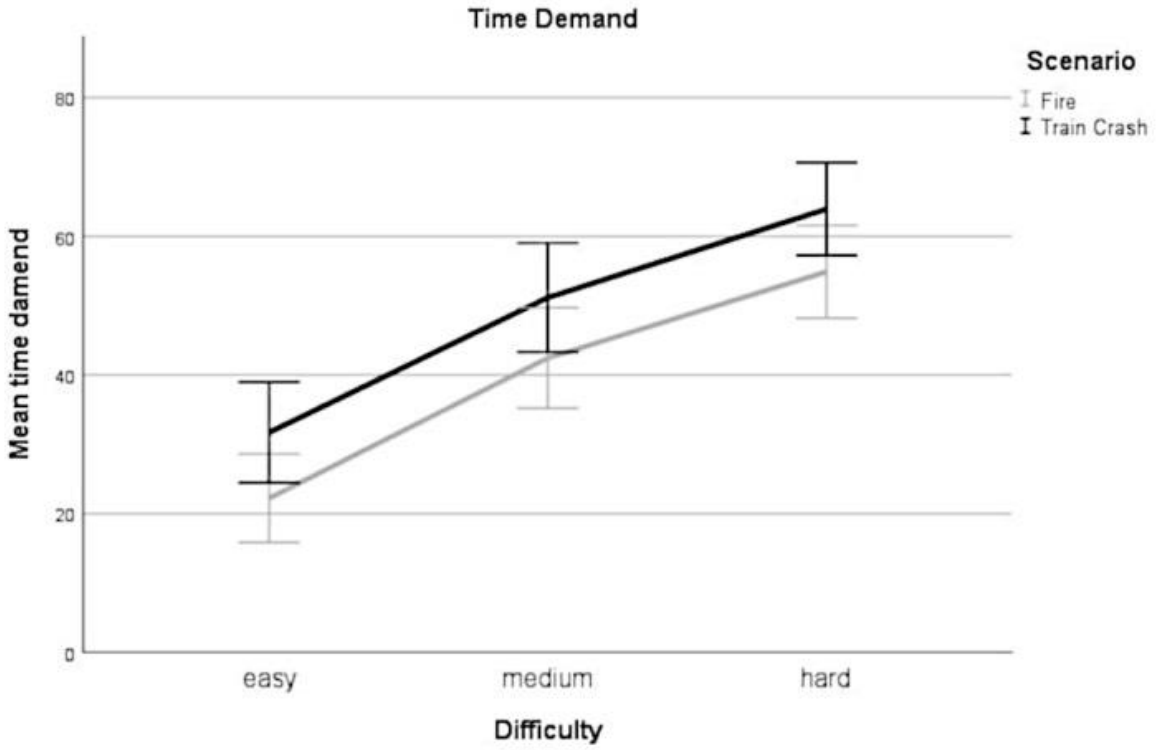


Figure 4. Time Demand. Mean perceived time demand for all levels of difficulty (*easy*, *medium*, *hard*) for each scenario (*Fire*, *Train Crash*). Error bars depict +/- 2 *SE*, which corresponds to 95% *CI*.

3.2.1.3 Effort - difficulty

Linear mixed-effect analysis showed a significant main effect of difficulty on the subjective rating of effort ($\chi^2(2) = 105.97$, $p < .001$).The models' total explanatory power was substantial (conditional $R^2 = 0.73$, marginal $R^2 = 0.13$). Within this model perceived mental *effort* was higher for *medium* difficulty levels compared to *low* difficulty levels, this effect can be considered as medium and significant (*beta* = 13.24, *SE* = 1.84, *std. beta* = 0.55, $p < .001$; see Figure 5), whereas perceived mental *effort* was higher for *hard* difficulty levels compared to *low* difficulty levels, this effect can be considered as very large and significant (*beta* = 21.01, *SE* = 1.84, *std. beta* = 0.87, $p < .001$; see Figure 5). Posthoc comparisons showed significant differences for all combinations of difficulty levels. Participants rated their effort higher during levels with higher experimentally induced task difficulty.

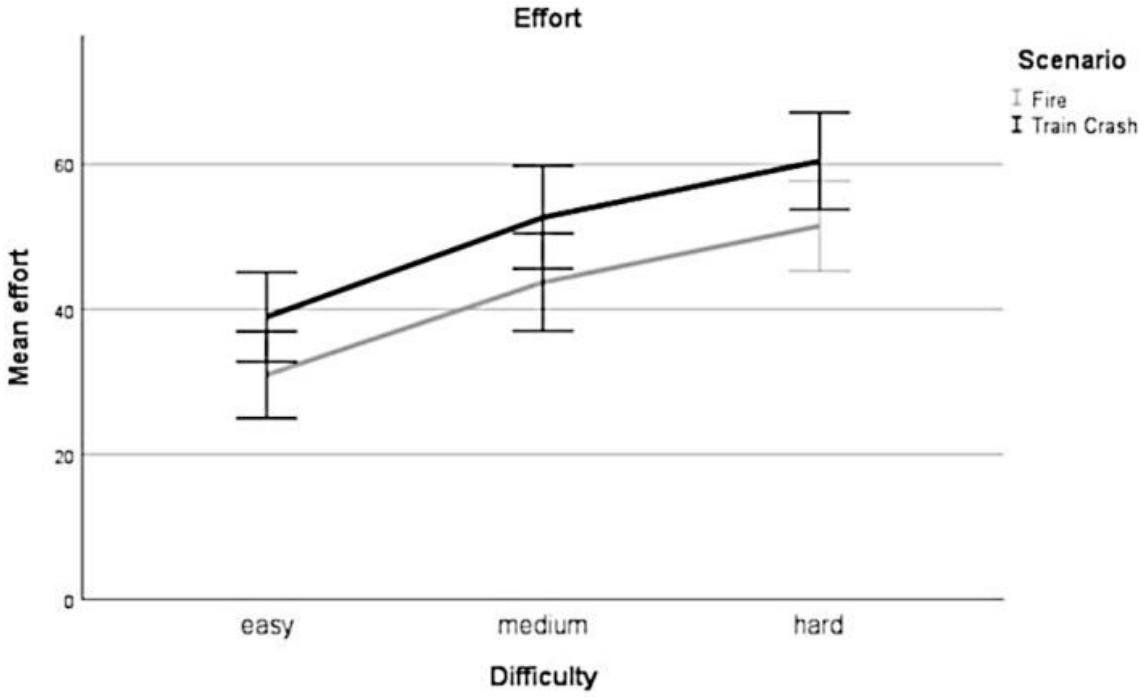


Figure 5. Effort. Mean perceived effort for all levels of difficulty (*easy*, *medium*, *hard*) for each scenario (*Fire*, *Train Crash*). Error bars depict +/- 2 *SE*, which corresponds to 95% *CI*.

3.2.1 Performance - difficulty

To evaluate, whether the levels of task difficulty are also reflected in gaming performance we fitted a logistic mixed-effect model on the relationship between the binary indicator of whether the game was completed successfully or not and the three difficulty levels. As we were primarily interested in the effect of difficulty levels, we considered difficulty as a fixed effect and added random intercepts for participants and scenarios. The generalized linear mixed-effect analysis revealed a significant main effect of difficulty ($\chi^2(2) = 115.39$, $p < .001$). The models' total explanatory power was substantial (conditional $R^2 = 0.67$, marginal $R^2 = 0.49$). Within this model we found that gaming performance was poorer for *medium* difficulty levels compared to *low* difficulty levels, this effect can be considered as large and significant (*beta* = -3.77, *SE* = 0.72, *std. beta* = -3.77, $p < .001$; see Figure 6); gaming performance was poorer for *hard* difficulty levels compared to *low* difficulty levels, this effect can be considered as large and significant (*beta* = -5.22, *SE* = 0.78, *std. beta* = -5.2, $p < .001$; see Figure 6). Post-hoc comparisons also showed significant differences for

all combinations of difficulty levels. Participants performed more poorly during levels with higher experimentally induced task difficulty.

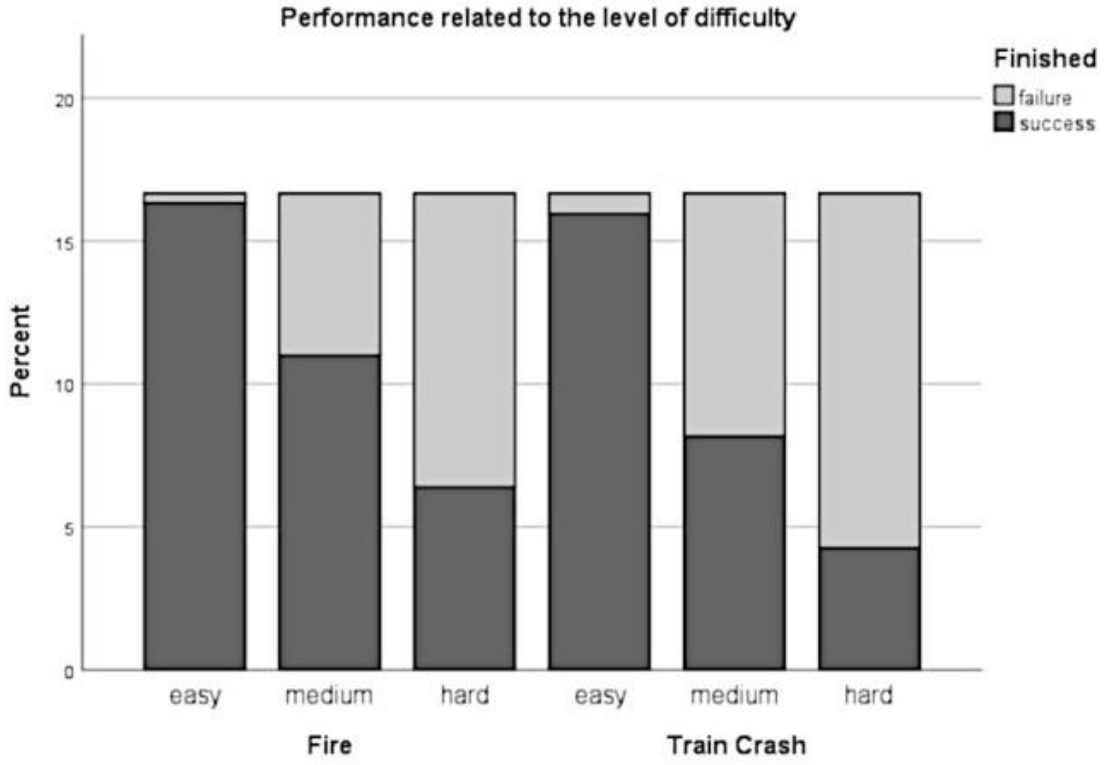


Figure 6. Performance in relation to the level of difficulty. Stacked histogram showing the percentage of successes/failures (i.e. whether the participants were able to rescue all victims and extinguish all fires within a defined time limit) for all levels of difficulty (*easy*, *medium*, *hard*) over both scenarios (*Fire* and *Train Crash*).

3.3 Subjective ratings vs. in-game metrics

To verify whether the calculated in-game metrics is able to predict the subjectively experienced cognitive load of participants we ran a linear mixed-effect model separately for each metric and the NASA-TLX item. As we were primarily interested in the relation between the in-game metrics and the subjective ratings regardless of the gaming scenario, we defined in-game metrics as fixed factors and added random intercepts for participants and scenarios.

3.3.1 Mental demand - normalized gaming time

Linear mixed-effect analysis indicated a significant effect of *normalized gaming time* ($\chi^2(1) = 104.33$, $p < .001$) on self-reported mental demand. The models' total explanatory power was substantial (conditional $R^2 = 0.83$, marginal $R^2 = 0.10$). Within this model, higher *normalized gaming times*, i.e. participants who took longer to finish the level or even failed, was associated significantly with higher perceived mental demand, this effect can be considered as small (*beta* = 38.73, *SE* = 3.38, *std. beta* = 0.31, $p < .001$).

3.3.2 Time demand - normalized gaming time

Linear mixed-effect analysis revealed a significant effect of *normalized gaming time* ($\chi^2(1) = 141.08$, $p < .001$) on self-reported time demand. The models' total explanatory power was substantial (conditional $R^2 = 0.72$, marginal $R^2 = 0.38$). Within this model we found that higher *normalized gaming times*, i.e. participants who took longer to finish a level or even failed, was significantly associated with higher perceived time demand, this effect can be considered as medium (*beta* = 0.91, *SE* = 5.59, *std. beta* = 0.60, $p < .001$); The effect of scenario was not significant (*beta* = 1.16, *SE* = 1.80, *std. beta* = 0.04, $p = .517$).

3.3.3 Effort - normalized gaming time

Linear mixed-effect analysis showed a significant effect of *normalized gaming time* ($\chi^2(1) = 125.65$, $p < .001$) on self-reported mental demand. The models' total explanatory power was substantial (conditional $R^2 = 0.73$, marginal $R^2 = 0.19$). Within this model we found that higher *normalized gaming time*, i.e. participants who took longer to finish the level or even failed, was significantly associated with higher perceived effort, this effect can be considered as medium (*beta* = 56.86, *SE* = 4.42, *std. beta* = 0.43, $p < .001$).

3.3.4 Mental demand – initial TADD

Linear mixed-effect analysis revealed a significant effect of *initial TADD* ($\chi^2(1) = 13.74$, $p < .001$) on self-reported mental demand. The models' total explanatory power was substantial (conditional $R^2 = 0.76$, marginal $R^2 = 0.01$). Within this model we found that higher *initial TADD*, i.e. participants who took longer to allocate the available personnel to the tasks to be done to during the first *action block*, was significantly related to higher perceived mental demand, this effect can be considered as very small (*beta* = 13.71, *SE* = 3.64, *std. beta* = 0.12, $p < .001$).

3.3.5 Time demand – initial TADD

Linear mixed-effect analysis identified a significant effect of *initial TADD* ($\chi^2(1) = 31.31$, $p < .001$) on self-reported time demand. The models' total explanatory power was substantial (conditional $R^2 = 0.45$, marginal $R^2 = 0.07$). Within this model we found that higher *initial TADD*, i.e. participants who took longer to allocate the available personnel to the tasks to be done to during the first *action block*, was associated with higher perceived time demand, this effect can be considered as small and significant (*beta* = 38.10, *SE* = 6.60, *std. beta* = 0.27, $p < .001$).

3.3.6 Effort – initial TADD

Linear mixed-effect analysis revealed a significant effect of *initial TADD* ($\chi^2(1) = 22.88$, $p < .001$) on self-reported mental demand. The models' total explanatory power was substantial (conditional $R^2 = 0.61$, marginal $R^2 = 0.04$). Within this model we found that higher *initial TADD*, i.e. participants who took longer to allocate the available personnel to the tasks to be done to during the first *action block*, was significantly associated with higher perceived effort, this effect can be considered as very small (*beta* = 23.89, *SE* = 4.88, *std. beta* = 0.30, $p < .001$).

3.3.7 Mental demand – mean TADD

Linear mixed-effect analysis showed a significant effect of *mean TADD* ($\chi^2(1) = 11.93$, $p < .001$) on self-reported mental demand. The models' total explanatory power was substantial (conditional $R^2 = 0.76$, marginal $R^2 = 0.01$). Within this model higher *mean TADD*, i.e. participants who in average took longer to allocate the available personnel to the tasks to be done to during all defined *action blocks*, was significantly associated with higher perceived effort, was significantly linked to higher perceived *mental demand*, this effect can be considered as very small (*beta* = 30.62, *SE* = 8.73, *std. beta* = 0.11, $p < .001$).

3.3.8 Time demand – mean TADD

Linear mixed-effect analysis revealed a significant effect of *mean TADD* ($\chi^2(1) = 8.83$, $p = .003$) on self-reported time demand. The models' total explanatory power was substantial (conditional $R^2 = 0.40$, marginal $R^2 = 0.02$). Within this model higher *mean TADD*, i.e. participants who in average took longer to allocate the available personnel to the tasks to be done to during all defined *action blocks*, was associated with higher perceived time demand, this effect can be considered as very small and significant (*beta* = 49.37, *SE* = 16.38, *std. beta* = 0.15, $p < .01$).

3.3.9 Effort – mean TADD

Linear mixed-effect analysis identified a significant effect of *mean TADD* ($\chi^2(1) = 13.31$, $p < .001$) on self-reported mental demand. The models' total explanatory power was substantial (conditional $R^2 = 0.59$, marginal $R^2 = 0.02$). Within this model higher *mean TADD*, i.e. participants who in average took longer to allocate the available personnel to the tasks to be done to during all defined *action blocks*, was significantly related with higher perceived effort, this effect can be considered as very small (*beta* = 43.97, *SE* = 11.85, *std. beta* = 0.15, $p < .001$.

3.4 Performance vs. in-game metrics

To verify whether the calculated in-game metrics would be able to predict the final performance of a given difficulty level, we ran a generalized linear mixed-effect models separately for the in-game metric *initial TADD* as well as for *mean TADD* and the binary indicator identifying whether the participants were able to extinguish all fires and transport all injured persons to the hospital (success) or not (failure). Since *normalized gaming time* basically was a performance measure, it predicts gaming success perfectly and cannot be used as a predictor variable in the mixed model. As we were primarily interested in the relation between in-game metrics and performance regardless of gaming scenario, we defined in-game metrics as fixed factors and added random intercepts for participants and scenarios.

3.4.1 Performance – initial TADD

Generalized linear mixed-effect analysis revealed a significant effect of *initial TADD* ($\chi^2(1) = 28.96$, $p < .001$) on performance. The models' total explanatory power was moderate (conditional $R^2 = 0.22$, marginal R^2 =of 0.14). Within this model we found that higher *initial TADD* was significantly linked to lower performance, this effect can be considered as small (*beta* = -3.86, *SE* = 0.79, *std. beta* = -0.76, $p < .001$; see Figure 7).

3.4.2 Performance – mean TADD

Generalized linear mixed-effect analysis indicated a significant effect of mean TADD on performance ($\chi^2(1) = 10.21$, $p < .001$). The models' total explanatory power was weak (conditional $R^2 = 0.13$, marginal $R^2 = 0.07$). Within this model the higher initial TADD was significantly associated with lower performance, this effect can be considered as very small (beta = -5.54, SE = 1.85, std. beta = -0.47, $p < .01$), whereas the effect of scenario was not significant (beta = -0.34, SE = 0.26, std. beta = -0.34, $p < .190$).

3.5 Performance prediction

To verify whether the *initial TADD* may be suitable for real-time or near-real-time prediction of performance (i.e., finished level successfully vs. failed) of the given level we used linear discriminant analyses with Leave-One-Subject-Out Cross-Validation. These demonstrated a 67.38% accuracy in scenario *Fire* and a 64.53% in the *Train Crash* scenario. However, permutation tests comparing the models' performance with the performance of models predicting randomly permuted outcomes showed that only for the *Train Crash* scenario this was significantly above the score of random models (*Fire*: random models mean accuracy: 67.28%, $p = .886$; *Train Crash:* random models mean accuracy: 55.37%, $p < .001$).

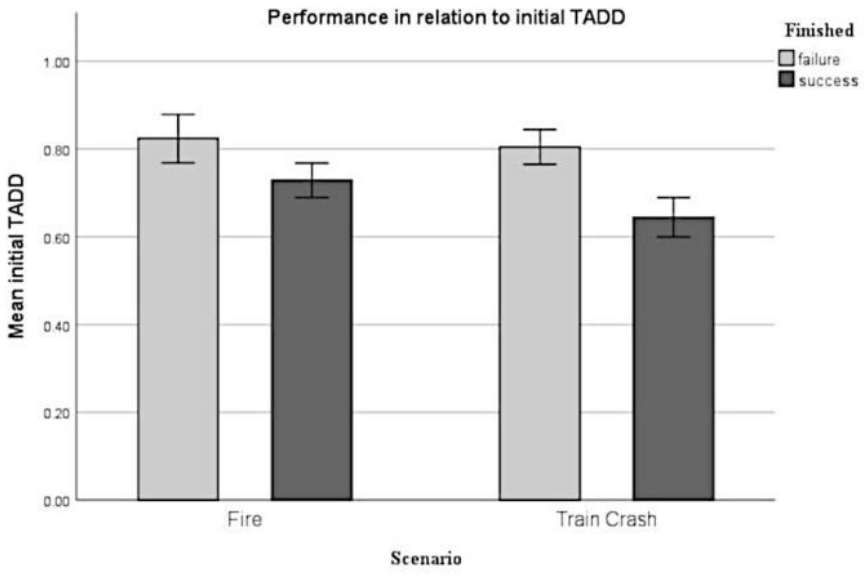


Figure 7. Performance in relation to initial TADD. Error bars depict +/- 2 SE, shich corresponds to 95% CI.

A linear discrimination analysis using the three NASA-TLX subscales in the performance scenario showed an average accuracy of 73.04%. A permutation test showed that this accuracy was significantly higher than models with randomly permuted outcomes (random models mean accuracy: 53.81%, $p < .001$). However, a permutation test showed that the accuracy of this model was not significantly different from the model using only our in-game metric ($p = .09$).

4 Discussion

The current study aimed at evaluating a practicable, parsimonious, and reliable approach for the online assessment of cognitive load in serious games, which is suitable for cognitive load prediction during realistic gaming setups with a similar structure to the game used in the study, i.e. a (real-time) strategy like serious game. Based on the time-based resource-sharing model of Barrouillet et al. (2004) we defined several in-game metrics (*initial TADD*, *mean TADD*, *normalized gaming time*) for describing the behavior and performance in an emergency simulation game. The results indicated that it seems indeed possible to use these simple in-game metrics to reliably assess and predict cognitive load based on a theory-driven approach. In the following, we critically discuss these results in greater detail.

First of all, we aimed at verifying whether the experimentally induced difficulty levels of the serious game were actually able to induce substantial differences in cognitive load for the participants. This manipulation check was an important prerequisite for examining our main scientific hypotheses. Results clearly indicated that increased difficulty (e.g., in terms of more personnel to coordinate and more tasks to execute under a constant time limit) indeed resulted in significantly higher subjective ratings of cognitive load accompanied by significantly poorer performance. This substantiated our expectations and indicated that the intended manipulation of cognitive load by means of game-difficulty levels worked as intended.

Further analyses showed that all three proposed in-game metrics can be considered valid to significantly predict self-reported workload as well as actual gaming success. In particular, *normalized gaming time* (i.e., ratio of actual playing time to total time available per game level) showed a more pronounced effect on subjective workload ratings as compared to *initial* and *mean TADD* (*temporal action density decay*; i.e. ratio of active playing time *burst* to total time available in the first *action block* and averaged over all *action blocks* defined per level). This may be due to the fact that the former measure was directly related to performance and thus most sensitive to subjectively experienced cognitive load. For instance, participants were able to develop a feeling for how well they have performed the game at the time of the survey and thus experienced failure may have resulted in higher perceived cognitive load as compared to known success (P. Hancock, 1989).

More interestingly, *initial TADD* showed a better predictive power as compared to *mean TADD* not only regarding subjective ratings of cognitive load but also in terms of the resulting performance. This suggests that early stages of gameplay may be more informative and thus more predictive for later gameplay outcomes than an aggregated score accumulated over a longer period of time. In this context, averaging *TADD* across the entire duration of the level seems to lead to a substantial loss of information for this metric.

A closer look at the construction of the game may help to better understand this difference. At the beginning of each level, a new emergency scenario was presented, and participants had to start assigning tasks to the available emergency personnel soon. That is, right at the beginning of the level participants had to orientate themselves in a completely new situation, to plan their rescue strategy, and to implement this strategy as quickly as possible. In addition, almost the entire rescue team had to be assigned to their tasks at this point, meaning that the first action block may have been significantly longer than all subsequent blocks, which were not as clearly defined due to more constant interactions with the game.

One possible explanation for the superiority of the *initial* over the *mean TADD* might be that as the game progressed, successful players realized that they were well in time and therefore experienced less time pressure. This might have resulted in longer *burst* and shorter *idle* phases, as they were not longer operating at their maximal speed, resulting in increased *TADD* ratios in the later stages of the level. Otherwise, it also seems conceivable that the initial orientation itself plays a crucial role in the outcome of the level. As better planning in early stages of a particular task was observed to be associated with better performance in various tasks (Saddler, Moran, Graham, & Harris, 2004) (Wang & Gibson Jr, 2010) (Capon, Farley, & Hulbert, 1994), *initial TADD* might also reflect more efficient planning to underlie decreased cognitive load during the initial *action block*. However, these assumptions need to be investigated in future studies.

The final aim of this study was to evaluate whether it would be possible to use in-game metrics for a real-time or near-real-time assessment of cognitive load and - based on this - a substantial performance prediction. The in-game metrics *normalized gaming time* as well as *mean TADD* represent summary measures, which can only be calculated retrospectively once a level has been completed. Thus, they cannot be used for predictive

purposes. In contrast, *initial TADD*, which was calculated during the first minutes of gameplay, significantly predicted gaming success - at least in the *Train Crash* scenario. Moreover, it could be shown that the prediction accuracy of a model using only this metric did not significantly differ from the model using selected NASA-TLX subscales as predictors for gaming performance. Interestingly, no significant prediction could be obtained for the less difficult scenario *Fire.* Importantly, however, this may have been influenced by a crucial data issue as far more participants succeeded in the scenario *Fire* than failed, whereas in the scenario *Train Crash* this relation was more balanced. Accordingly, the difference between the two scenarios may indicate an existing floor effect for the easy scenario, indicating that the use of this metric may be suitable only for situations eliciting phases of maximum cognitive load. Whether this assumption is correct must be investigated in future studies

In summary, the results of the current study indicated that gaming performance can be significantly predicted using *initial TADD* calculated from a short time interval at the very beginning of a new game level. This means that we were able to predict well above chance level whether the respective level would be completed successfully based on data acquired through the first tenth of the total gaming time. It is noteworthy that the quality of this prediction did not significantly differ from the prediction based on participants' retrospective and subjective ratings using the NASA-TLX that are informed by their experienced success of failure during game play. Hence, *initial TADD* seems to qualify well for a near-real-time adaptation of game flow, not requiring considerable computing power as it is the case for more data-driven approaches (e.g., neuronal networks or deep learning based on physiological data).

4.1 Methodological strengths and constraints

There are different analytical approaches to serious games (for review see: Zohaib, 2018), which are often based on data-driven probabilistic performance evaluations: (Magerko et al., 2006; Spronck et al., 2006; Zook & Riedl, 2012). Simple performance data, however, often seemed insufficient for estimating cognitive and emotional states of users, such as attention, cognitive load or emotional responses. Therefore, these cognitive states

are often assessed using (neuro-)physiological data (for reviews see: Kivikangas et al., 2011), which are, however, relatively complex and laborious to acquire and computationally intensive to evaluate and are thus not always suitable for real-world applications outside the laboratory. Importantly, though, the current study demonstrated that assessment/prediction of cognitive load using simple in-game metrics is feasible. We think that there are two crucial constraints for this approach to be successful: First, a theoretically informed top-down development and second the application within an appropriate test environment.

As regards the former, we are confident that a theoretical top-down approach may be key to find parsimonious, but still reliable and generalizable solutions. Therefore, a suitable theoretical framework should be chosen in the first place. In our case, the TBRS model (Barrouillet et al., 2004) specifically emphasizes the role of time pressure as the origin of cognitive load, therefore seeming to be particularly useful for predicting workload in time-critical situations such as serious game scenarios similar to the current one, i.e. (real-time) strategy games and simulations.

With respect to the latter, the development of an appropriate testing environment is essential. As, for instance, the TBRS model was originally evaluated on very specific tasks with strong time pressure induced through pre-defined pace, we evaluated whether its predictions may generalize to more realistic applied situations. In this way, we derived two critical aspects of a test situation to make these predictions work: time pressure and time-limited blocks of tasks. By considering these aspects, we designed a gaming environment that allowed for testing the proposed metrics.

4.2 Limitations and open questions

The methodological strengths and constraints of our study, however, can also be considered as limitations because it may not be possible to generalize the proposed metrics to all possible gaming situations. Presumably, they may well be used in settings with inherent time-limits and time pressure, where participants are exposed to new situations and have to manage various tasks and resources as it is the case in real-time strategy games. Other examples with similar task structures may comprise complex surgery tasks, assembly lines or time-critical emergency situations in the context of control tasks. Further

testing will be required in the future to substantiate the predictive power of proposed in-game metrics in this type of situation and, possibly, to adapt the computation of these metrics appropriately.

Therefore, the current study suggests a promising perspective, but at the same time raises several questions to be explored in the future. For instance, it is not clear whether and how the predictive power of the proposed metrics is related to the given time pressure and whether they can, therefore, be used in scenarios that are less time-critical. On the other hand it is possible that collected in-game metrics might be affected by factors other than cognitive load, such as motor processes related to the experience of the player with provided game controls, for instance. Since we used a conventional computer mouse as the only game control, we are confident that all participants were used to it and therefore the results obtained are valid in this respect. However, such general physiological processes should be taken into account and evaluated before proposed in-game metrics are generalized to different contexts. Furthermore, it should be evaluated more thoroughly why the *initial TADD* showed better performance as compared to the *mean TADD*. It might be possible, for instance, that the predictive value of the *mean TADD* (or the mean of the first few TADDs) can be improved by using other gaming situations or by sharpening the definition of *action blocks*.

4.3 Implications and future perspectives

The use of simple in-game metrics for measuring cognitive load and thus deriving performance prediction yields several advantages. First, our results suggest that psychological constructs, which have traditionally been assessed explicitly using either paper-pencil or computerized questionnaires, may well be estimated more implicitly using in-game metrics, that is without causing interruption to the task at hand (cf. stealth assessment: Shute & Kim, 2014). Second, whereas the use of more complex psychophysiological measures would come with additional computational and procurement costs, systems that operate on simple in-game metrics may be made more easily accessible to the general public. More complex systems relying on resources such as neural networks are computationally rather expensive and might require substantial

computing power. In contrast, simpler models for cognitive load estimation such as the one used in the current study may be easily run in parallel to the actual game on any PC without significant consumption of computing resources. Third, also complex multimodal measurement systems, which operate with sophisticated algorithms and integrate data from physiological and behavioral sources in research laboratories, may benefit from the development of simpler in-game metrics as these may be added to these more complex algorithms quite easily, thereby leading to improved classification accuracy in the future. Finally, the substitution of more complex probabilistic algorithms through simpler but reliable metrics (whenever possible) might lead to simplifications of complex models, while at the same time expanding their availability and usage. However, we feel that this may only be achieved when substantial evidence for relevant in-game metrics is based on theory rather than data alone.

4.4 Conclusion

The present study indicated that parsimonious, but theoretically well-founded in-game metrics can be used to estimate users' current cognitive load and, based on this, predict future gaming performance within the first tenth of the total gaming time. We applied our approach to a serious game simulating a time-critical emergency situation and requiring the management of emergency personnel. The game included different scenarios with three levels of difficulty each inducing corresponding levels of cognitive load. Based on proposed in-game metrics we were able to predict whether the respective level would be completed successfully or not well above chance level. Interestingly, the quality of this prediction did not differ significantly from a prediction based on participants' retrospective and subjective ratings using the NASA-TLX questionnaire. To achieve this we used a rather simple model that interprets behavioral data in the light of the TBRS theoretical approach (Barrouillet et al., 2004). Based on its parsimony and the corresponding low computational power required, this model can be easily incorporated into games to create an adaptive system. Further, the measure and models introduced in this study could be used in conjunction with other adaptive features to design even more comprehensive adaptive systems that can predict performance more effectively and accurately. Taken together, our results provide promising first evidence that needs to be substantiated in future research to determine

whether it is suitable for more general reliable assessments of players' cognitive load and for respective real-time adaptations of games or game-based learning environments.

Appendix

Table 3

Correlations matrix of variables considered in the present study

	2	3	4	5	6	7	8	*9*	*10*	*11*
1. Difficulty	- .636**	.265**	.529**	.397**	.000	.000	.000	.804**	.455**	.479**
2. Gaming success	-	-.322**	-.590**	-.465**	-.226**	-.147	.142	-.807**	-.342**	-.439**
NASA-TLX										
3. Mental demand		-	.747**	.900**	.190*	.122	-.004	.343**	.139	.185*
4. Time demand			-	.835**	.128	.087	-.026	.650**	.254**	.402**
5. Effort				-	.214*	.147	-.035	.482**	.201*	.269**
Covariates										
6. Age					-	-.053	-.233**	.176*	.210*	.171*
7. Sex						-	-.182*	.141	-.055	-.050
8. Gaming expertize							-	-.129	-.097	-.135
In-game metrics										
9. Normalized gaming time								-	.460**	.507**
10. Mean TADD									-	.606**
11. Initial TADD										-

Note: Pearson 2-tailed correlations. **: Correlation is significant at the 0.01 level; *: Correlation is significant at the 0.05 level. To give an overall impression over all collected values we conducted bivariate correlations presented in Table 2. The purpose of this summary is to give a first very general impression of the relations between the parameters, as presented values neither has been corrected for multiple comparisons, nor have repeated measurements been taken into account.

References

Anderson, K. J. (1994). Impulsivity, caffeine, and task difficulty: a within-subjects test of the Yerkes-Dodson law. Personal. Individ. Differ. 16, 813–829.

Appel, T., Sevcenko, N., Wortha, F., Tsarava, K., Moeller, K., Ninaus, M., et al. (2019). "Predicting cognitive load in an emergency simulation based on behavioral and physiological measures" in International Conference on Multimodal Interaction. eds. W. Gao, H. M. L. Meng, M. Turk, S. R. Fussell, B. Schuller, and Y. Song, et al. (New York, United States: Association for Computing Machinery), 154–163.

Babiloni, F. (2019) "Mental workload monitoring: new perspectives from neuroscience" in Human mental workload: Models and applications. H-WORKLOAD 2019. Communications in computer and information science. Vol. 1109. eds. L. Longo and M. Leva (Cham: Springer), 3–19.

Barrouillet, P., Bernardin, S., and Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. J. Exp. Psychol. Gen. 133, 83–100. doi: 10.1037/0096-3445.133.1.83

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., and Camos, V. (2007). Time and cognitive load in working memory. J. Exp. Psychol. Learn. Mem. Cogn. 33, 570–585. doi: 10.1037/0278-7393.33.3.570

Barrouillet, P., and Camos, V. (2015). Working memory: Loss and reconstruction. eds. H. Roediger, J. Pomerantz, A. D. Baddeley, V. Bruce, and J. Grainger (New York: Psychology Press).

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48. doi: 10.18637/jss.v067.i01

Berthold, A., and Jameson, A. (1999). "Interpreting symptoms of cognitive load in speech input" in UM99 user modeling. CISM international centre for mechanical sciences (Courses and Lectures). Vol. 407. ed. J. Kay (Vienna: Springer), 235–244.

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., et al. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. Comput. Educ. 94, 178–192. doi: 10.1016/j.compedu.2015.11.003

Brünken, R., Seufert, T., and Paas, F. (2010). "Measuring cognitive load" in Cognitive load theory. eds. R. Moreno and R. Brünken (Cambridge: Cambridge University Press), 181–202.

Camos, V., Portrat, S., Vergauwe, E., and Barrouillet, P. (2007). "The cognitive cost of executive functions" in Paper presented at the Joint Meeting of the EPS and the Psychonomic Society: Edinburgh (Great-Britain); July 4-7, 2007.

Capon, N., Farley, J. U., and Hulbert, J. M. (1994). Strategic planning and financial performance: more evidence. J. Manag. Stud. 31, 105–110. doi: 10.1111/j.1467-6486.1994.tb00335.x

Case, R., Kurland, D. M., and Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. J. Exp. Child Psychol. 33, 386–404.

Chang, C. -C., Warden, C. A., Liang, C., and Lin, G. -Y. (2018). Effects of digital game-based learning on achievement, flow and overall cognitive load. Australas. J. Educ. Technol. 34, 155–167. doi: 10.14742/ajet.2961

Chen, H., Cohen, P., and Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. Comm. Statist. Simul. Comput. 39, 860–864. doi: 10.1080/03610911003650383

Csikszentmihalyi, M. (1987). Das flow-Erlebnis: Jenseits von Angst und Langeweile: Im Tun aufgehen. Klett-Cotta.

Cummings, M. L., and Nehme, C. E. (2009). “Modeling the impact of workload in network centric supervisory control settings” in Paper presented at the 2nd Annual Sustaining Performance Under Stress Symposium; College Park MD; February 25, 2009.

Daneman, M., and Carpenter, P. A. (1980). Individual differences in working memory and reading. J. Mem. Lang. 19, 450–466.

Eggemeier, F. T., Shingledecker, C. A., and Crabtree, M. S. (1985). “Workload measurement in system design and evaluation” in Proceedings of the Human Factors Society Annual Meeting. Vol. 29. Los Angeles, CA: SAGE Publications Sage CA, 215–219.

Eggemeier, F. T., Wilson, G. F., Kramer, A. F., and Damos, D. L. (1991). “Workload assessment in multi-task environments” in Multiple-task performance. ed. D. Damos (London, Washington, DC: Taylor & Francis), 207–216.

Fan, J., and Smith, A. P. (2017). “The impact of workload and fatigue on performance” in Human mental workload: Models and applications. H-WORKLOAD 2019. Communications in computer and information science. Vol. 726. eds. L. Longo and M. Leva (Cham: Springer), 90–105.

Freire, M., Serrano-Laguna, Á., Manero, B., Martínez-Ortiz, I., Moreno-Ger, P., and Fernández-Manjón, B. (2016). “Game learning analytics: learning analytics for serious games” in Learning, design, and technology. eds. M. J. Spector, B. B. Lockee, and M. D. Childress (Cham: Springer), 1–29.

Funder, D. C., and Ozer, D. J. (2019). Evaluating effect size in psychological research: sense and nonsense. Adv. Methods Pract. Psychol. Sci. 2, 156–168. doi: 10.1177/2515245919847202

Geng, X., and Yamada, M. (2020). An augmented reality learning system for Japanese compound verbs: study of learning performance and cognitive load. Smart Learn. Environ. 7, 1–19. doi: 10.1186/s40561-020-00137-4

Gerjets, P. H., and Hesse, F. W. (2004). When are powerful learning environments effective? The role of learner activities and of students’ conceptions of educational technology. Int. J. Educ. Res. 41, 445–465. doi: 10.1016/j.ijer.2005.08.011

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., and Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. Front. Neurosci. 8:385. doi: 10.3389/fnins.2014.00385

Gopher, D., and Braune, R. (1984). On the psychophysics of workload: why bother with subjective measures? Hum. Factors 26, 519–532.

Haerle, S. K., Daly, M. J., Chan, H. H., Vescan, A., Kucharczyk, W., and Irish, J. C. (2013). Virtual surgical planning in endoscopic skull base surgery. Laryngoscope 123, 2935–2939. doi: 10.1002/lary.24004

Hancock, P. (1989). The effect of performance failure and task demand on the perception of mental workload. Appl. Ergon. 20, 197–205. doi: 10.1016/0003-6870(89)90077-x

Hancock, G., Hancock, P., and Janelle, C. (2012). The impact of emotions and predominant emotion regulation technique on driving performance. Work 41, 3608–3611. doi: 10.3233/WOR-2012-0666-3608

Hart, S. G. (2006). “NASA-task load index (NASA-TLX); 20 years later” in Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Vol. 50. Los Angeles, CA: Sage Publications CA, 904–908.

Hart, S. G., and Staveland, L. E. (1988). Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv. Psychol. 52, 139–183.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. (2014). Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS. Front. Hum. Neurosci. 7:935. doi: 10.3389/fnhum.2013.00935

Hernández-Sabaté, A., Albarracín, L., Calvo, D., and Gorgorió, N. (2016). “EyeMath: identifying mathematics problem solving processes in a RTS video game” in International Conference on Games and Learning Alliance. GALA 2016. Lecture Notes in Computer Science. Vol. 10056. eds. R. Jeuring and R. Veltkamp (Cham: Springer), 50–59.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. Biom. J. 50, 346–363. doi: 10.1002/bimj.200810425

Ikehara, C. S., and Crosby, M. E. (2005). “Assessing cognitive load with physiological sensors” in Proceedings of the 38th Annual Hawaii International Conference on System Sciences. New York: IEEE.

Johannsen, G. (1979). “Workload and workload measurement” in Mental workload. Vol. 8. ed. N. Moray (Boston: Springer), 3–11.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. Educ. Psychol. Rev. 19, 509–539. doi: 10.1007/s10648-007-9054-3

Kiili, K., Lindstedt, A., and Ninaus, M. (2018). “Exploring characteristics of students emotions, flow and motivation in a math game competition” in Paper presented at the GamiFIN. GamiFIN Conference 2018; May 21-23, 2018, Pori, Finland.

Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., et al. (2011). A review of the use of psychophysiological methods in game research. J. Gaming Virtual Worlds 3, 181–199. doi: 10.1386/jgvw.3.3.181_1

Kohlmorgen, J., Dornhege, G., Braun, M., Blankertz, B., Müller, K. -R., Curio, G., et al. (2007). “Improving human performance in a real operating environment through real-time mental workload detection” in Toward brain-computer interfacing. Vol. 409422. eds. G. Dornhege, J. d. R. Millan, T. Hinterverger, K.-R. Müller, and M. D. Childress (Cambridge, Massachusetts, London, England: MIT Press), 409–422.

Kramer, A. F. (1991). “Physiological metrics of mental workload: a review of recent progress” in Multiple-task performance. ed. N. Moray (London, Washington, DC: Taylor & Francis), 279–328.

Lenth, R., Singmann, H., Love, J., Buerkner, P., and Herve, M. (2019). Package “emmeans”: Estimated Marginal Means, aka Least-Squares Means. Compr. R Arch. Netw Available at: https://cran.r-project.org/web/packages/emmeans/

Lépine, R., Bernardin, S., and Barrouillet, P. (2005). Attention switching and working memory spans. Eur. J. Cogn. Psychol. 17, 329–345. doi: 10.1080/09541440440000014

Liefooghe, B., Barrouillet, P., Vandierendonck, A., and Camos, V. (2008). Working memory costs of task switching. J. Exp. Psychol. Learn. Mem. Cogn. 34, 478–494. doi: 10.1037/0278-7393.34.3.478

Lim, Y. M., Ayesh, A., and Stacey, M. (2015). “Using mouse and keyboard dynamics to detect cognitive stress during mental arithmetic” in Intelligent Systems in Science and Information 2014. SAI 2014. Studies in Computational Intelligence. Vol. 591. eds. K. Arai, S. Kapoor, and R. Bhatia (Cham: Springer), 335–350.

Linton, P., Jahns, D., and Chatelier, P. (1978). Operator workload assessment model: An evaluation of a VF/VA-V/STOL system. AGARD Methods to Assess Workloads 12 p (SEE N 78-31745 22-54).

Magerko, B., Stensrud, B. S., and Holt, L. S. (2006). Bringing the schoolhouse inside the box-a tool for engaging, individualized training. Available at: https://apps.dtic.mil/sti/pdfs/ADA481593.pdf (Accessed February 12, 2021).

Magnusdottir, E. H., Borsky, M., Meier, M., Johannsdottir, K., and Gudnason, J. (2017). Monitoring cognitive workload using vocal tract and voice source features. Period. Polytech. Electr. Eng. Comput. Sci. 61, 297–304. doi: 10.3311/PPee.10414

Makowski, D., and Lüdecke, D. (2019). The report package for R: Ensuring the use of best practices for results reporting. CRAN. Available at: https://github.com/easystats/report (Accessed February 12, 2021).

Meshkati, N. (1988). Toward development of a cohesive model of workload. Adv. Psychol. 52, 305–314.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. 63, 81–97.

Montani, F., Vandenberghe, C., Khedhaouria, A., and Courcy, F. (2020). Examining the inverted U-shaped relationship between workload and innovative work behavior: the role of work engagement and mindfulness. Hum. Relat. 73, 59–93. doi: 10.1177/0018726718819055

Muratet, M., Torguet, P., and Jessel, J.-P. (2009). “Learning programming with an RTS based serious game” in Serious games on the move. eds. O. Petrovic and A. Brand (Vienna: Springer), 181–192.

Nebel, S., and Ninaus, M. (2019). “New perspectives on game-based assessment with process data and physiological signals” in Game-based assessment revisited. eds. D. Ifenthaler and Y. Kim (Cham: Springer), 141–161.

Niederhauser, D. S., Reynolds, R. E., Salmen, D. J., and Skolmoski, P. (2000). The influence of cognitive load on learning from hypertext. J. Educ. Comput. Res. 23, 237–255. doi: 10.2190/81BG-RPDJ-9FA0-Q7PA

Ninaus, M., Witte, M., Kober, S. E., Friedrich, E. V., Kurzmann, J., Hartsuiker, E., et al. (2013). “Neurofeedback and serious games” in Psychology, pedagogy, and assessment in serious games. Vol. i. eds. E. T. M. Connolly, T. Boyle, G. Hainey, P. Baxter, and P. Moreno-ger (USA: IGI Global), 82–110.

O'Donnell, R., and Eggemeier, F. (1986). "Workload assessment methodology" in Handbook of perception and human performance. Volume 2. Cognitive processes and performance. eds. K. R. Boff, L. Kaufman, and J. P. Thomas (John Wiley and Sons, Inc.).

Orru, G., and Longo, L. (2018). "The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and Germane loads: a review" in Human mental workload: Models and applications. H-WORKLOAD 2018. Communications in computer and information science. eds. L. Longo and M. Leva (Cham: Springer), 23–48.

Paas, F. G., and Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. Educ. Psychol. Rev. 6, 351–371.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., and Perry, R. P. (2010). Boredom in achievement settings: exploring control–value antecedents and performance outcomes of a neglected emotion. J. Educ. Psychol. 102, 531–549. doi: 10.1037/a0019243

Portrat, S. (2008). Working memory and executive functions: The time-based resource-Sharin account. Dijon: Université de Bourgogne.

Promotion Software GmbH. (1999). World of Emergency. Available at: from Promotion Software GmbH website: https://www.world-of-emergency.com/?lang=en (Accessed August 26, 2019).

R Core Team (2020). R: A Language and Environment for Statistical Computing. Available at: https://www.R-project.org/

Reid, G. B., and Nygren, T. E. (1988). The subjective workload assessment technique: a scaling procedure for measuring mental workload. Adv. Psychol. 52, 185–218.

Richards, K. C., Enderlin, C. A., Beck, C., McSweeney, J. C., Jones, T. C., and Roberson, P. K. (2007). Tailored biobehavioral interventions: a literature review and synthesis. Res. Theory Nurs. Pract. 21, 271–285. doi: 10.1891/088971807782428029

Ruiz, N., Liu, G., Yin, B., Farrow, D., and Chen, F. (2010). "Teaching athletes cognitive skills: detecting cognitive load in speech input" in Proceedings of HCI 2010 24; September 6–10, 2010; 484–488.

Saddler, B., Moran, S., Graham, S., and Harris, K. R. (2004). Preventing writing difficulties: the effects of planning strategy instruction on the writing performance of struggling writers. Exceptionality 12, 3–17. doi: 10.1207/s15327035ex1201_2

Salomon, G. (1984). Television is "easy" and print is "tough": the differential investment of mental effort in learning as a function of perceptions and attributions. J. Educ. Psychol. 76, 647–658.

Scerbo, M. W. (1996). "Theoretical perspectives on adaptive automation" in Automation and human performance: Theory and applications. eds. R. Parasuraman and M. Mouloua (CRC Press), 37–64.

Sheridan, T. B., and Simpson, R. (1979). Toward the definition and measurement of the mental workload of transport pilots. Available at: https://dspace.mit.edu/handle/1721.1/67913 (Accessed March 12, 2021).

Shute, V. J., and Kim, Y. J. (2014). “Formative and stealth assessment” in Handbook of research on educational communications and technology. eds. J. Spector, M. Merrill, J. Elen, and M. Bishop (New York, NY: Springer), 311–321.

Simons, A., Wohlgenannt, I., Weinmann, M., and Fleischer, S. (2020). Good gamers, good managers? A proof-of-concept study with Sid Meier’s civilization. Rev. Manag. Sci. 1–34. doi: 10.1007/s11846-020-00378-0

Smith-Jackson, T. L., and Klein, K. W. (2009). Open-plan offices: task performance and mental workload. J. Environ. Psychol. 29, 279–289. doi: 10.1016/j.jenvp.2008.09.002

Spronck, P., Ponsen, M., Sprinkhuizen-Kuyper, I., and Postma, E. (2006). Adaptive game AI with dynamic scripting. Mach. Learn. 63, 217–248. doi: 10.1007/s10994-006-6205-6

Susi, T., Johannesson, M., and Backlund, P. (2007). Serious games: An overview. Available at: http://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-1279 (Accessed September 7, 2018).

Sweller, J., Van Merrienboer, J. J., and Paas, F. G. (1998). Cognitive architecture and instructional design. Educ. Psychol. Rev. 10, 251–296.

Temple, J. G., Dember, W. N., Warm, J. S., Jones, K. S., and LaGrange, C. M. (1997). “The effects of caffeine on performance and stress in an abbreviated vigilance task” in Proceedings of the Human Factors and Ergonomics Society Annual Meeting; October 1997; 1293–1297.

Van Rossum, G., and Drake, F. L. (2009). PYTHON 2.6 Reference Manual.

Veltman, J., and Jansen, C. (2005). The role of operator state assessment in adaptive automation. Available at: https://apps.dtic.mil/sti/citations/ADA455055 (Accessed February 12, 2021).

Vygotsky, L. S. (1980). Mind in society: The development of higher psychological processes. eds. M. Cole, V. John-Steiner, S. Scribner, and E. Souberman (Cambridge, London: Harvard University Press).

Walter, C., Rosenstiel, W., Bogdan, M., Gerjets, P., and Spüler, M. (2017). Online EEG-based workload adaptation of an arithmetic learning environment. Front. Hum. Neurosci. 11:286. doi: 10.3389/fnhum.2017.00286

Wang, Y.-R., and Gibson, G. E. (2010). A study of preproject planning and project success using ANNs and regression models. Autom. Constr. 19, 341–346. doi: 10.1016/j.autcon.2009.12.007

Watters, P. A., Martin, F., and Schreter, Z. (1997). Caffeine and cognitive performance: the nonlinear Yerkes–Dodson law. Hum. Psychopharmacol. Clin. Exp. 12, 249–257.

Welford, A. (1978). Mental work-load as a function of demand, capacity, strategy and skill. Ergonomics 21, 151–167.

Yerkes, R. M., and Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. J. Comp. Neurol. Psychol. 18, 459–482.

Yuksel, B. F., Oleson, K. B., Harrison, L., Peck, E. M., Afergan, D., Chang, R., et al. (2016). “Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state” in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; May 2016; 5372–5384.

Yurko, Y. Y., Scerbo, M. W., Prabhu, A. S., Acker, C. E., and Stefanidis, D. (2010). Higher mental workload is associated with poorer laparoscopic performance as measured by the NASA-TLX tool. Simul. Healthc. 5, 267–271. doi: 10.1097/SIH.0b013e3181e3f329

Zhou, T., Cha, J. S., Gonzalez, G., Wachs, J. P., Sundaram, C. P., and Yu, D. (2020). Multimodal physiological signals for workload prediction in robot-assisted surgery. ACM Trans. Hum. Robot Interact. 9, 1–26. doi: 10.1145/3368589

Zohaib, M. (2018). Dynamic difficulty adjustment (DDA) in computer games: a review. Adv. Hum. Comput. Interact. 2018, 1–12. doi: 10.1155/2018/5681652

Zook, A. E., and Riedl, M. O. (2012). “A temporal data-driven player model for dynamic difficulty adjustment” in Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. Available at: https://ojs.aaai.org/index.php/AIIDE/article/view/12504; October 8-12, 2012; Campus of Stanford University, Palo Alto, California, USA.

3 Study 2

Please cite as:

Sevcenko, N., Shopp, B., Dresler, T., Ehlis A.-C., Ninaus, M., Moeller, K., & Gerjets, P. (2021). Neural correlates of cognitive load while playing an emergency simulation game: a functional near-infrared spectroscopy (fNIRS) study. *IEEE Transactions on Games.. DOI: 10.1109/TG.2022.3142954*

Neural correlates of cognitive load while playing an emergency simulation game: a functional near-infrared spectroscopy (fNIRS) study

Authors:

Natalia Sevcenko (2,3), Betti Schopp (7), Thomas Dresler (4,7), Ann-Christine Ehlis (4,7), Manuel Ninaus (1,4,6), Korbinian Moeller (1,5), Peter Gerjets (1,4)

Affiliations:

1 - Leibniz Institut für Wissensmedien, Germany

2 - Department of Psychology, University of Tuebingen, Germany

3 - Daimler Trucks AG, Germany

4 - EAD Graduate School & Research Network, University of Tuebingen, Germany

5 - Centre for Mathematical Cognition, School of Science, Loughborough University, Loughborough, UK

6 - Department of Psychology, University of Innsbruck, Austria

7 - Department of Psychiatry & Psychotherapy, University Hospital Tuebingen, Tuebingen, Germany

Keywords: cognitive load, fNIRS, adaptivity, emergency, simulation, time-critical, Barrouillet, PFC, DLPFC, IFG

Abstract

Determining cognitive load seems crucial for the development of training environments for life- and time-critical emergencies. Empirical evidence suggests that functional near-infrared spectroscopy (fNIRS) provides reliable results for determining cognitive load based on multiple repetitions of relatively short cognitive tasks and that fNIRS measures of the prefrontal cortex are sensitive to changes in cognitive load. At the same time, it remains unclear how to use this technique for assessing cognitive load during a prolonged single-trial activity including heterogeneous tasks. In this study, we propose a novel approach to measure cognitive load during a computer-based emergency simulation game. To this end, we defined specific time periods, determined based on simulation log-data interpreted in light of Barrouillets' time-based resource-sharing model. Further, to validate our approach, we compared cortical oxygenation in prefrontal cortex areas, i.e. dorsolateral prefrontal cortex (DLPFC) and left inferior frontal gyrus (IFG), during predefined time periods. We found significant associations between cognitive load and oxygenation within the DLPFC depending on the chosen time slot, whereas no such dependencies were found for the IFG. Although requiring further investigation in terms of reliability and generalizability, the presented approach seems promising evidence that fNIRS might be suitable for the assessment of cognitive load beyond classical experimental set-ups.

1 Introduction

Daily life is becoming increasingly automated: Autopilots, speed and lane assistants, automated manufacturing, etc. help us to deal with well-controlled and predictable settings. In contrast, however, time-critical situations, such as driving a car in heavy traffic or performing a complex surgery still require a qualified human operator, preferably trained for such challenges. As it is not trivial to set up an analog training environment for life- and time-critical emergencies, computer-aided simulations, and digital training scenarios can be used advantageously in this area (Johnson, Rodrigues, Gubbala, & Weibel, 2018; Kincaid, Donovan, & Pettitt, 2003; Liu et al., 2020). Along with the potential to simulate dangerous time-critical situations, digital training scenarios also allow for the collection of individual data, which can be used to model cognitive or emotional states of the user (Nebel & Ninaus, 2019) including cognitive load.

Cognitive load is considered to reflect “how hard the brain is working to meet task demands” (Ayaz et al., 2012). According to the most influential theoretical view, it relates to user performance in a way shaped as “inverted-U” (Yerkes & Dodson, 1908), also closely related to the concept of “flow” (Csikszentmihalyi, 1987; Kiili, Lindstedt, & Ninaus, 2018), meaning that performance usually decreases in cognitive over- and underload conditions (e.g., Anderson, 1994; Montani, Vandenberghe, Khedhaouria, & Courcy, 2020; Watters, Martin, & Schreter, 1997; Yerkes & Dodson, 1908). Consequently, human-machine interaction such as any computer-based training might be optimized using a monitoring system capable of detecting variations in cognitive load (Orru & Longo, 2018), which seems feasible in real-world training environments (Appel et al., 2019; Gerjets, Walter, Rosenstiel, Bogdan, & Zander, 2014).

The literature describes four main categories of techniques for assessing cognitive load (Brünken, Seufert, & Paas, 2010; Eggemeier, Wilson, Kramer, & Damos, 1991; Johannsen, 1979; Scerbo, 1996), although it needs to be considered that each approach is associated with specific strengths and limitations. First, subjective measures are collected using self-reported questionnaires (e.g., NASA-TLX: Hart & Staveland, 1988; SWAT: Reid & Nygren, 1988). They are inexpensive and reliable (O’Donnell & Eggemeier, 1986), but are not capable of tracking subtle variations in cognitive load online over time. Second, performance-based approaches evaluate fluctuations in human performance, i.e. task outcome measures, and relate these to changes in psychological constructs such as cognitive load. This category of measurement

techniques appears intuitively to be the most obvious and direct. On the other hand, it is often impossible to obtain the necessary data until the actual task has already been completed. Moreover, it's hard to determine whether observed variations in performance have occurred due to changes in cognitive load or some other factors (Brünken et al., 2010). Third, behavioral measurements evaluate differences in interaction behavior (e.g., speech patterns or mouse usage) with the training system during usage (Berthold & Jameson, 1999; Ikehara & Crosby, 2005; Lim, Ayesh, & Stacey, 2014; Magnusdottir, Borsky, Meier, Johannsdottir, & Gudnason, 2017; Ruiz, Liu, Yin, Farrow, & Chen, 2010; Yap, Epps, Ambikairajah, & Choi, 2011). These measures are usually unobtrusive and inexpensive, do not require additional equipment, and potentially allow for continuous monitoring of cognitive states. However, behavioral patterns can also be influenced by factors unrelated to cognitive load (e.g., emotions or stress). Finally, physiological approaches are based on the evidence that changes in cognitive states are accompanied by physiological changes (Ahmad, Malik, Kamel, & Reza, 2016; Buchwald et al., 2019; Fowler, Nesbitt, & Canossa, 2019; Johannsen, 1979; Liang, Liang, Qu, & Yang, 2018; McDuff, Gontarek, & Picard, 2014). Their advantage is that they allow continuous recording of data and thus might be used for online adaptation of training environments. At the same time, monitoring of physiological signals often requires expensive and complicated equipment and sophisticated filtering and analysis procedures. Moreover, different types of physiological measures differ considerably in their obtrusiveness and practicability in real-life settings. While sensors for heart rate variability (HRV), electrodermal activity (EDA), or eye-tracking can be used in a comparably discreet way, electroencephalography (EEG) or functional magnetic resonance imaging (fMRI) are less practical or even impracticable in real-life situations because of their signal sensitivity and immobility (for an overview see Ninaus et al., 2013).

According to Brunken, Plass, and Leutner (2003) we can further categorize physiological measurements into indirect and direct methods, based on the type of relation of cognitive load and observed variables. From this perspective, measurements such as pupil dilation or HRV are only indirectly related to cognitive load, as they can be co-influenced by other factors such as emotional response or stress, whereas imaging techniques such as fMRI, EEG, positron emission tomography (PET), and functional near-infrared spectroscopy (fNIRS) can be considered direct methods, as they usually

assess hemodynamic cortical activation during task execution[2]. This approach seems to be very promising, as it can be applied in real-time independently from any internal data of the training software (e.g., mouse movements or training scores). On the other hand, as mentioned above, the major limitation of most neurophysiological and -imaging techniques is that they are expensive, complex, and immobile, which significantly limits their use in ecologically valid real-life studies. In this respect, fNIRS, which is described in the following section, offers some advantages over other neuroimaging techniques and appears promising in this field.

1.1 Functional near-infrared spectroscopy

Near-infrared spectroscopy was first presented in the fundamental work of Jobsis (1977). It represents a non-invasive neuroimaging method for measuring cerebral oxygenation/hemodynamics, which relies on the mechanism of neurovascular coupling and optical spectroscopy (for review see: Fallgatter, Ehlis, Wagener, Michel, & Herrmann, 2004; Herold, Wiegel, Scholkmann, & Müller, 2018; Strangman, Boas, & Sutton, 2002).

Hereby, near-infrared light with strictly defined wavelengths in the range between 600 and 1000 nm is emitted through the scalp of the participant. It penetrates up to approx. 2 cm (Strangman, Li, & Zhang, 2013) deep into the tissue (depending on the distance between light source and detector) and thus reaches outer cortical gray matter, where the emitted light is partially absorbed by oxygenated (HbO2) and deoxygenated (HbR) hemoglobin molecules of blood cells that differ significantly in their absorption spectra. The residual reflection of the respective wavelength is received by a detector, with a usual inter-optode distance of about 3 cm (Herold et al., 2018; Strangman et al., 2013). Based on the difference between emitted and detected light (see Cope et al., 1988), the relative concentration of HbO2 and HbR in the brain tissue underlying the mean distance between light transmitter and its detector is calculated, allowing inferences about local changes in blood flow.

[2] To be perfectly precise, all these methods can also be considered as indirect, because they deduce cognitive activity through blood flow or electrical activation. However, as there is no more direct method of capturing it today, we retain this terminology below.

Neuronal activation in a particular brain area increases its metabolic needs and leads to an increase in local cerebral blood flow (Hoge et al., 1999; Kim, Rostrup, Larsson, Ogawa, & Paulson, 1999). This response is called “functional hyperemia” and is mediated by mechanisms of neurovascular coupling (Nippert, Biesecker, & Newman, 2018), resulting in an increase in HbO2 with a simultaneous decrease in HbR in the respective area. Based on these observations, a temporal change in detected HbO2 and HbR hemoglobin concentration allows conclusions about changes in local brain activation. Therefore, depending on the measurement area, it seems possible to draw online conclusions about which cognitive processes take place in the brain at a certain point of time. When attempting to evaluate changes in cognitive activation related to, for instance, cognitive load, areas of the prefrontal cortex (PFC) are typically regarded (Herold et al., 2018).

1.1.1 Prefrontal cortex. The PFC is part of the neocortex and is located in the anterior part of the frontal lobe. The percentage of the PFC from the total volume of the cerebral cortex varies considerably between animal species, ranging from 3.5% in a cat to 29% in humans, suggesting that this area has developed rather late in evolutionary history (Fuster, 2015) and thus might underlie higher cognitive functions. Indeed, the PFC is connected with a “vast array of other cerebral structures” (Fuster, 2001, p. 319) such as the brainstem, thalamus, basal ganglia, limbic system, and a number of neocortical regions that serve various aspects of sensory and motor functions. Documented cases of PFC lesions often report a “personality change” that manifests itself either in a general reduction in activity or the disinhibition of behavior without affecting the patients’ intelligence, speech, or memory (see Kammer & Karnath, 2006).

Taken together, the evidence suggests that the PFC plays a modulatory role in the path between stimulus and response. The integrative theory of prefrontal cortex function by Miller and Cohen (2001) suggests that the PFC generates special “bias” signal patterns that alter further stimulus processing either by blocking or amplifying desired processing paths and thus adapting reflectively produced behaviors to the current context. For example, if on the way to work (which is an automated behavior) we see someone lying in the street, we have the choice of interrupting the automated behavior or not, which would depend on our experience and the context, including whether we have enough time for this or whether we see that the person is already being

helped, etc. This decision-making process - involving action often summarized as “cognitive control” or “executive functions” - seems associated with PFC involvement along with at least partial influences of working memory (Thier, 2006). As such, the PFC seems an obvious location to place fNIRS optodes when measuring cognitive load. This placement also has an obvious technical advantage. The scalp in the forehead area and above the forehead is usually less hairy, which makes the measurement more reliable (Murkin & Arango, 2009).

1.1.2 Strengths and limitations. When comparing the temporal and spatial resolution of neuroimaging methods, fNIRS provides solid results in both domains (Parasuraman & Rizzo, 2006, see Figure 1), even when comparing short segments of data (Strangman, Goldstein, Rauch, & Stein, 2006).

The great strength of the method is that, depending on the fNIRS device used, experimental tasks can be performed while sitting, standing, or even in motion, which makes the obtained results ecologically far more valid than results from experiments in which participants have to lie down, as during fMRI measurements. Furthermore, compared to EEG, this method is less susceptible to motion artifacts, electro-oculographic and facial electromyographic activity, as well as electrical environmental noise - which might be particularly problematic in neuronal measurements of human-machine interactions (see Derosière, Mandrick, Dray, Ward, & Perrey, 2013). As such, fNIRS has been successfully used for investigating cognitive and emotional processes during gaming (e.g., Kober, Wood, Kiili, Moeller, & Ninaus, 2020; Witte, Ninaus, Kober, Neuper, & Wood, 2015).

However, fNIRS is not completely free of artifacts and the recorded data must be pre-processed for analysis. The measurement method is sensitive to changing light conditions, which must be taken into account when planning a study. Moreover, the hair of the participant must be carefully pushed to the side at each fNIRS measurement optode, as they can strongly influence the quality of the signal (McIntosh, Shahani, Boulton, & McCulloch, 2010; Pringle, Roberts, Kohl, & Lekeux, 1999).

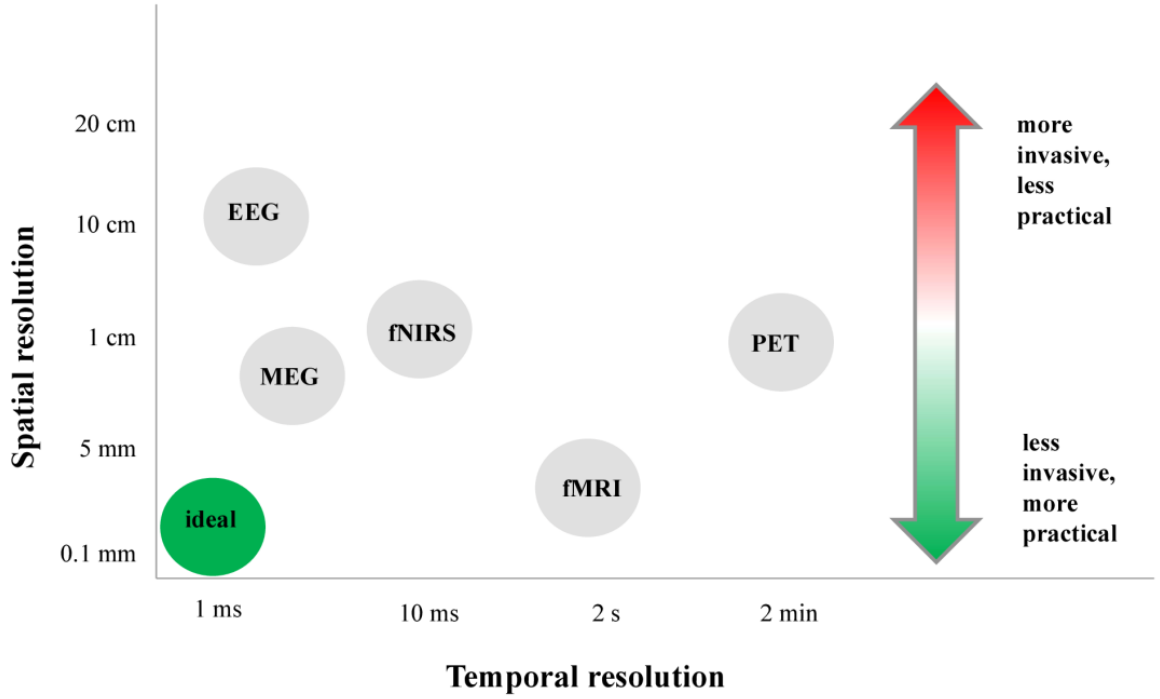


Figure 1. Resolution space of brain imaging techniques (cf. Parasuraman & Rizzo, 2006, p. 7).

A solid body of evidence supports that fNIRS measures of PFC are sensitive to changes in cognitive load. Several experiments documented an increase of PFC activation with increasing difficulty of *n-back* (Ayaz, Izzetoglu, Bunce, Heiman-Patterson, & Onaral, 2007; Fishburn, Norr, Medvedev, & Vaidya, 2014; Herff et al., 2014; Herrmann et al., 2007; Li, Gong, Gan, & Luo, 2005; Smith & Jonides, 1997) and *Stroop* tasks (Ehlis, Herrmann, Wagener, & Fallgatter, 2005; Xu et al., 2017).

These fundamental studies substantiated the usage of fNIRS for the measurement of cortical activity in laboratory settings. However, in contrast to a well-structured laboratory experiment, typical real-life challenges contain a multitude of heterogeneous events (e.g., visual recognition, emotion, and memorization) that occur simultaneously and concurrently, making it difficult to impossible to assign hemodynamic cortical activation to specific events. That raises the question of whether and how technology can be used for more complex experiments with heterogeneous and more ecologically valid tasks, as they typically happen in everyday life. First such attempts (see some examples below) have already been made.

For example, Ayaz et al. (2012) examined the hemodynamics (HbDiff computed as the difference between HbO2 and HbR) during an air traffic management task in which participants had to handle 6, 12, or 18 air forces, respectively, for each of two specified communication types (either voice or data). The researchers found increased activity within the left PFC (i.e., in inferior frontal gyrus and medial PFC) corresponding to defined levels of difficulty during short task sequences as compared to pre-task resting periods.

Bruno et al. (2018) also observed increased activity of bilateral dorsolateral PFC with increasing task difficulty using fNIRS (HbO2) for a simulated driving task. Therein, task difficulty was increased for so-called “incongruent” tasks, where turning the steering wheel to the right resulted in an opposite movement of the car (to the left).

In another study by Unni, Ihme, Jipp, and Rieger (2017), an elegant design was used to integrate a standard *n-back* routine (*0-* up to *4-back*) into a driving task using a virtual reality driving simulator. Task difficulty was adjusted by asking participants to adjust their speed according to the speed signal that appeared before n steps (i.e. in the *1-back* condition participants had to adapt their speed to the last speed sign, while in the *4-back* condition they had to maintain the speed prescribed by the 4th last sign). The observed changes in HbR in bilateral inferior frontal areas and bilateral temporo-occipital areas were found to reflect cognitive load induced by the adapted *n-back* task.

These examples indicate that fNIRS can be a useful method for conducting ecologically valid realistic experiments. However, although these studies investigated realistic tasks, they mainly used methods that are only applicable in laboratory settings. In all of these studies, hemodynamic cortical activation was measured during short periods of high cognitive load, and data were aggregated over a large number of equivalent repetitions. This method is very common. Herold et al. (2018) reported in their methodological review of 35 studies that fNIRS research typically either used block or event-based design, which is hardly feasible in real-life situations, or provided simple comparisons between baseline and post-treatment measurements.

This methodology provides reliable results but does not answer the question of whether fNIRS technology can also be used to measure cognitive load online during long-term heterogeneous tasks that are not repeated several times, as happens often in real-life. And because such tasks usually involve different mental activities, which can

overlap in time and be executed in different sequences depending on participants' strategy, it is hard to tell exactly what specific timeframes should be used when evaluating cognitive load. Another common approach as described by (Gerjets et al., 2014) consists of using machine learning methods based on behavioral and (neuro-) physiological data, which seems to be quite feasible for real-live adaptation, leaving open the question of how to generalize the model, because such a data-driven approach cannot be easily generalized to different subjects and situations.

Against this background, one might wonder whether an appropriate theoretical approach might help to resolve these issues considering the specific situation in which cognitive load should be measured. Because we focus on the measurement of cognitive load during the simulation of time-critical emergencies, a well-evaluated (Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007; Camos, Portrat, Vergauwe, & Barrouillet, 2007; Lépine, Bernardin, & Barrouillet, 2005; Liefooghe, Barrouillet, Vandierendonck, & Camos, 2008; Portrat, 2008) time-based resource-sharing (TBRS) model developed by Barrouillet et al. Barrouillet, Bernardin, and Camos (2004) might provide a suitable theoretical framework.

1.2 Time-based resource-sharing model of cognitive load

Barrouillet et al. (2004) emphasized that, in addition to task complexity, cognitive load is strongly dependent on the time available for the task at hand, which is particularly relevant for time-critical situations. According to the TBRS Model cognitive load depends on the proportion of time, "during which attention is captured in such a way that the storage of information is disturbed", which is not trivial to determine. Nevertheless, in their recent paper, Sevcenko, Ninaus, Wortha, Moeller, and Gerjets (2021) have shown that this model can be successfully applied to predicting cognitive load in time-critical serious games. Thereby they used the TBRS model to define a behavioral metric based on the ratio of specific time intervals. Relying on log data, they found that the game flow can be divided into so-called *action blocks*, i.e. time-segments of dense actions (*burst*) followed by some waiting time (*idle*). When assuming that during the *burst* periods participants are operating at their cognitive limit - which appears plausible under time-critical emergency situations- their cognitive load should be comparably high (i.e. at their personal maximum), and thus cognitive load of an *action block* can be estimated as the

relation of the duration of the *burst* phase to the total duration of the inspected block (*burst* + *idle*). This way, behavioral metric *temporal action density decay (TADD)* was proposed and validated in 47 participants. Interestingly, it was discovered that *TADD* reflecting this relation within the first action block (*initial TADD*) was a more valid measure of cognitive load as compared with averaged TADD sequence calculated over a whole level. This means that a simple behavioral metric at the beginning of a level significantly predicted the success of the whole level. The fact that *initial TADD* was based on data collected very early on during the game process, makes it potentially useful for real-time adaptation systems.

This apparently provides us with a measurement foundation that is i) theory-driven and can thus be generalized to any time-critical resource management situation; ii) is based on relatively short time intervals (*burst* & *idle*); iii) can be calculated from the initial phase of training and is thus suitable for real-time adaptation. However, which time slots should be used when applying this analytical approach using fNIRS methodology? *Idle* time intervals can become too short (or even equal to zero), for example, if the player fails to act fast enough and the first occupied emergency personnel finishes his task before the player has completed all ongoing task assignments. This makes *idle* time intervals unqualified for fNIRS analysis. On the other hand, *burst* periods seem sufficiently long and comparable, because during the first *burst* interval all participants are completing nearly the same tasks. However, also because all participants are supposed to operate at their limit under time-critical conditions, we can expect similar neuronal activity during this phase. In this article, we aimed at exploring these options and evaluate whether the direct observation of cortical hemodynamics during the initial *burst* phase and the idle phase directly afterward can provide additional insights into the nature and assessment of cognitive load.

1.3 Present study

In this study, we pursue two main questions. We are interested in whether fNIRS technology is feasible for cognitive load detection in realistic time-critical emergency situations realized with a game; and if so, which time intervals should be used, considering the TBRS Model of Barrouillet et al. (2004) and previous results. In

particular, we evaluate cortical hemodynamics in the PFC region for specified time slots addressing whether the corresponding hemodynamic cortical activation is related (i) to the task difficulty, (ii) to the achieved performance, (iii) and to the subjective perception of cognitive load. To allow for variance in cognitive load, we use a computer-based time-critical emergency simulation game which requires management of time-critical situations and realizes different levels of difficulty, as validated using NASA-TLX questionnaire. Two scenarios of the simulation represent two different emergency situations, in which participants have to perform the same predefined number of actions by managing a predefined number of emergency personnel and equipment. These prerequisites are constant for all participants, although the exact chronological order of task execution and, thus, the resulting effectiveness depends on the individual strategy used.

2 Methods

The study was carried out as part of a larger project that included several other physiological measures such as cardiac measurements, galvanic skin response, and eye-tracking. The aim of the study was to investigate whether the fNIRS methodology would be feasible for detecting cognitive load in realistic environments.

2.1 Participants

In this study, we present data of 27 volunteers (18 females, 9 males) aged between 20 and 49 years (M = 25.9; SD = 7.2). Data of further 20 participants were excluded from the present analysis due to the poor quality of the fNIRS recording. All participants were right-handed, spoke fluent German, were recruited via an online database, and were compensated monetarily for their time expenditure. They reported no neurological, mental or cardiovascular disorders, and did not take any psychotropic medication. The local ethics committee approved the study and written informed consent was obtained from all participants prior to the experiment.

2.2 Task

All participants played an adapted version of a game-based simulation of different emergency situations (Emergency: Promotion Software GmbH, 1999), in which they had to coordinate emergency personnel, such as paramedics, emergency doctors, and firefighters, as well as auxiliary items such as ambulances, fire trucks, and fire truck ladders to rescue victims and extinguish fires.

After getting familiar with the games' simulation routine by playing a tutorial and a training scenario, participants were confronted with two experimental scenarios: *Fire* and *Train Crash*. Each scenario included three levels of difficulty: easy, *medium*, and *hard*. These were defined by varying the number of tasks to be performed and the number of personnel to be coordinated within a given period of time. Because increased task density would require not only more actions but also better planning, coordination, and prioritization, we expected this to generate a concomitant increase in cognitive load. The time pressure was imposed by setting time limits for the levels as well as time bars that

showed how fast a victim would die if not helped. For a detailed summary of all simulation parameters, scenario descriptions, and validation of difficulty levels see Sevcenko et al. (2021).

2.3 Experimental setup and design

The experiment was implemented in a quiet room under constant light conditions. The Emergency game simulation was presented on a 16''notebook driven at a screen resolution rate of 1920 x 1080 using a conventional computer mouse as the only interaction tool. Data for cortical hemodynamics were acquired on an additional notebook as can be seen in Figure 2, using a portable NIRSPORT-2 device (NIRX Medical Technologies). The simulation started with an introductory training sequence, which was followed by two experimental scenarios: *Fire* and *Train Crash*. At the end of each of the three levels per scenario, NASA-TLX scores were collected. Each participant executed the defined sequence of levels only once, which lasted about one hour including the training phase.

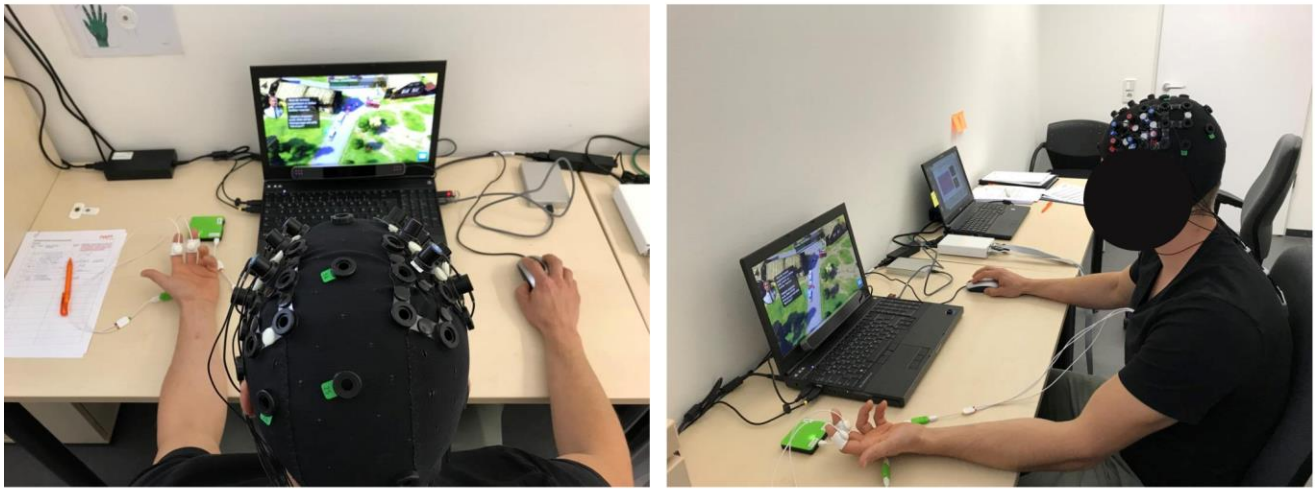


Figure 2. Experimental setup.

2.3 Measured variables

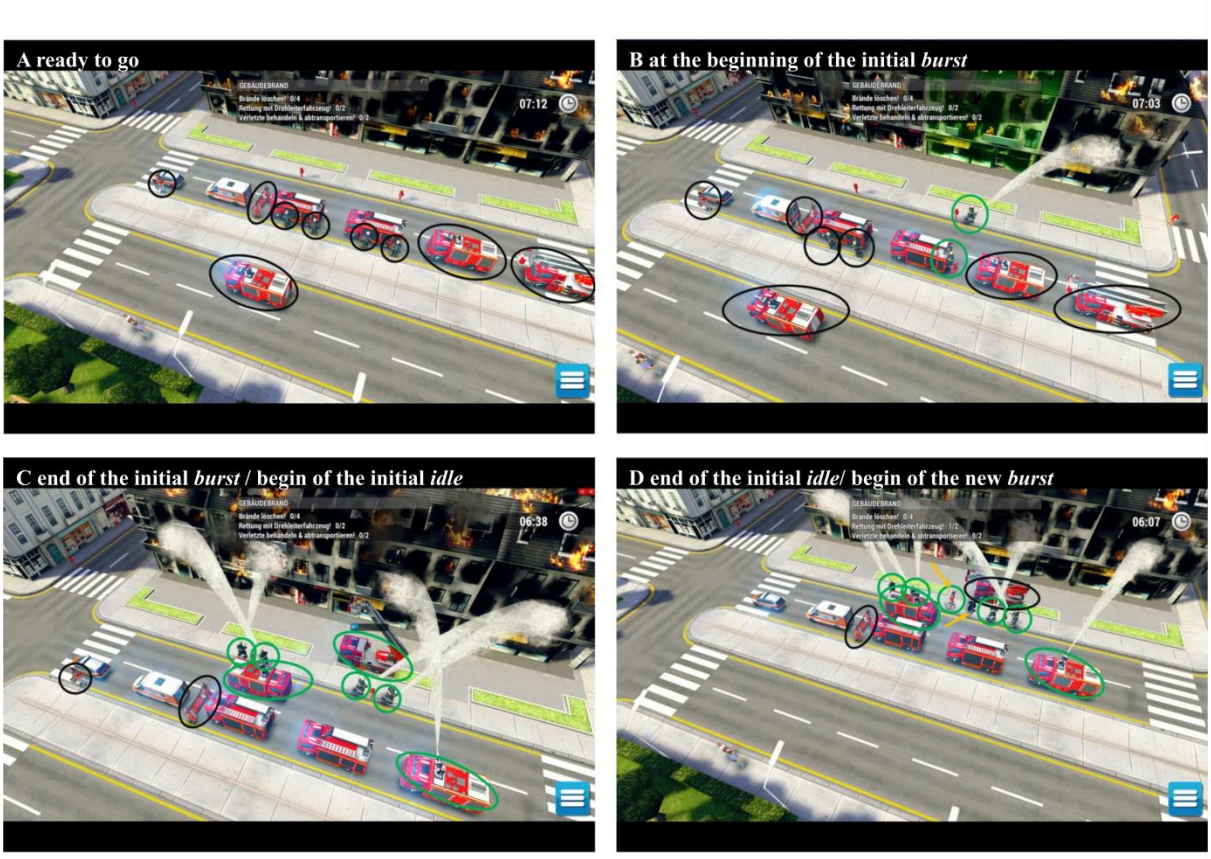
For estimating cognitive load we used a continuous multichannel recording of cortical hemodynamics during time slots that were derived from participants' behavior as described below. These measurements were subsequently associated with level

difficulty, performance data, and subjective estimation of cognitive load, collected by a subset of items taken from the NASA-TLX questionnaire (i.e., *mental demand*, *time demand*, and *effort*). Additional data such as age and sex were acquired prior to the experiment using a self-report survey.

2.3.1 Level difficulty and performance data. For each level a *real difficulty* score was defined as the percentage of participants, who failed to complete the level, this means, could not extinguish all the fires and transport all injured persons to the hospital within the defined time limit. Individuals' performance per level was represented by the binary indicator of whether the level was completed successfully or not.

2.3.2 NASA-TLX. Subjective estimation of cognitive load was acquired using subscales of the multidimensional NASA-TLX (Hart & Staveland, 1988) questionnaire, which consists of six items/dimensions rated on a 21-level scale (0 to 100 points with steps of 5). These dimensions correspond to various theories distinguishing between physical, mental, and emotional facets of operators' load (Hart, 2006). In the current study, we used the subscales addressing the mental facet, i.e. *mental demand*, *temporal demand*, and *effort*, which represents a common procedure when investigating specific facets of workload (Haerle et al., 2013; Temple, Dember, Warm, Jones, & LaGrange, 1997).

2.3.3 Time periods. Hemodynamic cortical data were analyzed based on *burst* and *idle* time periods related to the *initial TADD* metric proposed by Sevcenko et al. (2021), which were designed to predict cognitive load during time-critical situations where tasks have to be prioritized and resources managed. According to this approach, the time series of participants' actions during each level was divided into so-called *action*



blocks, consisting of active (*burst*) and waiting (*idle*) periods. During the *burst* period, participants manage their emergency personnel, and after the last available personnel are assigned a task, the *idle* interval occurs and lasts until the first personnel is available again. Figure 3 illustrates four exemplary turning points in the run-up to the emergency situation, which determine the initial *burst* and *idle* phases.

Figure 3. Behavioral phases during a game, explained using the *Fire* scenario as an example. Inactive emergency forces are marked with black circles, active - with green ones.
A: the game level starts, all emergency personnel are ready to go and still inactive.
B: the initial *burst* phase begins with the first task, assigned to the emergency force by the player, and lasts until all available personnel are engaged actively. In this example, the emergency doctor and paramedics are not applicable because there are no injured persons waiting for the treatment.
C: the initial *burst* phase ends as soon as the last applicable emergency force is engaged, emergency doctor and paramedics are still not applicable. This time-point marks also the beginning of the initial *idle* phase, in which the player has to wait until some emergency personnel gets free or new tasks appear. This example shows an initial *idle* phase as all

available emergency personnel are already active and the player has to wait until the person is rescued from the burning building and the next task will appear.
D: the initial *idle* phase ends as soon as some emergency personal becomes available. In this example, the initial *idle* phase ends as soon as the rescued person appears laying down on the street. At this moment, an emergency doctor becomes available and can be assigned to treat the patient; at the same time, also the ladder truck becomes inactive and must be instructed to rescue the next person.

In this study, we analyzed hemodynamic data obtained during four subsequent time slots. The first time slot initial *burst* starts with the first user action and its duration corresponds to the duration of the initial *burst* phase, which varied between participants. The three subsequent time slots start with the end of the corresponding initial *burst* and last 20 seconds each (t0 – t20: starting directly after the initial *burst* phase; t20 – t40: starting 20 seconds after the end of the initial *burst*; t40 – t60: starting 40 seconds after the end of the *burst*). This configuration is shown in Figure 4.

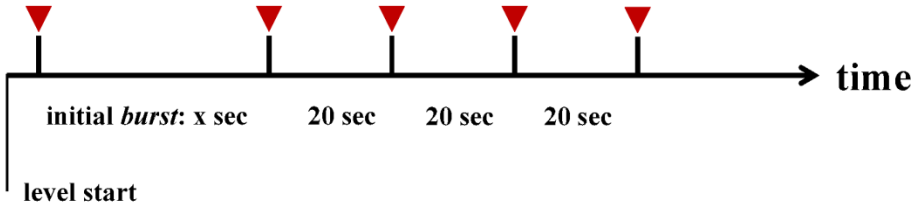


Figure 4. Graphical representation of the investigated time periods.

Because we assumed that all participants would be working at their cognitive limit during the *burst* phase, we expected no effects of task difficulty, performance and subjective cognitive load on neuronal activation during this time. After the *burst* phase is over, participants who experienced high cognitive load were supposed to have only a very short *idle* phase if any. Consequently, during the 20 seconds following the initial *burst*, participants who experience low cognitive load should still be in the initial *idle* phase, whereas for participants with high cognitive load, the next *burst* should already start. Assuming that hemodynamic cortical activation during the initial *burst* phase differs from the activation during the initial *idle* phase, we expected to find effects on neuronal activity during this time. As time progresses, we expected more heterogeneity in the data, this means, depending on their strategies more and more participants would start with a new

burst phase while others were still or again in an *idle* phase. Therefore, we expected smaller or even no significant effects in the later stages of the level.

2.4 FNIRS Imaging Procedure

Participants' PFC oxygenation was recorded using a portable NIRSPORT-2 device (NIRX Medical Technologies), which works with two wavelengths (750mn & 859 nm) and provides time series of relative HbO2 and HbR concentration changes (Cope et al., 1988). We used 8 light emitters and 8 detectors, resulting in 20 channels, which were placed into electrode caps CUCMS-56/58 (EASYCAP) and adjusted to the Cz and Fpz positions according to the 10-20 system (Jasper, 1958). The probeset (see Figure 5) covered the dorsolateral PFC and left inferior frontal gyrus (IFG), the latter being part of the PFC that is involved in speech processing (Riecker, Mathiak, Grodd, Hertrich, & Ackermann, 2005), including the production of the inner dialogue (McGuire et al., 1996). The distance between a light emitter and its corresponding detector (i.e., the inter-optode distance, the middle of which corresponds to a measurement channel), was approx. 3 cm. Data recording was performed at a sampling rate of 7.8125 Hz. The data were preprocessed with MATLAB version 2017a as described in section 2.5.

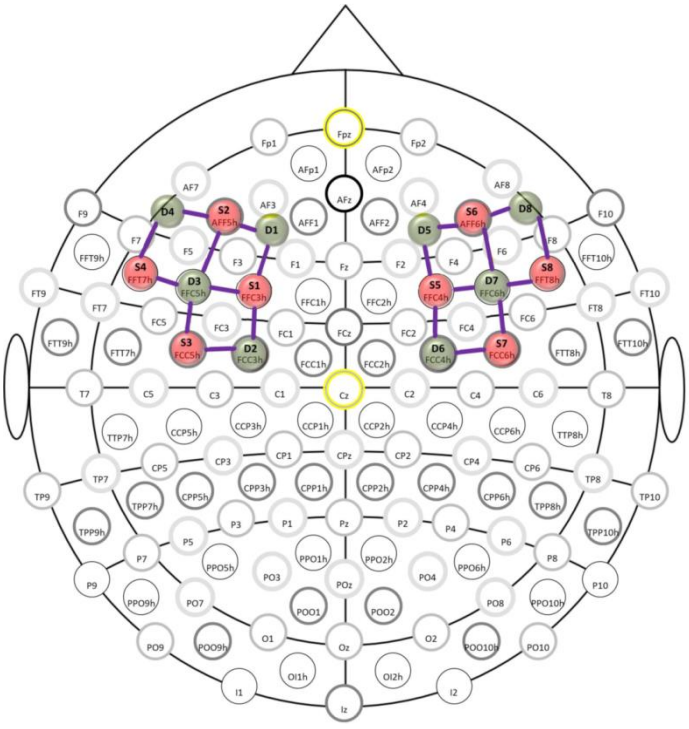


Figure 5. fNIRS probeset generated by NIRx Montage Editor (NIRX Medical Technologies). Left: Topological placement used of light emitters (filled in green), detectors (filled in red), and channels (purple) on a standard 10-20 EEG system. Right: Location of the easycap adjustment positions: Cz and Fpz (pink).

2.5 Data analysis

2.5.1. FNIRS data pre-processing. As mentioned above, despite their relative robustness, fNIRS recordings are not completely free of artifacts that can be caused by the participants' physiology (respiration, cardiac activity) and movements, which might cause relative motion between optical fibers and scalp (see Cooper et al., 2012; Pinti et al., 2018). The preprocessing steps were conducted in the following way.

First, data were corrected for high amplitude movement artifacts by the TDDR correction (Fishburn, Ludlum, Vaidya, & Medvedev, 2019), followed by bandpass filtering (0.01 - 0.1 Hz) and a correlation-based signal improvement (CBSI) (Cui, Bray, & Reiss, 2010). After that, visual data inspection for outlier-channels was conducted, and on average 2.2 channels per participant were interpolated with their surrounding channels. A following PCA-based Gaussian kernel filter was used for global signal correction (Zhang, Noah, & Hirsch, 2016) and data were z-standardized. Finally, we calculated the normalized area under the curve (nAUC) for this signal by placing its integral in relation to the duration of the investigated time period.

2.5.2 ROI definition. After pre-processing was completed, we combined 20 channels into the following 6 regions of interest (ROI), illustrated by Figure 6: DLPFC left (DLPFC-L), DLPFC right (DLPFC-R), and IFG (IFG). ROI definition was carried out on a data-driven basis combined with theoretical assumptions about the different sections of the PFC and their respective functions. Thereby, we visually inspected for all channels a timeframe from 5 - 30 sec after the initial *burst* phase had started. Channels were combined based on the anatomical topology, theoretical knowledge and hemodynamic activation shown during this period.

2.5.3 Statistical Analyses. We employed linear mixed-effect analyses using statistical software R (R Core Team, 2020) with the lme4 package (Bates, Mächler,

Bolker, & Walker, 2014). The assumptions of homoscedasticity and normality were verified by visual inspection of residual plots. The *p*-values were obtained by likelihood ratio tests of the full model (with the effect of the investigated parameter) tested against a reduced model (without its effect). In case of a significant result, further model analyses were applied using the *report* package of Makowski, Lüdecke, and Ben-Schachar (2020). Standardized parameters were obtained by fitting a model on a standardized version of the dataset.

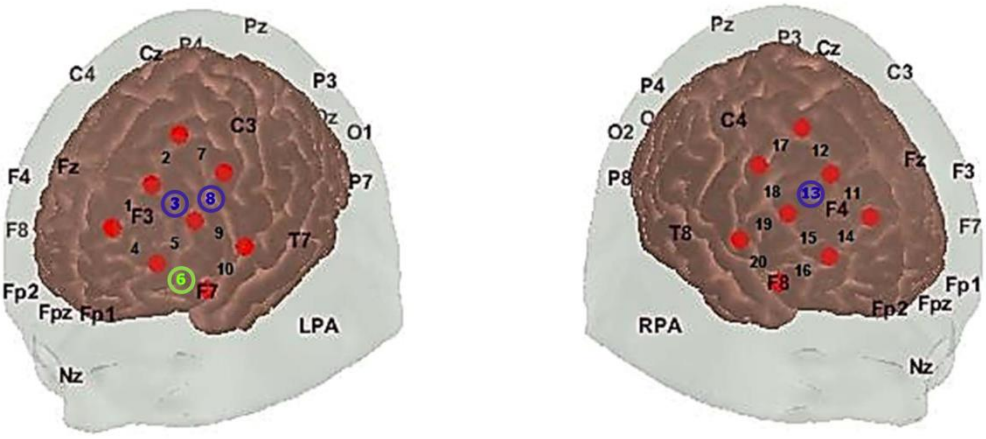


Figure 6. Anatomical allocation of the channels to the brain regions and ROI definition. DLPFC left: channels 3, 8; DLPFC right: channels 13; IFG: 6. The figure was generated by MATLAB based Atlas Viewer (Aasted et al., 2015) with manually added channel numbers.

3 Results

In this study we examined hemodynamic cortical activation within specified time windows during a time-critical emergency simulation. Thereby, we aimed to evaluate whether these observations can be used to assess cognitive load induced by the simulation design. In particular, for four defined time periods (initial *burst* and following t0 – t20, t20 – t40, and t40 – t60), we examined the impact of *real difficulty* of the level (percentage of participants who failed to complete all tasks within the specified time limit) and the participants' *performance* per level (binary indicator of whether the level was completed successfully or not) on the hemodynamic cortical activation level (nAUC: time-normalized integral of the CBSI signal) in a specified ROI. Furthermore, we evaluated the impact of nAUC on subjective ratings of cognitive load for the whole simulation level, acquired via NASA-TLX questionnaire.

The structure of the following sections corresponds to the structure of Table 1, which provides a general overview of all achieved results. Thereby we proceed sequentially through the table from top to bottom, providing detailed statistics for all specified time periods.

3.0 Manipulation check

Supporting the validity of the simulation design, we found a significant positive effect of 'scenario' on perceived cognitive load in relation to all acquired items of NASA-TLX questionnaire. This reflects that the scenario "*Train Crash*" was perceived as more challenging than "*Fire*" regarding the subjective rating of *mental demand*, *time demand* and *effort*: *mental demand* ($\chi^2(1) = 9.11$, $p = .002$, *beta* = 5.56, $p < .01$); *time demand* ($\chi^2(1) = 4.49$, $p = .034$, *beta* = 10.25, $p < .01$), *effort* ($\chi^2(1) = 4.49$, $p = .034$, *beta* = 10.49, $p < .01$).

Table 1

Graphical summary of significant effects

	DLPFC		IFG
	left	right	
burst			
nAUC ~ real difficulty	+	+	
nAUC ~ performance		-	
mental_demand ~ nAUC			
time_demand ~ nAUC		+	
effort ~ nAUC			
t0 – t20			
nAUC ~ real difficulty			
nAUC ~ performance			
mental_demand ~ nAUC	-		
time_demand ~ nAUC	-		
effort ~ nAUC	-	-	
t20 – t40			
nAUC ~ real difficulty		-	
nAUC ~ performance			
mental_demand ~ nAUC		-	
time_demand ~ nAUC			
effort ~ nAUC			
t40 – t60			
nAUC ~ real difficulty			
nAUC ~ performance			
mental_demand ~ nAUC			
time_demand ~ nAUC			
effort ~ nAUC			

Note: Significant results are listed as + for positive and – for negative effects. *Real difficulty* (related to level): percentage of participants, who did not complete the level. *Performance* (related to participant): binary indicator of whether the level was completed successfully or not. *nAUC*: time-normalized CBSI integral. Regions of interest (ROI) are labelled as follows: DLPFC for dorsolateral prefrontal cortex; IFG for inferior frontal gyrus.

3.1 Initial *burst*

3.1.1 *Real difficulty*. To evaluate the effect of *real difficulty* on neuronal activity within defined ROIs, we employed linear mixed-effect analysis with intercepts for participants entered as random effects and *real difficulty* as a fixed effect. We found a significantly positive effect of *real difficulty* on hemodynamic cortical activation in DLPFC-L and DLPFC-R ROIs, meaning that higher levels of *real difficulty* were associated with higher hemodynamic activity. In contrast, no significant association was found between *real difficulty* and hemodynamic cortical activation within the IFG ROI. For detailed statistics see Table 2.

Table 2

nAUC ~ real difficulty during initial burst

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	4.49	0.034	0.19	0.02	2.24	< 0.05
DLPFC-R	6.43	0.011	0.55	0.02	2.82	< 0.05
IFG	0.21	0.210				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

3.1.2 Performance. To determine the effect of participants' performance on neuronal activity within the defined ROIs, we used a linear mixed-effect analysis with intercepts for participants entered as random effects and the binary indicator of whether the level was successfully completed or not as a fixed effect. We found a significant main effect of performance on the nAUC only within the DLPFC-R ROI. This effect was significantly negative, meaning that higher level of performance was associated with lower hemodynamic activity. For detailed statistics see Table 3.

Table 3

nAUC ~ performance during initial burst

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	0.27	0.602				
DLPFC-R	4.03	0.045	0.54	0.01	-1.51	< 0.05
IFG	0.20	0.657				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

3.1.3 Perceived cognitive load. To evaluate the effect of participants' neuronal activity on subjective ratings of cognitive load, we employed a linear mixed-effect analysis with intercepts for participants entered as random effects and scenario and nAUC within investigated ROI as fixed effects. The analysis was performed for all combinations of ROIs and selected NASA-TLX subscales.

We found no significant associations between perceived cognitive load and hemodynamic cortical activation in terms of *mental demand* and *effort* (see Tables 4 & 6). Regarding *time demand*, a significant positive effect of neuronal activation within DLPFC-R ROI was found, implying that participants exhibiting higher neuronal activation in these cortex areas also perceived higher time pressure (see Table 5).

Table 4

mental demand ~ nAUC during initial burst

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	0.07	0.799				
DLPFC-R	1.00	0.317				
IFG	0.03	0.865				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

.

Table 5

time demand ~ nAUC during initial burst

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	3.25	0.710				
DLPFC-R	5.12	0.024	0.40	0.07	0.87	< 0.05
IFG	0.15	0.701				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

Table 6

effort ~ nAUC during initial burst

	$\chi^2(1)$	*p*	cond. R^2	marg. R^2	*beta*	*p*
DLPFC-L	0.67	0.411				
DLPFC-R	0.97	0.324				
IFG	0.00	0.991				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

3.2 t0 – t20

3.2.1 *Real difficulty*. To determine the effect of *real difficulty* on neuronal activity, we employed linear mixed-effect analysis with intercepts for participants entered as random effects and *real difficulty* as a fixed effect for each of the specified ROIs separately. We found no significant effects within any of the ROIs (see Table 7).

3.2.2 Performance. To determine the effect of participants' performance on neuronal activity within the defined ROIs, we used a linear mixed-effect analysis with intercepts for participants entered as random effects and the binary indicator of whether the level was successfully completed or not as a fixed effect. We found no significant effects within any of the ROIs (see Table 8).

Table 7

nAUC ~ real difficulty during t0 – t20

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	1.29	0.256				
DLPFC-R	0.38	0.539				
IFG	0.01	0.904				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

Table 8

nAUC ~ performance during t0 – t20

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	0.02	0.890				
DLPFC-R	0.09	0.761				
IFG	0.91	0.340				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

3.2.3 Perceived cognitive load. To determine the effect of participants' neuronal activity on subjective ratings of cognitive load, we employed a linear mixed-effect analysis with intercepts for participants entered as random effects and scenario and nAUC within investigated ROI as fixed effects. The analysis was performed for all combinations of ROIs and selected NASA-TLX items.

We found a significant negative effect of hemodynamic cortical activity only within DLPFC-L ROI on the perceived *mental demand* (see Table 9), implying that higher

hemodynamic cortical activation directly after the *burst* phase was associated with the experience of lower *mental demand.* Likewise, subjective *Time demand* was significantly negatively associated with hemodynamic cortical activation in DLPFC-L ROI, i.e. the increased neuronal activity was related to the experience of less time pressure, whereas no significant effect for other POIs were found (see Table 10). Perceived *effort* was significantly associated with hemodynamic cortical activation within both ROIs related to DLPFC, again no effect of IFG ROI was observed (see Table 11). This effect was negative, meaning that the increased neuronal activity was related to the experience of less effort.

Table 9

mental demand ~ nAUC during t0-t20

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	4.34	0.037	0.76	0.03	-0.69	< 0.05
DLPFC-R	2.68	0.102				
IFG	0.00	0.950				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

Table 10

time demand ~ nAUC during t0-t20

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	5.90	0.015	0.41	0.06	-1.50	< 0.05
DLPFC-R	0.18	0.671				
IFG	0.83	0.361				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

Table 11

effort ~ nAUC during t0-t20

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	5.86	0.015	0.56	0.04	-1.11	< 0.05
DLPFC-R	7.33	0.007	0.57	0.05	-1.03	< 0.01
IFG	0.83	0.361				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

3.3 t20 – t40

3.3.1 *Real difficulty* of. To evaluate the effect of *real difficulty* on neuronal activation, we employed linear mixed-effect analysis with intercepts for participants entered as random effects and *real difficulty* as a fixed effect for each ROI respectively.

We found a significant negative association between *real difficulty* and cortical hemodynamics only within DLPFC-R ROI, meaning that higher *real difficulty* was associated with lower hemodynamic activation in these cortical areas. No significant effects were wound for other ROIs (see Table 12).

Table 12

nAUC ~ real difficulty during t20 – t40

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	0.77	0.380				
DLPFC-R	5.93	0.015	0.61	0.01	-2.35	< 0.05
IFG	1.39	0.238				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

3.3.2 Performance. To determine the effect of participants' performance on neuronal activity within the defined ROIs, we used a linear mixed-effect analysis with intercepts for participants entered as random effects and the binary indicator of whether the level was successfully completed or not as a fixed effect. We found no significant effects within any of the ROIs (see Table 13).

Table 13

nAUC ~ performance during t20 – t40

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	0.78	0.376				
DLPFC-R	0.95	0.330				
IFG	1.43	0.231				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

3.3.3 Perceived cognitive load. To determine the effect of participants' neuronal activity on subjective ratings of cognitive load, we employed a linear mixed-effect analysis with intercepts for participants entered as random effects as well as scenario and nAUC within investigated ROI as fixed effects. The analysis was performed for all combinations of ROIs and selected NASA-TLX items.

We found a significant negative association between nAUC and perceived *mental demand* within the DLPFC-R ROI, indicating that higher hemodynamic cortical activation in this cortical area during t20 – t40 after the *burst* phase was associated with lower perceived *mental demand*. We found no significant effects within any other ROIs in this regard (see Table 14). Likewise, no significant effects were found regarding perceived *time demand* and *effort* in all ROIs during t20 – t40 after the *burst* phase (see Table 15 &16).

Table 14

mental demand ~ nAUC during t20-t40

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	0.10	0.655				
DLPFC-R	5.78	0.016	0.77	0.04	-0.60	< 0.05
IFG	0.36	0.550				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

Table 15

time demand ~ nAUC during t20-t40

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	1.72	0.190				
DLPFC-R	0.32	0.569				
IFG	2.27	0.132				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

Table 16

effort ~ nAUC during t20-t40

	$\chi^2(1)$	*p*	cond. R^2	marg. R^2	*beta*	*p*
DLPFC-L	0.03	0.856				
DLPFC-R	1.63	0.202				
IFG	1.51	0.219				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

3.4.1 40 – t60

3.4.2 *Real difficulty*. To evaluate the effect of *real difficulty* on hemodynamic cortical activation within defined ROIs, we employed linear mixed-effect analysis with intercepts for participants entered as random effects. For this time period, we found no significant effects within any of the ROIs (see Table 17)

3.4.3 Performance. To analyze the relationship between the participants' performance and hemodynamic cortical activation, we applied a linear mixed-effect analysis with intercepts for participants entered as random effects and the binary indicator of whether the level was successfully completed or not as a fixed effect. We found no significant effects within any of the ROIs (see Table 18).

Table 17

nAUC ~ real difficulty during t40-t60

	$\chi^2(1)$	*p*	cond. R^2	marg. R^2	*beta*	*p*
DLPFC-L	0.01	0.918				
DLPFC-R	1.08	0.300				
IFG	0.80	0.371				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

Table 18

nAUC ~ real difficulty during t40-t60

	$\chi^2(1)$	*p*	cond. R^2	marg. R^2	*beta*	*p*
DLPFC-L	0.01	0.992				
DLPFC-R	0.29	0.593				
IFG	0.13	0.719				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

3.4.4 Perceived cognitive load. To determine the effect of participants' neuronal activity on subjective ratings of cognitive load, we employed a linear mixed-effect analysis with intercepts for participants entered as random effects and scenario and nAUC within investigated ROI as fixed effects. The analysis was performed for all combinations of ROIs and selected NASA-TLX items. Within this time window we found no significant effects in any of the ROIs neither regarding perceived *mental demand* (see Table 18) nor regarding *time demand* (see Table 19), or *effort* (see Table 20).

Table 19

mental demand ~ nAUC during t40-t60

	$\chi^2(1)$	*p*	cond. R^2	marg. R^2	*beta*	*p*
DLPFC-L	0.87	0.352				
DLPFC-R	1.66	0.197				
IFG	0.38	0.537				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

Table 20

time demand ~ nAUC during t40-t60

	$\chi^2(1)$	*p*	cond. R^2	marg. R^2	*beta*	*p*
DLPFC-L	0.73	0.393				
DLPFC-R	0.19	0.663				
IFG	0.69	0.407				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

Table 21

effort ~ nAUC during t40-t60

			cond.	marg.		
	$\chi^2(1)$	*p*	R^2	R^2	*beta*	*p*
DLPFC-L	0.58	0.448				
DLPFC-R	0.78	0.377				
IFG	0.17	0.678				

Note: cond. R^2: conditional R^2 representing models total explanatory power, marg. R^2: marginal R^2 representing explanatory power of fixes effects. Regions of interest (ROI) are labelled as follows: *DLPFC* for dorsolateral prefrontal cortex; *IFG* for inferior frontal gyrus; *L* for left; *R* for right.

4 Discussion

Determining cognitive load seems crucial for the development of training environments for life- and time-critical emergencies. Here, the use of fNIRS methodology may provide reliable results in this respect not only in the laboratory but also in ecologically more valid experiments, using averaged data of repeated measures of relatively cognitive tasks. At the same time, it remains unclear how to use fNIRS for assessing cognitive load during a prolonged single-trial activity involving heterogeneous tasks. In particular, because averaging over a too long period of time may well obscure potential differences in hemodynamic cortical activation, the question arises which time windows may be used sensibly to obtain reliable data on experienced cognitive load.

This study describes an attempt to solve this challenge by following a theory-driven top-down approach for investigating realistic time-critical emergency situations from the perspective of resource management. We used a game-based emergency simulation requiring time-critical coordination of emergency personnel to induce different levels of cognitive load. Considering Barrouillet et al. (2004) TBRS Model and previous results, we hypothesized that time periods directly following the initial *burst* phase might be well suited for reliable cognitive load assessment with fNIRS. Contrarily, we did not expect to find significant associations between hemodynamic cortical activation and cognitive load for the initial *burst* phases themselves in which participants should perform at the maximum of their individual capacity as well as for the later course of the simulation.

In accordance with our expectations and despite the fact that simulation levels varied significantly in terms of *real difficulty* and subjective cognitive load, only a few or no significant effect on hemodynamic cortical activation was observed during latter time intervals: t20 – t40 and t40 – t60 seconds after the end of the initial *burst*. These results substantiate that it is not always possible to determine differences in cognitive load by comparing random time intervals over different difficulty levels over a long and complex realistic task. Because a realistic cognitive task usually consists of heterogeneous subtasks, it cannot be guaranteed that all operators perform these tasks in the same order and using the same strategy. This leads to a lot of “noise” in neuronal activity when different mental tasks are performed continuously or even overlapping at different times by different participants making it almost impossible to pinpoint influences of specific variables.

This problem might be solved by searching for comparable time slots using the global knowledge about the nature of the performed task. As hypothesized, we found a significant association between hemodynamic cortical activation within different areas of DLPFC immediately after the initial *burst* phase (t0 - t20) and the subjective assessment of cognitive load of the entire level (by NASA-TLX). This association seems robust, as it appeared for all investigated subscales (*mental demand*, *time demand*, and *effort*). Surprisingly, participants who showed stronger hemodynamic cortical activation within this time frame perceived the entire level as less demanding for them and vice versa. This pattern of results provides an interesting insight into the nature of the *idle* phase following the initial *burst* phase. While the simulation log captured that all participants acted similarly, when managing initial resources during the first *burst* (i.e., allocating emergency personnel to their tasks) and thus may have relied on similar cognitive processes, there is no information about their cognitive engagement during the *idle* phase, because no logable behavior was performed in this time period. In principle, in this phase, they could either wait passively for the next *burst* or use this time for active monitoring and planning, which might result in better performance later on and thus to a reduction in the subjectively experienced cognitive load (Hancock, 1989). However, as no significant association between neuronal activity and final performance was observed for this time slot, another explanation seems more likely. It is conceivable that cognitively challenged participants tended to use the *idle* pause to relax as they might get tired of maintaining attention, which may have led to reduced hemodynamic cortical activation (Nihashi et al., 2019). In contrast, more successful players might use this phase to monitor the situation and plan ahead their next steps. Nevertheless, answering this question seems to require further investigation.

Because we assumed that during the initial *burst* phase all participants would operate at their cognitive maximum, we expected no differences in hemodynamic cortical activation between participants and *real difficulty* at this time. Surprisingly, we found a considerable positive association between hemodynamic cortical activation within wide areas of the DLPFC ROI and *real difficulty* of the level as well as with perceived time pressure during the initial *burst* phase. In contrast, negative associations with the actual performance were found within the right DLPFC. First of all, these results may be interpreted as substantiation of the “inverted-U” shaped association of cognitive load and performance (Csikszentmihalyi, 1987; Yerkes & Dodson, 1908), which assumes that the cognitive load should be kept within a certain range to obtain the best results.

Furthermore, the initial *burst* phase, which takes place at the very beginning of the level, appears nevertheless to be well suited to determine the degree of task difficulty and could therefore be used for real-time adaptation of training simulations. At the same time, however, further investigation is needed to determine whether this effect might persist under more stressful conditions. It is conceivable that our preliminary assumptions were correct, but the induced time pressure was not sufficient to make all participants work at their cognitive limit. In this case, a *burst* time slot would be well suitable for measuring cognitive load in situations with low to medium time pressure, while for measuring cognitive load under high time pressure other options need to be identified.

All above-mentioned results refer to neuronal activity within broader ranges of the DLPFC. In this study, we found almost no association of IFG activation and participants' cognitive load. This may be due to the nature of the experimental task, which did not require difficult calculations or other cognitive manipulations necessitating an inner dialogue. Despite this result, the empirical evidence indicates that this region may be sensitive to changes in cognitive load (Ayaz et al., 2012; de Fockert & Theeuwes, 2012) and should therefore be investigated in future research.

4.1 Methodological strengths and constraints

Analytic approaches for simulated training environments are often based on data-driven probabilistic evaluations (Magerko, Stensrud, & Holt, 2006; Spronck, Ponsen, Sprinkhuizen-Kuyper, & Postma, 2006; Zook & Riedl, 2012). However, these seem insufficient for modeling cognitive user states, which can be "directly" (Brunken et al., 2003) assessed via neurophysiological methods (for review see: Kivikangas et al., 2011). Compared to other neuroimaging techniques, the use of the fNIRS methodology seems advantageous due to its relatively low cost, high mobility, robustness against various artifacts, and reliability when used on averaged data, whereas single-user single-trial evaluations still remain challenging (Herold et al., 2018; Scholkmann et al., 2014). This study takes a step in this direction by proposing a theory-based approach to the choice of time frames suitable for assessing cognitive load in realistic single-trial time-critical emergency situations related to resource management. Unlike data-driven approaches such as machine learning, this approach might be generalized to similar environments, which however raises further questions. As mentioned above, we have to

investigate whether a finer clustering of training scenarios is needed for the selection of the appropriate metrics, brain regions, and time frames for data collection. Also, in this constellation fNIRS cannot be used as a stand-alone methodology, because it needs log data from the simulation to calculate the required time periods, which would result in relatively complex assessment installation. Thus the obtained results might be used to improve laboratory systems with the aim of conducting mobile, ecologically valid experiments.

4.2 Conclusion

In this study, we presented a simple method to determine time periods that qualify for cognitive load detection in a single-trial time-critical resource-management emergency situation using the fNIRS method in combination with TBRS theory. Detection of proposed time periods is based on log data and can be easily run in the background. We found significant associations between cognitive load and neuronal activity within DLPFC during chosen time periods, whereas only a few or no significant effects were observed during later time intervals, substantiating that it is not always possible to determine differences in cognitive load by comparing random periods of time. We found no significant dependencies within IFG. These results illustrate how knowledge of task structure may be used advantageously for the identification of cognitive load. Although requiring further investigation in terms of reliability and generalizability, the presented approach seems promising evidence that fNIRS might be suitable for the assessment of cognitive load beyond classical experimental set-ups.

References

Aasted, C. M., Yücel, M. A., Cooper, R. J., Dubb, J., Tsuzuki, D., Becerra, L., Petkov, M. P., Borsook, D., Dan, I., & Boas, D. A. (2015). Anatomical guidance for functional near-infrared spectroscopy: AtlasViewer tutorial. *Neurophotonics, 2*(2), 020801.

Ahmad, R. F., Malik, A. S., Kamel, N., & Reza, F. (2016). *Machine learning approach for classifying the cognitive states of the human brain with functional magnetic resonance imaging (fMRI).* Paper presented at the 2016 6th International Conference on Intelligent and Advanced Systems (ICIAS).

Anderson, K. J. (1994). Impulsitivity, caffeine, and task difficulty: A within-subjects test of the Yerkes-Dodson law. *Personality and Individual Differences, 16*(6), 813-829.

Appel, T., Sevcenko, N., Wortha, F., Tsarava, K., Moeller, K., Ninaus, M., Kasneci, E., & Gerjets, P. (2019). *Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures.* Paper presented at the 2019 International Conference on Multimodal Interaction.

Ayaz, H., Izzetoglu, M., Bunce, S., Heiman-Patterson, T., & Onaral, B. (2007). *Detecting cognitive activity related hemodynamic signal for brain computer interface using functional near infrared spectroscopy.* Paper presented at the Neural Engineering, 2007. CNE'07. 3rd International IEEE/EMBS Conference on.

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.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General, 133*(1), 83.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 570.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Berthold, A., & Jameson, A. (1999). Interpreting symptoms of cognitive load in speech input. In *UM99 user modeling* (pp. 235-244): Springer.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational psychologist, 38*(1), 53-61.

Brünken, R., Seufert, T., & Paas, F. (2010). Measuring cognitive load.

Bruno, J. L., Baker, J. M., Gundran, A., Harbott, L. K., Stuart, Z., Piccirilli, A. M., Hosseini, S. H., Gerdes, J. C., & Reiss, A. L. (2018). Mind over motor mapping: Driver response to changing vehicle dynamics. *Human brain mapping, 39*(10), 3915-3927.

Buchwald, M., Kupiński, S., Bykowski, A., Marcinkowska, J., Ratajczyk, D., & Jukiewicz, M. (2019). *Electrodermal activity as a measure of cognitive load: a methodological approach.* Paper presented at the 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA).

Camos, V., Portrat, S., Vergauwe, E., & Barrouillet, P. (2007). *The cognitive cost of executive functions*. Paper presented at the Joint Meeting of the EPS and the Psychonomic Society, Edinburgh (Great-Britain).

Cooper, R., Selb, J., Gagnon, L., Phillip, D., Schytz, H. W., Iversen, H. K., Ashina, M., & Boas, D. A. (2012). A systematic comparison of motion artifact correction techniques for functional near-infrared spectroscopy. *Frontiers in neuroscience, 6*, 147.

Cope, M., Delpy, D., Reynolds, E., Wray, S., Wyatt, J., & Van der Zee, P. (1988). Methods of quantitating cerebral near infrared spectroscopy data. In *Oxygen Transport to Tissue X* (pp. 183-189): Springer.

Csikszentmihalyi, M. (1987). *Das flow-Erlebnis: jenseits von Angst und Langeweile: im Tun aufgehen*: Klett-Cotta.

Cui, X., Bray, S., & Reiss, A. L. (2010). Functional near infrared spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics. *Neuroimage, 49*(4), 3039-3046.

de Fockert, J. W., & Theeuwes, J. (2012). Role of frontal cortex in attentional capture by singleton distractors. *Brain and cognition, 80*(3), 367-373.

Derosière, G., Mandrick, K., Dray, G., Ward, T. E., & Perrey, S. (2013). NIRS-measured prefrontal cortex activity in neuroergonomics: strengths and weaknesses. *Frontiers in human neuroscience, 7*, 583.

EASYCAP. EASYCAP EEG Recording Caps and Related Products. Retrieved 19.02.2021, from

Eggemeier, F. T., Wilson, G. F., Kramer, A. F., & Damos, D. L. (1991). Workload assessment in multi-task environments. *Multiple-task performance*, 207-216.

Ehlis, A.-C., Herrmann, M., Wagener, A., & Fallgatter, A. (2005). Multi-channel near-infrared spectroscopy detects specific inferior-frontal activation during incongruent Stroop trials. *Biological psychology, 69*(3), 315-331.

Fallgatter, A., Ehlis, A., Wagener, A., Michel, T., & Herrmann, M. (2004). Near-infrared spectroscopy in psychiatry. *Der Nervenarzt, 75*(9), 911.

Fishburn, F. A., Ludlum, R. S., Vaidya, C. J., & Medvedev, A. V. (2019). Temporal Derivative Distribution Repair (TDDR): A motion correction method for fNIRS. *Neuroimage, 184*, 171-179.

Fishburn, F. A., Norr, M. E., Medvedev, A. V., & Vaidya, C. J. (2014). Sensitivity of fNIRS to cognitive state and load. *Frontiers in human neuroscience, 8*, 76.

Fowler, A., Nesbitt, K., & Canossa, A. (2019). *Identifying Cognitive Load in a Computer Game: An exploratory study of young children.* Paper presented at the 2019 IEEE Conference on Games (CoG).

Fuster, J. (2001). The prefrontal cortex—an update: time is of the essence. *Neuron, 30*(2), 319-333.

Fuster, J. (2015). *The prefrontal cortex*: Academic Press.

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., & Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in neuroscience, 8*, 385.

Haerle, S. K., Daly, M. J., Chan, H. H., Vescan, A., Kucharczyk, W., & Irish, J. C. (2013). Virtual surgical planning in endoscopic skull base surgery. *The Laryngoscope, 123*(12), 2935-2939.

Hancock, P. (1989). The effect of performance failure and task demand on the perception of mental workload. *Applied Ergonomics, 20*(3), 197-205.

Hart, S. G. (2006). *NASA-task load index (NASA-TLX); 20 years later.* Paper presented at the Proceedings of the human factors and ergonomics society annual meeting.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183): Elsevier.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2014). Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS. *Frontiers in human neuroscience, 7*, 935.

Herold, F., Wiegel, P., Scholkmann, F., & Müller, N. G. (2018). Applications of functional near-infrared spectroscopy (fNIRS) neuroimaging in exercise–cognition science: a systematic, methodology-focused review. *Journal of clinical medicine, 7*(12), 466.

Herrmann, M. J., Walter, A., Schreppel, T., Ehlis, A. C., Pauli, P., Lesch, K. P., & Fallgatter, A. (2007). D4 receptor gene variation modulates activation of prefrontal cortex during working memory. *European Journal of Neuroscience, 26*(10), 2713-2718.

Hoge, R. D., Atkinson, J., Gill, B., Crelier, G. R., Marrett, S., & Pike, G. B. (1999). Linear coupling between cerebral blood flow and oxygen consumption in activated human cortex. *Proceedings of the National Academy of Sciences, 96*(16), 9403-9408.

Ikehara, C. S., & Crosby, M. E. (2005). *Assessing cognitive load with physiological sensors.* Paper presented at the Proceedings of the 38th annual hawaii international conference on system sciences.

Jasper, H. (1958). Report of the committee on methods of clinical examination in electroencephalography. *Electroencephalogr Clin Neurophysiol, 10*, 370-375.

Jobsis, F. F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science, 198*(4323), 1264-1267.

Johannsen, G. (1979). Workload and workload measurement. In N. Moray (Ed.), *Mental Workload* (pp. 3-11): Springer.

Johnson, J. G., Rodrigues, D. G., Gubbala, M., & Weibel, N. (2018). *Holocpr: Designing and evaluating a mixed reality interface for time-critical emergencies.* Paper presented at the Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare.

Kammer, T., & Karnath, H.-O. (2006). Manifestationen von Frontalhirnschädigungen. In *Neuropsychologie* (pp. 489-500): Springer.

Kiili, K., Lindstedt, A., & Ninaus, M. (2018). *Exploring characteristics of students' emotions, flow and motivation in a math game competition.* Paper presented at the GamiFIN.

Kim, S. G., Rostrup, E., Larsson, H. B., Ogawa, S., & Paulson, O. B. (1999). Determination of relative CMRO2 from CBF and BOLD changes: significant increase of oxygen consumption rate during visual stimulation. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 41*(6), 1152-1161.

Kincaid, J. P., Donovan, J., & Pettitt, B. (2003). Simulation techniques for training emergency response. *International Journal of Emergency Management, 1*(3), 238-246.

Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., & Ravaja, N. (2011). A review of the use of psychophysiological methods in game research. *journal of gaming & virtual worlds, 3*(3), 181-199.

Kober, S. E., Wood, G., Kiili, K., Moeller, K., & Ninaus, M. (2020). Game-based learning environments affect frontal brain activity. *PloS one, 15*(11), e0242573.

Lépine, R., Bernardin, S., & Barrouillet, P. (2005). Attention switching and working memory spans. *European Journal of Cognitive Psychology, 17*(3), 329-345.

Li, C., Gong, H., Gan, Z., & Luo, Q. (2005). *Monitoring of prefrontal cortex activation during verbal n-back task with 24-channel functional NIRS imager.* Paper presented at the Optics in Health Care and Biomedical Optics: Diagnostics and Treatment II.

Liang, Y., Liang, W., Qu, J., & Yang, J. (2018). *Experimental study on EEG with different cognitive load.* Paper presented at the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC).

Liefooghe, B., Barrouillet, P., Vandierendonck, A., & Camos, V. (2008). Working memory costs of task switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(3), 478.

Lim, Y. M., Ayesh, A., & Stacey, M. (2014). *Using mouse and keyboard dynamics to detect cognitive stress during mental arithmetic.* Paper presented at the Science and Information Conference.

Liu, Y., Lan, Z., Tschoerner, B., Virdi, S. S., Cui, J., Li, F., Sourina, O., Zhang, D., Chai, D., & Müller-Wittig, W. (2020). *Human Factors Assessment in VR-based Firefighting Training in Maritime: A Pilot Study.* Paper presented at the 2020 International Conference on Cyberworlds (CW).

Magerko, B., Stensrud, B. S., & Holt, L. S. (2006). *Bringing the schoolhouse inside the box-a tool for engaging, individualized training*. Retrieved from

Magnusdottir, E. H., Borsky, M., Meier, M., Johannsdottir, K., & Gudnason, J. (2017). Monitoring cognitive workload using vocal tract and voice source features. *Periodica Polytechnica Electrical Engineering and Computer Science, 61*(4), 297-304.

Makowski, D., Lüdecke, D., & Ben-Schachar, M. (2020). Automated reporting as a practical tool to improve reproducibility and methodological best practices adoption. *J. Open Source Softw, 5*, 2815.

McDuff, D., Gontarek, S., & Picard, R. (2014). *Remote measurement of cognitive stress via heart rate variability.* Paper presented at the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

McGuire, P., Silbersweig, D., Murray, R., David, A., Frackowiak, R., & Frith, C. (1996). Functional anatomy of inner speech and auditory verbal imagery. *Psychological medicine, 26*(1), 29-38.

McIntosh, M. A., Shahani, U., Boulton, R. G., & McCulloch, D. L. (2010). Absolute quantification of oxygenated hemoglobin within the visual cortex with functional near infrared spectroscopy (fNIRS). *Investigative ophthalmology & visual science, 51*(9), 4856-4860.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience, 24*(1), 167-202.

Montani, F., Vandenberghe, C., Khedhaouria, A., & Courcy, F. (2020). Examining the inverted U-shaped relationship between workload and innovative work behavior: The role of work engagement and mindfulness. *Human Relations, 73*(1), 59-93.

Murkin, J. M., & Arango, M. (2009). Near-infrared spectroscopy as an index of brain and tissue oxygenation. *British journal of anaesthesia, 103*(suppl_1), i3-i13.

Nebel, S., & Ninaus, M. (2019). New perspectives on game-based assessment with process data and physiological signals. In *Game-Based Assessment Revisited* (pp. 141-161): Springer.

Nihashi, T., Ishigaki, T., Satake, H., Ito, S., Kaii, O., Mori, Y., Shimamoto, K., Fukushima, H., Suzuki, K., & Umakoshi, H. (2019). Monitoring of fatigue in radiologists during prolonged image interpretation using fNIRS. *Japanese journal of radiology, 37*(6), 437-448.

Ninaus, M., Witte, M., Kober, S. E., Friedrich, E. V., Kurzmann, J., Hartsuiker, E., Neuper, C., & Wood, G. (2013). Neurofeedback and serious games. In E. T. M. Connolly, T. Boyle, G. Hainey, P. Baxter, & Moreno-ger (Eds.), *Psychology, Pedagogy, and Assessment in Serious Games* (Vol. i, pp. 82-110). USA: IGI Global.

Nippert, A. R., Biesecker, K. R., & Newman, E. A. (2018). Mechanisms mediating functional hyperemia in the brain. *The Neuroscientist, 24*(1), 73-83.

NIRX Medical Technologies. Retrieved 15.10.2020, from website: https://nirx.net/nirsport

O'Donnell, R., & Eggemeier, F. (1986). Workload assessment methodology. Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance. KR Boff, L. Kaufman and JP Thomas. In: John Wiley and Sons, Inc.

Orru, G., & Longo, L. (2018). *The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and Germane loads: a review.* Paper presented at the International Symposium on Human Mental Workload: Models and Applications.

Parasuraman, R., & Rizzo, M. (2006). Introduction to Neuroergonomics. In.

Pinti, P., Aichelburg, C., Gilbert, S., Hamilton, A., Hirsch, J., Burgess, P., & Tachtsidis, I. (2018). A review on the use of wearable functional near-infrared spectroscopy in naturalistic environments. *Japanese Psychological Research, 60*(4), 347-373.

Portrat, S. (2008). *Working memory and executive functions: The Time-Based Resource-Sharin account* Université de Bourgogne. Dijon.

Pringle, J., Roberts, C., Kohl, M., & Lekeux, P. (1999). Near infrared spectroscopy in large animals: optical pathlength and influence of hair covering and epidermal pigmentation. *The Veterinary Journal, 158*(1), 48-52.

Promotion Software GmbH. (1999). World of Emergency. Retrieved August 26, 2019, from Promotion Software GmbH website: https://www.world-of-emergency.com/?lang=en

R Core Team. (2020). R: A Language and Environment for Statistical Computing. Retrieved from https://www.R-project.org/

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In *Advances in psychology* (Vol. 52, pp. 185-218): Elsevier.

Riecker, A., Mathiak, K., Grodd, W., Hertrich, I., & Ackermann, H. (2005). Functional MRI reveals two distinct cerebral networks subserving speech motor control. *The Journal of the Acoustical Society of America, 117*(4), 2574-2574.

Ruiz, N., Liu, G., Yin, B., Farrow, D., & Chen, F. (2010). Teaching athletes cognitive skills: detecting cognitive load in speech input. *Proceedings of HCI 2010 24*, 484-488.

Scerbo, M. W. (1996). Theoretical perspectives on adaptive automation. In R. Parasuraman & M. Mouloua (Eds.), *Automation and Human Performance: Theory and Applications* (pp. 37-64): CRC Press.

Scholkmann, F., Kleiser, S., Metz, A. J., Zimmermann, R., Pavia, J. M., Wolf, U., & Wolf, M. (2014). A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *Neuroimage, 85*, 6-27.

Sevcenko, N., Ninaus, M., Wortha, F., Moeller, K., & Gerjets, P. (2021). *Measuring Cognitive Load Using In-Game Metrics Of A Serious Simulation Game*. osf.io/b87ag.

Smith, E. E., & Jonides, J. (1997). Working memory: A view from neuroimaging. *Cognitive psychology, 33*(1), 5-42.

Spronck, P., Ponsen, M., Sprinkhuizen-Kuyper, I., & Postma, E. (2006). Adaptive game AI with dynamic scripting. *Machine Learning, 63*(3), 217-248.

Strangman, G., Boas, D. A., & Sutton, J. P. (2002). Non-invasive neuroimaging using near-infrared light. *Biological psychiatry, 52*(7), 679-693.

Strangman, G., Goldstein, R., Rauch, S. L., & Stein, J. (2006). Near-infrared spectroscopy and imaging for investigating stroke rehabilitation: test-retest reliability and review of the literature. *Archives of physical medicine and rehabilitation, 87*(12), 12-19.

Strangman, G. E., Li, Z., & Zhang, Q. (2013). Depth sensitivity and source-detector separations for near infrared spectroscopy based on the Colin27 brain template. *PloS one, 8*(8), e66319.

Temple, J. G., Dember, W. N., Warm, J. S., Jones, K. S., & LaGrange, C. M. (1997). *The effects of caffeine on performance and stress in an abbreviated vigilance task.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Thier, P. (2006). Die funktionelle Architektur des präfrontalen Kortex. In *Neuropsychologie* (pp. 471-478): Springer.

Unni, A., Ihme, K., Jipp, M., & Rieger, J. W. (2017). Assessing the driver's current level of working memory load with high density functional near-infrared spectroscopy: a realistic driving simulator study. *Frontiers in human neuroscience, 11*, 167.

Watters, P. A., Martin, F., & Schreter, Z. (1997). Caffeine and cognitive performance: The nonlinear Yerkes–Dodson law. *Human Psychopharmacology: Clinical and Experimental, 12*(3), 249-257.

Witte, M., Ninaus, M., Kober, S. E., Neuper, C., & Wood, G. (2015). Neuronal correlates of cognitive control during gaming revealed by near-infrared spectroscopy. *PloS one, 10*(8), e0134816.

Xu, X., Deng, Z.-Y., Huang, Q., Zhang, W.-X., Qi, C.-z., & Huang, J.-A. (2017). Prefrontal cortex-mediated executive function as assessed by Stroop task performance associates with weight loss among overweight and obese adolescents and young adults. *Behavioural brain research, 321*, 240-248.

Yap, T. F., Epps, J., Ambikairajah, E., & Choi, E. H. (2011). *Voice source features for cognitive load classification.* Paper presented at the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

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.

Zhang, X., Noah, J. A., & Hirsch, J. (2016). Separation of the global and local components in functional near-infrared spectroscopy signals using principal component spatial filtering. *Neurophotonics, 3*(1), 015004.

Zook, A. E., & Riedl, M. O. (2012). *A temporal data-driven player model for dynamic difficulty adjustment.* Paper presented at the Eighth Artificial Intelligence and Interactive Digital Entertainment Conference.

4 Study 3

The following is an author manuscript of an article submitted to Journal on Multimodal User Interfaces, available online under https://www.springer.com/journal/12193.

Please cite as:

Sevcenko, N., Appel, T., Ninaus, M., Wortha, F., Moeller, K., & Gerjets, P. (2022). Theory-based approach for assessing cognitive load during time-critical resource-managing human-computer interactions: an eye-tracking study. *Submitted manuscript*

Theory-based approach for assessing cognitive load during time-critical resource-managing human-computer interactions: an eye-tracking study

Authors:

Natalia Sevcenko (1, 2), Tobias Appel (3), Manuel Ninaus (4, 7),
Korbinian Moeller (5, 6, 7), Peter Gerjets (6, 7)

Affiliations:

1 - Daimler Truck AG, Stuttgart, Germany

2 - Psychology, Faculty of Science, Eberhard Karl University, Tuebingen, Germany

3 - Hector Research Institute of Education Sciences and Psychology, Faculty of Economics and Social Sciences, Tuebingen, Germany

4 - Department of Psychology, University of Innsbruck, Austria

5 - Centre for Mathematical Cognition, School of Science, Loughborough University, Loughborough, UK

6 - Leibniz-Institut fuer Wissensmedien, Tuebingen, Germany

7 - LEAD Graduate School & Research Network, University of Tuebingen, Germany

keywords: cognitive load, adaptation, human-machine, HCI, time-critical, serious game, cognitive ergonomics

Abstract

Computerized systems are taking on increasingly complex tasks. Consequently, monitoring automated computerized systems is becoming increasingly demanding for human operators, which is particularly relevant in time-critical situations. A possible solution might be adapting human-computer interfaces (HCI) to the operators' cognitive load. Here, we present a novel approach for theory-based measurement of cognitive load based on tracking eye movements of 42 participants while playing a serious game simulating time-critical situations that required resource management at different levels of difficulty. Gaze data was collected within narrow time periods, calculated based on log data interpreted in the light of the time-based resource-sharing model. Our results indicated that eye fixation frequency, saccadic rate, and pupil diameter significantly predicted task difficulty, while performance was best predicted by eye fixation frequency. Subjectively perceived cognitive load was significantly associated with the rate of microsaccades. Moreover our results indicated that more successful players tended to use breaks in gameplay to actively monitor the scene, while players who use these times to rest are more likely to fail the level. The presented approach seems promising for measuring cognitive load in realistic situations, considering adaptation of HCI.

1 Introduction

Our daily lives are becoming more and more automated, with computerized systems such as autopilots, speed and lane assistants, robotic surgeons, etc. taking on increasingly complex tasks. Thus, a human operator is supported considerably by such systems on the one hand. However, monitoring of such automated systems is becoming increasingly demanding due to the rising complexity of tasks they are capable of executing. This problem is particularly evident in unexpected time-critical situations, where the operator does not have time for a detailed analysis of the interface and therefore relies more than usual on cognitive ergonomics.

A possible solution to this challenge may be the development of intelligent human-computer interfaces (HCIs) capable of adapting their demands and appearance (e.g., minimizing displayed information when the operator is cognitively overloaded) to the situation as well as to the operators' cognitive and emotional states to optimize performance and prevent failures. The so-called cognitive load on the operator seems particularly relevant in this case, because it is considered to reflect the degree to which available cognitive resources are engaged in the task at hand (Babiloni, 2019) and thus can be used to predict operators' performance. Accordingly, detecting the actual level of cognitive load seems highly relevant in a variety of realistic settings (P. Gerjets et al., 2014), for example, to change the appearance of the interface according to the operators' cognitive load. Empirical evidence indicates that differently designed HCIs may cause different levels of cognitive load while performing the same task. As one example, Charabati et al. (2009) compared interfaces designed to monitor anesthesia parameters during a surgery and reported that participants rated their cognitive load significantly lower when using a mixed numerical–graphical interface compared to the numerical and advanced-graphical interfaces. Another example was provided by the study of Oviatt (2006) who found that students performed significantly better at solving mathematical problems when using a digital pen and paper interface compared to graphical tablet interfaces. At the same time, online adaptation of the environment to the operators' current cognitive load was shown to lead to a significant performance improvements (Walter et al., 2017). Fortunately, digital systems easily allow for collecting individual user data that may well be used to model cognitive or emotional states of the operator (Nebel & Ninaus, 2019) including but not limited to cognitive load. In the current study we investigated whether eye-tracking features collected during narrow time

intervals, which were predefined using a theoretical approach, may be used successfully to measure cognitive load during time-critical emergency simulation.

1.1 Cognitive Load

The concept of cognitive load is based on the realization that cognitive resources are limited (G. A. Miller, 1956). It can be understood as the degree of “how hard the brain is working to meet task demands” (Ayaz et al., 2012, p. 36). At the same time, it needs to be considered that cognitive load evolves as a complex interplay between different task demands and mental processes (Babiloni, 2019), and thus represents a dynamic variable that fluctuates during task accomplishment.

An association between cognitive load and human performance was demonstrated in a variety of realistic settings such as e-learning (Oviatt, 2006; Walter et al., 2017), transportation (Fan & Smith, 2017a; G. Hancock et al., 2012; P. Hancock, 1989), office work (Aasted et al., 2015; Smith-Jackson & Klein, 2009) and medicine (Yurko et al., 2010). It has been observed , that this relation seems to be shaped like an “inverted-U” (Yerkes & Dodson, 1908) with best performance under medium cognitive load. Additionally, it seems to be associated with Csikszentmihalyis’ concept of “flow” (Csikszentmihalyi, 1975; Kiili et al., 2018), which characterizes a state of total concentration on the task at hand and also assumes that performance usually declines when cognitive demands are boring or overstraining(e.g., Anderson, 1994; Cummings & Nehme, 2009; Montani et al., 2020; Yerkes & Dodson, 1908). As such, this indicates that human-computer interaction might be optimized by keeping its’ operators’ cognitive load at a medium level (Orru & Longo, 2019).

Kohlmorgen et al. (2007) substantiated this consideration by showing that an adaptive reduction of cognitive load improved driving performance under real traffic conditions. Furthermore, Yuksel et al. (2016) reported on the benefits of adaptively increasing task complexity on learning performance in pianists. Moreover, Walter et al. (2017) observed significant improvement in learners’ math performance when using an electroencephalography (EEG)-based adaptive learning environment. However, it is worth noting that although the system used in this experiment did not have to be calibrated individually, improvements achieved were comparable to those obtained when using traditional error-based adaptation. Taken together, these results indicate that

online adaptation to cognitive load is beneficial as well as practicable. Thus, considering that cognitive load is a dynamic variable that fluctuates over time during task completion, we need a measurement method able to capture cognitive load at the early stages of task processing to adapt HCI to it in a timely manner, because the earlier a non-responsive cognitive state can be detected, the earlier HCI can be adapted accordingly.

Measurement techniques of cognitive load can be classified into four main categories: i) performance-based, ii) subjective, iii) behavioral, and iv) physiological measurements (Brünken et al., 2010; Eggemeier et al., 1991; Johannsen, 1979). *Performance-based* measurements rely upon user performance, for example, the rate of correct responses while solving a sequence of arithmetic tasks. These measurements are hardly applicable in the context of human-computer interaction (HCI), where intermediate results can seldom be identified. *Subjective* measurements are usually obtained by (standardized) questionnaires such as SWAT (Reid & Nygren, 1988) and NASA-TLX (Hart & Staveland, 1988). These measurements are well validated, easy to apply, and highly reliable. Unfortunately, they can primarily be collected after the task has already been completed and thus are hardly applicable for online assessment of cognitive load when aiming for timely HCI adaptation. *Behavioral* measurements rely on the analyses of differences in operators' interaction behavior with the system, such as mouse usage, click rates, etc. They potentially allow for online evaluation of fluctuations in cognitive load. However, behavioral measurements might be influenced by factors other than the task at hand such as attentional or motivational processes (Azcarraga & Suarez, 2013). Finally, *physiological* measurements (e.g., measurements of heart rate variability and electrodermal activity, electroencephalography (EEG), functional magnetic resonance imaging (fMRI), eye-tracking) relate physiological parameters to psychological constructs including cognitive load (FakhrHosseini & Jeon, 2019; Haapalainen, Kim, Forlizzi, & Dey, 2010; Johannsen, 1979; Liu, Walker, Friedman, Arrington, & Solovey, 2020). They seem very promising for the development of adaptive systems because they allow for continuous recording of the respective variables and thus online adaptation. However, they often require costly equipment and sophisticated methods of data analysis. Moreover, some physiological methods are hardly feasible in realistic HCI environments (for an overview see e.g., Ninaus et al., 2014) due to their immobility (e.g., fMRI) or high noise sensitivity (e.g., EEG).

1.2 Eye-Tracking

One physiological method which is gaining increasing popularity in this regard is eye-tracking. In particular the subtle realization of modern video-based eye-tracking systems (for an overview see: Hutton, 2019) makes this technique potentially promising for the commercial development of user-friendly adaptive HCIs. In this article, we focused on fixations, blinks, saccades, microsaccades, and pupil diameter as specific indices of participants' eye-fixation behavior because, as described in more detail below, evidence has shown that these features respond to fluctuations in cognitive load.

1.2.1 Fixations. Voluntarily controlled stable gazes lasting from 200 to 300 milliseconds to up to several seconds are called fixations. During these periods eyes stay relatively still, while the person processes information from the fixation area (Pouget, 2019). The relation between cognitive load and fixation time seems to depend on the task at hand. There is evidence that increased task complexity is associated with fewer but longer fixations (for reviews see: Clifton Jr et al., 2016; Rayner, 1998). For example, S. Chen et al. (2011) concluded that increased fixation duration and decreased fixation rate indicate increased attentional effort on a more demanding task. Similarly, De Rivecourt et al. (2008) found that increased task complexity is associated with longer fixations on the control instruments during simulated flight. Contrarily, however, Van Orden et al. (2001) observed fixation frequency to systematically increase with the visual complexity of a target classification task.

1.2.2 Saccades. Eye movements between two fixations that allow for exploration of the surroundings and attention control are called saccades. There is evidence that saccadic rate and length decrease with task difficulty (Nakayama et al., 2002).

1.2.3 Microsaccades. If our eyes would stay completely still during fixation, the visual image would gradually fade because neural response weakens with constant stimulation (Pouget, 2019). Microsaccades are small unintentional eye movements, which cover less than 1° of visual angle and prevent currently viewed visual information from fading. Evidence suggests that microsaccadic frequency increases with increasing visual complexity of the task at hand (Benedetto et al., 2011), whereas in non-visual

tasks microsaccadic rate seems to decrease and microsaccadic magnitude to increase with task difficulty (Gao et al., 2015; Siegenthaler et al., 2014).

1.2.4 Blinks. A commonly known function of blinking consists of keeping the eyeball moist and protecting it from physical damage. Besides that, in addition to microsaccades, blinking is also needed to prevent perceptual fading (Alexander & Martinez-Conde, 2019). Moreover, bursts of blinks seem to occur before and after periods of intense information processing (Siegle et al., 2008). Additionally, high blink rates were found associated with higher cognitive load (Nakayama et al., 2002).

1.2.5 Pupil dilation. This metric is most commonly considered in cognitive load research (for a general overview see: Andreassi, 2013). In states of high cognitive load, pupil diameter was repeatedly observed to increase proportionally both in visual and non-visual tasks (Fukuda et al., 2005; Klingner et al., 2011).

Taken together, eye-tracking seems a very promising technique well-suitable for assessing cognitive load online during HCI. Empirical evidence indicates that the eye-tracking measures listed above fluctuate dynamically over time (Siegle et al., 2008). Hence it might be advantageous to know the exact on- and off-set of each stimulus to be able to effectively differentiate between states of low and high cognitive load. While this premise is easy to achieve in a controlled laboratory setting, realistic HCI usually consists of a variety of interlocking tasks and stimuli that are impossible to analyze separately. Therefore, we chose a specific analytic approach. Instead of analyzing eye-tracking data for the entire time course of HCI, we focused on the analysis of most relevant time periods determined according to an established theoretical approach, allowing for better generalizability of these calculations to similar situations. In the context of time-critical interactions under severe time restraints, the time-based resource-sharing (TBRS) model briefly described below provided a suitable theoretical basis for assessing cognitive load.

1.3 Time-based resource-sharing model

The main idea proposed in the TBRS model by Barrouillet et al. (2004) is that, in addition to task complexity, cognitive load also strongly depends on available time. This is particularly relevant in time-critical situations. The model describes working memory as a core system of cognition consisting of two processes indispensable for the execution of a cognitive task: information storage and processing. According to the model, both components require attention, to switch between subtasks resulting in complex and time-critical interactions between them and eventually causing interruptions in the processing of subtasks. Based on these assumptions, TBRS predicts that cognitive load, and thus performance "depends on the proportion of time during which attention is captured in such a way that the storage of information is disturbed" (Barrouillet et al., 2004, p. 93). However, the authors acknowledge that it is not trivial to determine these time intervals.

In a recent study (Sevcenko et al., 2021), we addressed this challenge by applying the TBRS model to a serious game requiring time-critical action. Thereby, we identified characteristic patterns of player activity based on in-game log data and defined a behavioral metric as the ratio of the temporal duration of these patterns (for more details see Materials and Methods). Importantly, and observed that the proposed metric proved useful to predict participants' cognitive load. Moreover, because this metric was computed during the initial phase of interaction with the simulation it might offer the possibility for the potential development of smart HCI systems that can adapt to the operators' cognitive load early on during task execution and not only after task completion.

1.4 Present study

In this article, we present a novel theory-driven approach for targeted measurement of cognitive load using eye-tracking. In so doing, we used in-game log data interpreted in the light of the TBRS model to identify time periods relevant for analysis. We then related specific eye-tracking features acquired during these time periods to participants' performance and subjectively reported cognitive load, as well as task difficulty. To induce variance in cognitive load we used a serious game that

simulates time-critical emergencies requiring resource-management at different levels of difficulty.

In particular, we aimed at achieving two main goals. First, we considered the eye-tracking method to validate the theory-based approach suggested in Sevcenko et al. (2021). We expected that it should be possible to predict task difficulty, cognitive load, and performance based on eye-tracking features collected during predefined time intervals using TBRS Model. Second, with the help of eye-tracking data, we aimed at better understanding of what cognitive activities take place during specified time periods. For reasons of better readability, the detailed description of the analyzed time intervals, eye-tracking characteristics, and associated hypotheses are described in the following section.

2 Materials and Methods

The study was carried out as part of a larger project. Besides eye-tracking features described in this paper, it included measurements of behavioral in-game data, cardiac activity, galvanic skin response, and cortical hemodynamics, measured by functional near-infrared spectroscopy (Appel et al., 2019; Sevcenko et al., 2021).

2.1 Participants

In the following, we present data of 42 participants (31 females, 11 males) aged between 19 and 48 years (M = 24.3; SD = 5.4). Five participants were excluded from the analysis due to poor quality of their eye-tracking data. All participants spoke fluent German and were right-handed. They were recruited via an online database and compensated for their time expenditure. None of the participants reported neurologic, psychiatric, or cardiovascular disorders, and none of them were taking psychotropic medications. The study was approved by the local ethics committee and written informed consent was obtained prior to the experiment.

2.2 Task

Participants played an adapted version of the serious game (Emergency: Promotion Software GmbH, 1999), simulating time-critical emergencies. There were two different emergency scenarios with three different levels of difficulty each. During the game, participants had to coordinate different emergency personnel, such as emergency doctors, paramedics, and firefighters, as well as ambulances, fire- and ladder trucks to rescue victims and extinguish fires.

After familiarizing themselves with the task by playing a learning sequence, all participants completed two experimental scenarios: *Fire* and *Train Crash*. The learning sequence consisted of a short tutorial followed by a car accident scenario where participants had to free all victims from the crashed vehicles, provide first aid and then arrange their transport to hospital. The time limit for the training scenario was 5 minutes.

In the *Fire* scenario, participants had to extinguish a burning building block, rescue some residents from burning houses, provide first aid, and arrange their transport to hospital within a time limit of 7.5 minutes. The scenario *Train Crash* involved a train crashing into a building and causing a quick-spreading fire. The scenario required participants to free trapped passengers, provide first aid, and arrange their transport to hospital, as well as extinguish numerous fires. The time limit for each level of this scenario was 10 minutes.

Each scenario was presented at three difficulty levels: easy, medium, and hard, as defined by varying the number of tasks to be performed and the number of personnel to be coordinated (see Table 1). We expect that the increasing density actions required achieved in this way increased task demands in terms of planning, coordination, and prioritization, which should consequently lead to varying levels of cognitive load. Time pressure was additionally induced by setting time limits for levels.

2.3 Apparatus and experimental setup

The experiment was performed in a quiet room under constant light conditions (see Fig 1). The Emergency serious game was presented on a 16'' notebook driven at a screen resolution rate of 1920 x 1080. A conventional computer mouse was used as the only interaction device. Gaze data were recorded at 250 Hz using a SensoMotoric Instruments (SMI) RED250 eye tracker in combination with SMI Experiment Center 3.7.60 software installed on the same notebook. The eye tracker was calibrated using SMIs' integrated 9-point calibration procedure. The seating position of each participant was determined individually before the calibration of the eye tracker, without using a chin rest. The experiment began with the calibration followed by a baseline phase during which participants were asked to sit still and look at a fixation cross for 5 minutes to acquire baseline parameters of physiological measures. After that, participants completed an introductory learning sequence, followed by the two scenarios with their respective levels of difficulty presented in constant order (*Fire*: easy, medium, hard; *Train Crash*: easy, medium, hard). Subjective ratings of cognitive load experienced during the Emergency serious game were obtained intermittently after each level using the NASA-TLX questionnaire. The whole experiment lasted about one hour including training time.

Table 1 *Overview over the initial game parameters (Sevcenko et al., 2021).*

Scenario / Game Parameters	Difficulty		
	easy	medium	hard
Scenario: Fire			
Time limit (sec)	*450*	*450*	*450*
Tasks – total	*8+*	*13+*	*18+*
Victims	2	3	4
Fires	4+	7+	10+
Ladder Rescues	2	3	4
Resources –total	*9*	*12*	*15*
Doctors	1	2	2
Paramedics	1	2	2
Fire Fighters	4	4	6
Fire Trucks	2	3	4
Ladder Trucks	1	1	1
Scenario: Train Crash			
Time limit (sec)	*600*	*600*	*600*
Tasks – total	*20+*	*30+*	*40+*
Victims	10	15	20
Cars to cut	7	10	13
Fires	3+	5+	7+
Resources – total	*10*	*14*	*18*
Doctors	2	3	4
Paramedics	3	5	6
Fire Fighters	4	4	6
Fire Trucks	1	2	2

Note: The number of fires depended on players' performance and might grow. These cases are marked by the '+' sign.

2.4 Features

To estimate cognitive load, we used eye-tracking features known from the literature in this regard. These data were recorded during narrow time periods, which were calculated based on log data as described below. Hereafter, means of the respective eye-tracking measures for the respective time periods were associated with the level difficulty of the scenarios as well as participants' performance and their subjective ratings of cognitive load. In the following, we describe in more detail.

2.4.1 Difficulty and performance. For each level, a difficulty score was defined as the percentage of participants, who failed to complete the level within the predefined

time limit. Performance was reflected individually for each participant as the binary indicator of whether the level was completed successfully or not.

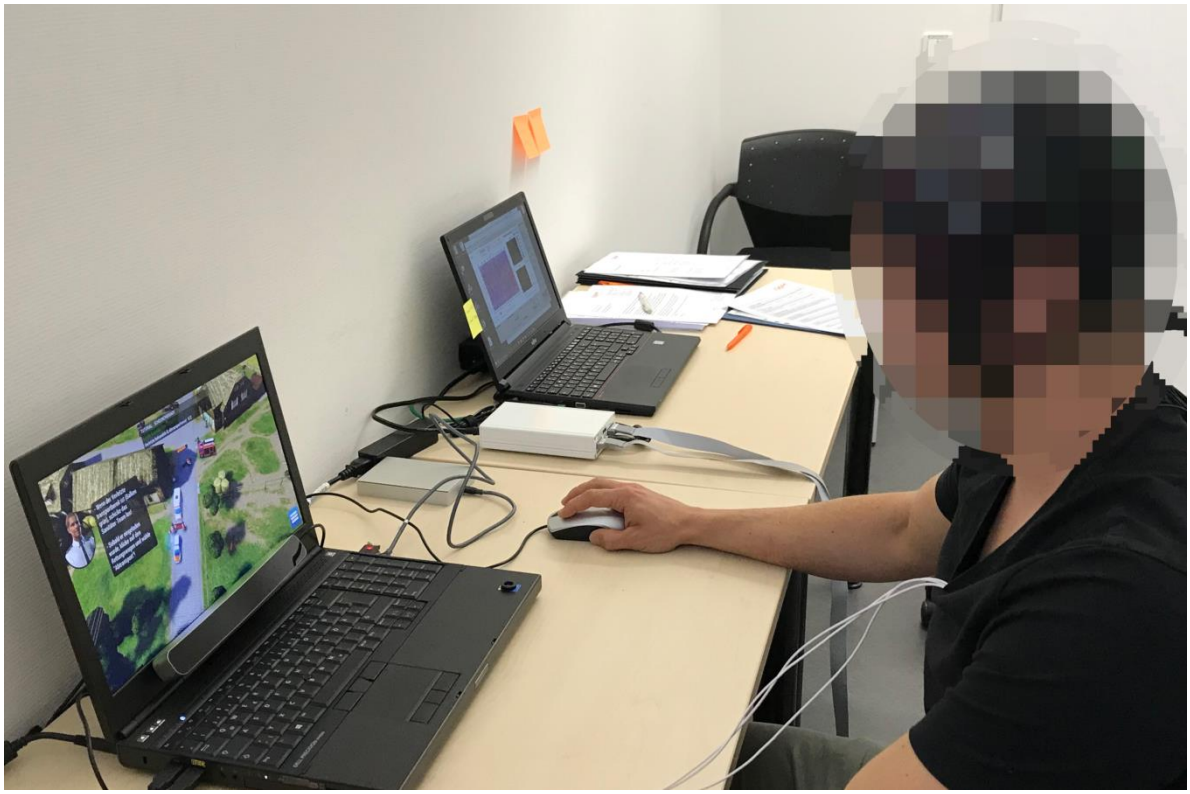


Fig. 1 Experimental setup.

2.4.2 Subjective rating of cognitive load. After each level, we asked participants to rate their subjectively experienced cognitive load by completing selected items of the NASA-TLX questionnaire (Hart & Staveland, 1988). The NASA-TLX consists of six items rated on a 21-level scale (0 to 100 points with steps of 5), and its' dimensions correspond to various theories distinguishing between physical, mental, and emotional demands imposed on the operator (Hart, 2006). In this study, we considered the three items addressing the mental facet of operators' load (i.e. mental demand, temporal demand, and effort) (Haerle et al., 2013; Temple et al., 1997).

2.4.3 Eye-Tracking features. Eye-tracking features were extracted using SMI Experiment Center 3.7.60 software. During preprocessing of pupillometric data, we removed all data points where pupil diameter was non-positive, because such artifacts typically indicate invalid data. We also did not consider data up to 100 ms immediately

before and after each blink, because during these periods the pupil is partially occluded by the eyelid or eyelashes and thus cannot be detected reliably (Kret & Sjak-Shie, 2019; Mathôt, Fabius, Van Heusden, & Van der Stigchel, 2018). Finally, we linearly interpolated small gaps of up to 50 ms to increase the amount of usable data. The rate of microsaccades was computed using the method proposed by Krejtz, Duchowski, Niedzielska, Biele, and Krejtz (2018). After preprocessing, we averaged the collected data over time, subtracted the respective baseline value, and z-standardized all features.

When analyzing gaze data, we focused on fixations, blinks, saccades, microsaccades, and pupil diameter because, as described above, evidence has shown that these features respond to variations in cognitive load. Previous evidence on the relationship between cognitive load and fixation rate is not entirely conclusive and seems to depend on the task at hand. For visual tasks it seems that the fixation rate increases with task difficulty (He et al., 2012; Van Orden et al., 2001), whereas non-visual task demands lead to a decrease in fixation rate (S. Chen et al., 2011; De Rivecourt et al., 2008). The same consideration is true for saccades and microsaccades. There is evidence that in non-visual tasks saccadic- and microsaccadic rate decreases with task difficulty (Gao et al., 2015; Nakayama et al., 2002; Siegenthaler et al., 2014), whereas visual complexity seems to increase saccadic and microsaccadic rates (Benedetto et al., 2011; He et al., 2012). The Emergency serious game was designed in such a manner that along with increasing the required number of actions, increasing levels of difficulty required better planning and coordination, which is a strong non-visual component. Therefore, we expected fixation frequency and saccadic and microsaccadic rates to increase in response to visual load and to decrease in response to non-visual cognitive load. Based on previous evidence we also expected pupil dilation (Fukuda et al., 2005; He et al., 2012; Klingner et al., 2011) and blinking rate (Nakayama et al., 2002; Siegle et al., 2008) to increase with increased difficulty.

2.4.4 Analyzed time periods. Eye-tracking features were collected during time periods related to the initial temporal action density decay (initial TADD) metric, which was proposed to predict cognitive load in time-critical situations where resources must be managed and tasks prioritized. According to this approach, such situations can be divided into a series of so-called action blocks consisting of active phases (burst), in which resources are managed, and waiting phases (idle), in which all resources are

occupied or unavailable and one must wait until a new task appears or a resource becomes available again (for a detailed example see Fig 2). In a recent study, the ratio of the length of the first detected burst (initial burst) to the length of the first action block (initial action block: occurred right at the beginning of each level), was shown to significantly predict performance (Sevcenko et al., 2021). In fact, it turned out that participants, who completed their tasks faster and therefore had to wait longer at the beginning of the level, were also significantly more likely to successfully complete the respective level. Thus, knowledge about participants' timing appears to be important for performance predictions, although this metric tells us nothing about cognitive processes actually occurring during this time.

In the present article, we determined initial burst and initial idle periods based on logged in-game activities for each participant. During the burst period, participants managed their emergency personnel, and after the last available personnel was assigned to a task, the idle interval started and lasted until the first personnel finished their tasks and were available again (for a detailed example see Fig 2). While participants' cognitive engagement during the initial burst can be estimated based on logged in-game actions, this method cannot be used for the initial idle period, because no actions are performed during this time. It is conceivable that some participants might use the initial idle for relaxation, which would be reflected in decreased cognitive load, whereas others might use this time for planning and visual screening of the scenery, which we expect to lead to an increased level of cognitive load as compared to the first group.

Based on these considerations, we expected that participants, who stay more active during the initial idle phase, will perform better and to be more likely to complete the level successfully. Regarding the initial burst phase, we expected all participants to work at their limit, that is, to have maximum cognitive load. For this reason, we did not expect any relationship between eye tracking data during the initial burst phase with task difficulty as well as their subjective assessment of cognitive load and their performance.

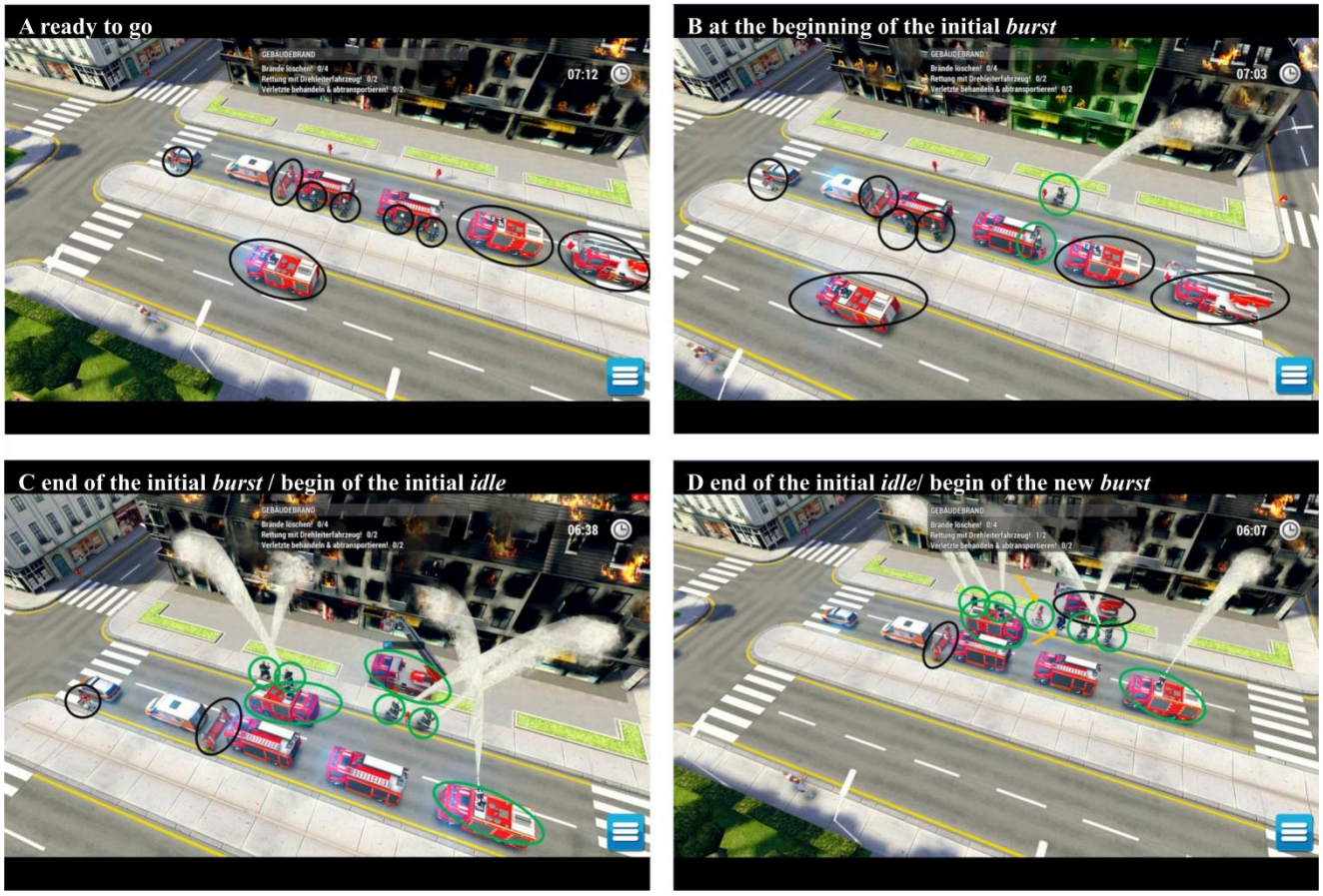


Fig. 2 Analyzed time periods using the *Fire* scenario as an example. Inactive task forces are marked with black circles, active - with green

A: At the beginning of the game level, all emergency personnel are ready and inactive.

B: The initial burst begins with the first task the player assigns and lasts until all available personnel are actively engaged. In this example, the emergency doctor and paramedics are not available for assignment because there are no injured people to be treated.

C: The initial burst ends as soon as the last available personnel are assigned, emergency doctor and paramedics cannot yet be assigned. This time also marks the beginning of the initial idle, in which the player must wait until some personnel become available again or until new tasks occur. In this example, all available personnel are already active and the player must wait until the person is rescued from the burning building, only then he can be treated by the doctor.

D: The initial idle ends when the first personnel are available again. In this example, the initial idle phase ends as soon as the rescued person appears lying on the road. At this moment, an emergency doctor becomes free and can be assigned to treat the patient; at the same time, the ladder truck also becomes free again and can be assigned to rescue the next person.

2.5 Statistical Analysis

We employed linear mixed-effect analyses using statistical software R (R Core Team, 2020) with the lme4 package (Bates et al., 2014). The p-values were obtained by likelihood ratio tests of the full model tested against a reduced model. Further model analyses were applied in case of a significant result, using the report package of D Makowski et al. (2020). Standardized parameters were obtained by fitting a model on a standardized version of the dataset.

3 Results

In this section, we present in detail associations between gaze data collected during initial burst, and initial idle time periods and (1) level difficulty, (2) participants' performance, (3) and subjective estimation of cognitive load. To ensure better readability and not overwhelm our readers with the vast amount of statistics, we have decided to only report significant results in detail.

3.1 Level Difficulty

First, we aimed at investigating whether level difficulty affected eye-tracing features during the initial burst and initial idle time periods. Therefore, we fitted linear mixed models for each combination of gaze features and time periods to predict gaze data. We considered difficulty as a fixed effect and added random intercepts for subjects.

3.1.1 Initial burst. During the initial burst period we found significantly negative association between level difficulty and *fixation frequency* ($\chi^2(1) = 7.26$, $p = .007$, *beta* = -0.20, 95% CI [-0.35, -0.06], $t(247) = -2.72$, $p < .01$; s*td. beta* = -0.05, 95% CI [-0.09, -0.01]) and *saccadic rate* ($\chi^2(1) = 3.92$, $p = .048$, *beta* = -0.15, 95% CI [-0.29, -2.18e-03], $t(247) = -1.99$, $p < .05$; *std. beta* = -0.04, 95% CI -0.08, -6.16e-04]), whereas effect on *pupil diameter* positive ($\chi^2(1) = 13.30$, $p < .001$, *beta* = 0.17, 95% CI [0.08, 0.26], $t(247) = 3.71$, $p < .001$; *std. beta* = 0.05, 95% CI [0.02, 0.08]). That is, during initial burst phase more challenging levels were significantly associated with fewer *fixations* and *saccades* as well as increased *pupil diameter* see Fig 3.

3.1.2 Initial idle. We found no significant association between gaze features and difficulty for any feature within the initial idle phase.

3.2 Performance

We fitted linear mixed-effect models for each combination of gaze features and time periods on the relationship between performance and gaze data. We added random intercepts for participants and considered performance as a fixed effect. We also considered the scenario as a fixed effect because we were interested in whether performance induces an effect in addition to the scenario.

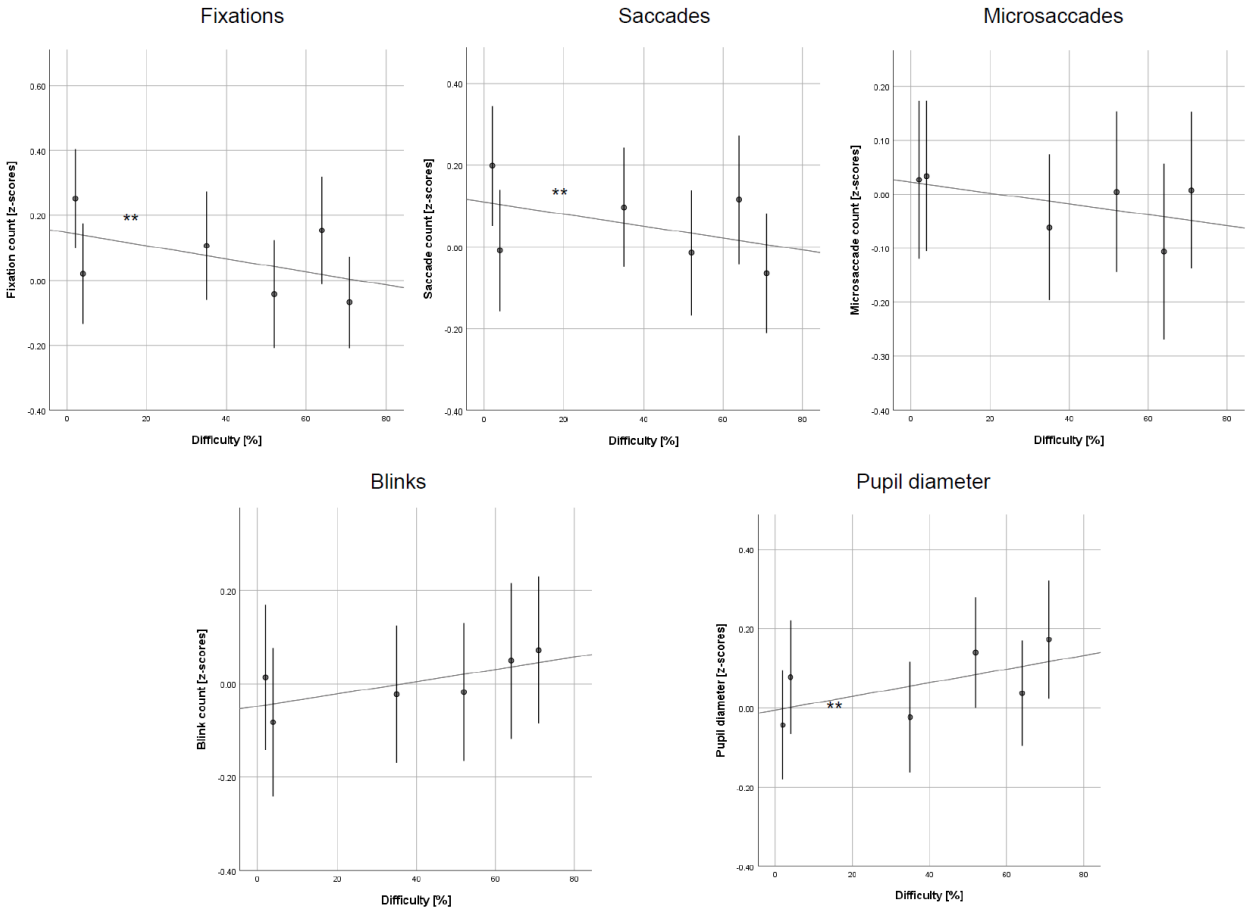


Fig. 3 Significant associations between level difficulty and eye-tracking features. Error bars depict +/- standard error.

3.2.1 Initial burst. During the burst phase we found a significantly negative effect of scenario on *fixation frequency* ($\chi^2(1) = 25.53$, $p < .001$, *beta* = -0.19, 95% CI [-0.27, -0.12], $t(247) = -5.03$, $p < .001$; *std. beta* = -0.19, 95% CI [-0.26, -0.12]). The same effect was found regarding *saccades* ($\chi^2(1) = 17.74$, $p < .001$, *beta* = -0.16, 95% CI [-0.23, -0.08], $t(247) = -4.11$, $p < .001$; *std. beta* = -0.16, 95% CI [-0.24, -0.09]). Whereas the

effect on *pupil diameter* was positive ($\chi^2(1) = 31.25$, $p < .001$, *beta* = 0.14, 95% CI [0.09, 0.19], $t(247) = 5.75$, $p < .01$; *std. beta* = 0.15, 95% CI [0.10, 0.21]). That is, while playing more challenging *Train Crash* scenario participants did significantly less *fixations* and *saccades*, while their *pupil diameter* was significantly increased compared to *Fire* scenario.

However, associations between gaze data and performance showed opposite, we found positive effect on *fixations* ($\chi^2(1) = 25.53$, $p < .001$, *beta* = 0.09, 95% CI [2.19e-03, 0.17], $t(247) = 2.01$, $p < .05$; *std. beta* = 0.09, 95% CI [2.14e-03, 0.17]), *saccades* ($\chi^2(1) = 4.34$, $p = .03$, *beta* = 0.09, 95% CI [5.89e-03, 0.18], $t(247) = 2.09$, $p < .05$; *std. beta* = 0.10, 95% CI [6.11e-03, 0.19]) and *microsaccades* ($\chi^2(1) = 6.22$, $p = .012$, *beta* = 0.012, 95% CI [0.03, 0.22], $t(247) = 2.51$, $p < .05$; *std. beta* = 0.13, 95% CI [0.03, 0.23]), along with negative effect on *blinks* ($\chi^2(1) = 5.47$, $p = .019$, *beta* = -0.10, 95% CI [-0.19, -0.02], $t(247) = -2.35$, $p < .05$; *std. beta* = -0.10, 95% CI [-0.19, -0.02]), meaning that participants who successfully completed the level showed significantly more *fixations*, *saccades* and *microsaccades*, but *blinked* significantly less often during the initial burst phase compared to unsuccessful participants (see Fig 4).

We found a significant positive effect of performance on *fixation frequency* ($\chi^2(1) = 29.53$, $p < .001$, *beta* = 0.09, 95% CI [2.19e-03, 0.17], $t(247) = 2.01$, $p < .05$; *std. beta* = 0.09, 95% CI [2.14e-03, 0.17]), *saccades* ($\chi^2(1) = 4.34$, $p = .037$, *beta* = 0.09, 95% CI [5.89e-03, 0.18], $t(247) = 2.09$, $p < .05$; *std. beta* = 0.10, 95% CI [6.11e-03, 0.19]), and *microsaccades* ($\chi^2(1) = 6.22$, $p = .013$, *beta* = 0.12, 95% CI [0.03, 0.22], $t(247) = 2.51$, $p < .05$; *std. beta* = 0.13, 95% CI [0.03, 0.23]), meaning that participants who failed the level also exhibited less *fixations*, *saccades* and *microsaccades* during the initial burst phase. In contrast, the effect on *blinks* was significantly negative, meaning that more successful participants *blinked* less during initial burst phase ($\chi^2(1) = 5.47$, p = .019, *beta* = -0.10, 95% CI [-0.19, -0.02], $t(247) = -2.35$, $p < .05$; *std. beta* = -0.10, 95% CI [-0.19, -0.02]).

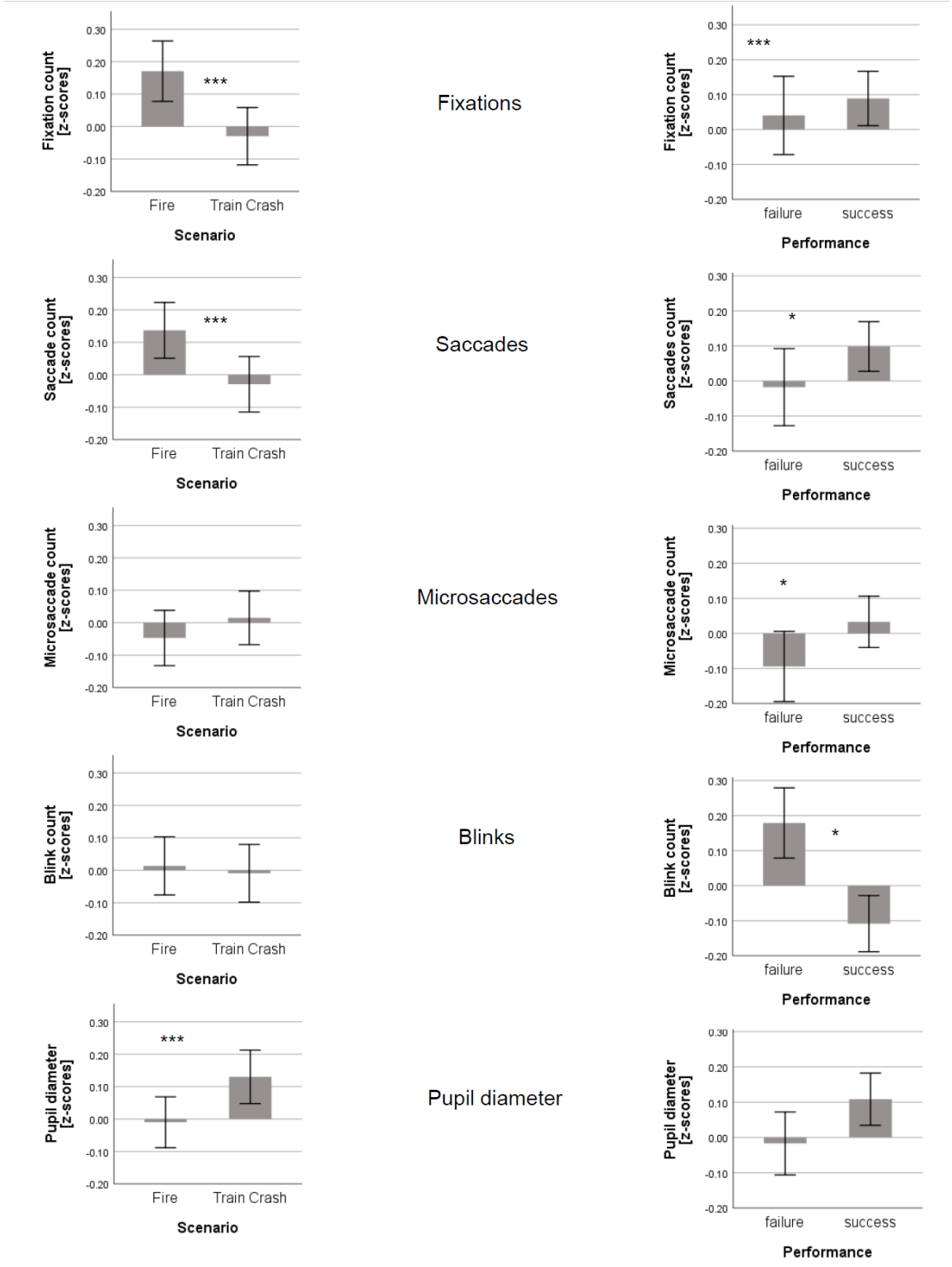


Fig. 4 Associations between gaze features, scenario, and performance. '***' = significance at 0.001 level; '**' = significance at the .01 level; '*' = is significance at.05 level. Error bars depict +/- standard error.

3.2.2 Initial idle. During initial idle only pupil diameter differed significantly between scenarios, whereas more challenging Train Crash scenario was associated with increased pupil diameter ($\chi^2(1) = 4.57$, $p = .032$, *beta* = 0.14, 95% CI [0.01, 0.27], $t(247) = 2.17$, $p < .05$; *std. beta* = 0.14, 95% CI [0.01, 0.28]). In contrast, performance was significantly positively associated with fixation frequency ($\chi^2(1) = 8.04$, $p = .005$, *beta* = 0.27, 95% CI [0.09, 0.46], $t(247) = 2.87$, $p < .01$; *std. beta* = 0.27, 95% CI [0.09, 0.45]), saccadic ($\chi^2(1) = 14.12$, $p < .001$, *beta* = 0.32, 95% CI [0.16, 0.49], $t(247) = 3.82$, $p < .001$; *std. beta* = 0.33, 95% CI [0.16, 0.50]), and microsaccadic rates ($\chi^2(1) = 14.79$, $p = .001$, *beta* = 0.37, 95% CI [0.18, 0.55], $t(247) = 3.92$, $p < .001$; *std. beta* = 0.37, 95% CI [0.19, 0.56]), meaning that participants who succeed the level showed significantly more fixations, saccades and microsaccades than participants who failed.

3.3 Subjective assessment of cognitive load

To investigate whether eye-tracking features are influenced by subjectively reported cognitive load we fitted linear mixed-effect models for each combination of the inspected NASA-TLX items and gaze features on the relationship between subjective cognitive load and gaze data which resulted in 15 models for each time period. We included random intercepts for participants, whereas scenario was considered a fixed factor.

For all considered NASA-TLX items we found a significant difference between scenarios, whereas the *Train Crash* scenario was perceived as more demanding than scenario *Fire* regarding *mental demand*, *time demand* and *effort* (*mental demand:* $\chi^2(1) = 19.19$, $p < .001$, *beta* = 6.39, 95% CI [3.60, 9.18], $t(247) = 4.48$, $p < .001$; *std. beta* = 0.28, 95% CI [0.16, 0.41]; *time demand:* $\chi^2(1) = 7.84$, $p = .005$, *beta* = 7.66, 95% CI [2.35, 12.97], $t(247) = 2.83$, $p < .01$; *std. beta* = 0.28, 95% CI [0.08, 0.47]; *effort:* $\chi^2(1) = 15.63$, $p < .001$, *beta* = 7.58, 95% CI [3.89, 11.27], $t(247) = 4.03$, $p < .001$; *std. beta* = 0.32, 95% CI [0.16, 0.47]).

3.3.1 Initial burst. Furthermore, in addition to the effect of scenario during the initial burst period we found a significant negative effect of *microsaccades* on all three items: *mental demand* ($\chi^2(1) = 19.19$, $p < .001$, *beta* = -4.27, 95% CI [-7.96, -0.59], $t(247)$

= -2.27, p < .05; *std. beta* = -0.18, 95% CI [-0.33, -0.02]), *time demand* ($\chi^2(1) = 7.84$, $p = .005$, *beta* = -5.17, 95% CI [-10.32, -0.03], t(247) = -1.97, $p < .05$; std. beta = -0.18, 95% CI [-0.35, -9.74e-04]), and *effort* ($\chi^2(1) = 15.63$, $p < .001$, *beta* = -6.71, 95% CI [-11.03, -2.40], t(247) = -3.05, $p < .01$; *std. beta* = -0.27, 95% CI [-0.44, -0.09]). That is, participant who reported to be more challenged in terms of mental demand, time demand and effort exhibited significantly less *microsaccades* than participants who rated their cognitive load lower.

3.3.2 Initial idle. Moreover we found significant negative effect of *microsaccades* on t*ime demand* ($\chi^2(1) = 7.84$, $p = .005$, *beta* = -5.27, 95% CI [-8.96, -1.57], $t(247) = -2.79$, $p < .01$; *std. beta* = -0.19, 95% CI [-0.32, -0.06]) and *effort* ($\chi^2(1) = 15.63$, $p < .001$, *beta* = -4.70, 95% CI [-7.43, -1.98], $t(247) = -3.38$, $p < .001$; *std. beta* = -0.19, 95% CI [-0.31, -0.08]). That is, participants who reported to be more challenged regarding *time demand* and *effort* did significantly less *microsaccades* during initial idle phase compared to less challenged participants. We found no significant effect regarding *mental demand*.

3.4 Additional Analyses

We hypothesized that more successful players should experience lower cognitive load, which should be reflected in their eye-tracking features - in particular, in lower ratings of subjective cognitive load among more successful participants. To evaluate this assumption, we fitted three linear mixed-effect models for each of the investigated NASA-TLX items to predict subjective ratings of cognitive load from performance. We included scenario and performance as fixed factors and participants as random intercepts in the models.

For all NASA-TLX items we found significantly positive association with scenario and significantly negative association with performance along with non-significant interaction effect (effect of scenario on mental demand: $\chi^2(1) = 19.19$, $p < .001$, *beta* = 5.37, 95% CI [2.91, 7.82], $t(247) = 4.29$, $p < .001$; *std. beta* = 0.24, 95% CI [0.13, 0.35]; effect of performance on mental demand: $\chi^2(1) = 58.26$, $p < .001$, *beta* = -11.71, 95% CI [-14.52, -8.89], $t(247) = -8.16$, $p < .001$; *std. beta* = -0.52, 95% CI [-0.64, -0.39]; effect of

scenario on time demand: $\chi^2(1) = 7.84$, $p = .005$, *beta* = 5.15, 95% CI [0.99, 9.32], $t(247) = 2.42$, $p < .05$; *std. beta* = 0.19, 95% CI [0.04, 0.34]; effect of performance on time demand: $\chi^2(1) = 11.60$, $p < .001$, *beta* = -28.71, 95% CI [-33.42, -24.00], $t(247) = -11.96$, $p < .001$; *std. beta* = -1.04, 95% CI [-1.21, -0.87]; effect of scenario on effort: $\chi^2(1) = 15.63$, $p < .001$, *beta* = 5.96, 95% CI [2.94, 8.98], $t(247) = 3.87$, $p < .001$; *std. beta* = 0.25, 95% CI [0.12, 0.38]; effect of performance on effort: $\chi^2(1) = 90.52$, $p < .001$, *beta* = -18.52, 95% CI [-21.97, -15.08], $t(247) = -10.54$, $p < .001$; *std. beta* = -0.78, 95% CI [-0.92, -0.63]). That is, subjective ratings of cognitive load were higher in a more challenging Train Crash scenario, and at the same time more successful players reported lower cognitive load.

4 Discussion

In this article, we presented a novel theory-driven approach to targeted eye-tracking based measurement of cognitive load in HCI. Our first goal was to investigate whether participants' cognitive load during a time-critical Emergency serious game can be estimated reliably based on gaze features collected at the beginning of a game session within initial burst and initial idle time periods. As a second goal, we aimed at deepening our understanding of what cognitive processes occur during initial burst and idle phases. To identify these time periods we used behavioral log data interpreted in the light of the TBRS model (Barrouillet et al., 2004) in line with a recent approach by Sevcenko et al. (2021).

In the following we will first describe in detail how the presented approach can be used for prediction of task difficulty, operators' performance, and subjectively perceived cognitive load, then we proceed to discuss which cognitive processes seem to happen during the initial idle phase and demonstrate correctness of the level construction of the used serious game. After that we present strengths and limitations of the study and briefly outline a possible direction of future research, followed by a general conclusion.

4.1 Difficulty, performance and cognitive load prediction

In general, the results of the present study substantiated our hypothesis that cognitive load might be predicted using eye-tracking features collected during time intervals related to initial TADD metric was confirmed. Indeed, we found significant associations between gaze features during the indicated time periods and difficulty, performance, and cognitive load, although some of these associations were unexpected. In line with our expectations, we found significant associations between performance and eye-tracking features during initial idle: successful participants did significantly more fixations, saccades, and microsaccades during this time period as compared to participants who failed the level. Additionally, we found a significant negative association between microsaccadic rates and subjective ratings of cognitive load for both initial burst and initial idle time periods. Other investigated gaze features showed no significant associations in this regard.

Contrary to our expectations, we found strong associations between eye-tracking features, difficulty, and performance when considering the initial burst time period. Because we assumed that all participants would play at their cognitive limit during the initial burst phase, we expected no effects during this time. Although gaze features were associated with task difficulty in the expected way, the observed association with performance was in the opposite direction. For instance, we observed that participants performed significantly fewer fixations during the more challenging levels, but this association was significantly less pronounced in more successful participants. The same pattern was also evident for other gaze features. Importantly, this finding seems sensible and might indicate that more successful players experienced lower cognitive load, which is reflected in their eye-tracking features. To test this assumption we conducted additional analyses (see Section 3.4 Additional Analyses) which showed exactly the same pattern of subjectively reported cognitive load ratings and thus further supported this account. As such, contrary to our expectations, the initial burst might be well suited to determine task difficulty and to-be-expected performance. One possible explanation for this finding is that time pressure, induced by the Emergency serious game, was not high enough to make all participants work on their cognitive limits. In this case, initial burst might be well suitable for measuring cognitive load in situations with low to medium time pressure, while for measuring cognitive load under high time pressure other options need to be identified. This hypothesis requires further investigation.

Taken together, our results support the idea that cognitive load and performance during HCI can be captured successfully based on gaze data collected during relatively narrow timeframes with the latter being inferred by a theory-driven approach and thus allowing for better generalizability to similar settings.

4.2 Cognitive processes during initial idle

Our second goal was to better understand what cognitive processes occur during initial idle, because no loggable in-game actions happen during this period. As expected, we found that successful participants performed significantly more fixations, saccades, and microsaccades during this time as compared to participants who failed the level. Based on this finding and previous evidence it seems that more successful participants

tend to use the initial idle time for more intensive visual exploration and monitoring of the game scene (Benedetto et al., 2011; He et al., 2012; Van Orden et al., 2001).

4.3 Manipulation Check

Last but not least, it is worth noting that our results substantiated that difficulty levels of the Emergency serious game were well constructed and effectively induced different levels of cognitive load. As expected, participants reported the *Train Crash* scenario to be cognitively more demanding than the *Fire* scenario. Likewise, the difficulty score which was calculated for each game level representing the percentage of participants who failed a level indicated that level difficulty increased as expected during the respective scenarios. Furthermore, we found significant negative associations between difficulty level and fixation as well as saccadic frequency, suggesting that game levels differed in non-visual cognitive demand components such as strategic planning (S. Chen et al., 2011; De Rivecourt et al., 2008; Nakayama et al., 2002).

4.4 Strengths and Limitations

Analytical approaches to HCIs are often based on data-driven probabilistic performance evaluations (Magerko et al., 2006; Spronck et al., 2006; Zook & Riedl, 2012), which sometimes seem insufficient for estimating operators' cognitive and emotional states. In such cases, (neuro-)physiological methods (for review see:Kivikangas et al., 2011) seem advantageous, although often relatively complex to acquire and evaluate. Another peculiarity of physiological methods is that most of them require physical contact with the operator and therefore can hardly be used for online commercial developments. Based on these considerations, we employed the eye-tracking method in this study, because modern video-based eye-tracking systems may not require physical contact and thus can be used in a very subtle way (for an overview see: Hutton, 2019).

Nevertheless, we think that the main strength of our approach is represented by our theoretically informed top-down development relying on an appropriate theoretical framework. In this way, we hope to foster generalizability of this method to similar situations, which might be less feasible in pure data-driven approaches. For this

purpose, we used the TBRS model (Barrouillet et al., 2004), which specifically emphasizes the role of time pressure in inducing cognitive load and therefore seems particularly suitable for predicting cognitive load during time-critical HCI. Moreover, the proposed method allows for early prediction of operator performance and can therefore be used for the development of interactive adaptive HCIs.

However, our approach is not free of limitations. First, the proposed method applies only to a relatively narrow family of situations or tasks with certain characteristics. Further research is needed to determine whether this approach can be applied or easily adapted to other contexts, e.g., interactions without time pressure. Second, as mentioned above, it is not clear whether different degrees of time pressure during HCI must be considered when using this method. Third, the recording frequency of the eye-tracker used for the study was 250 Hz, which might represent a technical limitation for instance by detecting microsaccades.

4.5 Conclusion

In this paper, we presented a novel theory-driven approach considering specific eye-tracking features to predict cognitive load during time-critical resource-managing situations in combination with TBRS theory. Eye-tracking data was collected during relatively narrow time windows at the beginning of the interaction with the simulation serious game and thus can be potentially used for real-time adaptation of human-computer interactions. Moreover, the detection of the time periods was based on log data and can be easily run in the background. Obtained results supported the proposed approach; eye-tracking features collected during the initial burst appeared to be well suited to predict performance and task difficulty. Fixations frequency, saccadic rate, and pupil diameter seem to be well suited to predict task difficulty during the initial burst phase. Fixation rate was the best indicator to predict performance during the initial burst. If an estimate of subjectively perceived cognitive load is required, the microsaccadic rate recorded during the initial action block might be a good option. These results illustrate how theoretic knowledge about the task structure may be used advantageously for the assessment of cognitive load. Although requiring further investigation in terms of reliability and generalizability, the presented approach seems promising for measuring

cognitive load in realistic time-critical HCI, considering adaptation to operators' mental needs.

References

Aasted, C. M., Yücel, M. A., Cooper, R. J., Dubb, J., Tsuzuki, D., Becerra, L., Petkov, M. P., Borsook, D., Dan, I., & Boas, D. A. (2015). Anatomical guidance for functional near-infrared spectroscopy: AtlasViewer tutorial. Neurophotonics, 2(2), 020801.

Alexander, R. G., & Martinez-Conde, S. (2019). Fixational Eye Movements. In Klein, C. & Ettinger, U. (Eds.), Eye Movement Research: An Introduction to its Scientific Foundations and Applications (pp. 73-115). Cham: Springer International Publishing.

Anderson, K. J. (1994). Impulsitivity, caffeine, and task difficulty: A within-subjects test of the Yerkes-Dodson law. Personality and Individual Differences, 16(6), 813-829.

Andreassi, J. L. (2013). Psychophysiology: Human behavior & physiological response (4 ed.). USA: Lawrence Erlbaub Associates.

Appel, T., Sevcenko, N., Wortha, F., Tsarava, K., Moeller, K., Ninaus, M., Kasneci, E., & Gerjets, P. (2019). Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures. In Gao, W., Meng, H. M. L., Turk, M., Fussell, S. R., Schuller, B., Song, Y., & Yu, K. (Eds.), 2019 International Conference on Multimodal Interaction (pp. 154-163). New York United States: Association for Computing Machinery.

Ayaz, H., Izzetoglu, M., Bunce, S., Heiman-Patterson, T., & Onaral, B. (2007). Detecting cognitive activity related hemodynamic signal for brain computer interface using functional near infrared spectroscopy. Paper presented at the Neural Engineering, 2007. CNE'07. 3rd International IEEE/EMBS Conference on.

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.

Azcarraga, J., & Suarez, M. T. (2013). Recognizing student emotions using brainwaves and mouse behavior data. International Journal of Distance Education Technologies (IJDET), 11(2), 1-15.

Babiloni, F. (2019). Mental Workload Monitoring: New Perspectives from Neuroscience. In Longo, L. & Leva, M. (Eds.), Human Mental Workload: Models and Applications. H-WORKLOAD 2019. Communications in Computer and Information Science (Vol. 1107, pp. 3-19). Cham: Springer.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. Journal of Experimental Psychology: General, 133(1), 83.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(3), 570.

Barrouillet, P., & Camos, V. (2015). Working memory: Loss and reconstruction (Roediger, H., Pomerantz, J., Baddeley, A. D., Bruce, V., & Grainger, J. Eds.). New York: Psychology Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Benedetto, S., Pedrotti, M., & Bridgeman, B. (2011). Microsaccades and exploratory saccades in a naturalistic environment. Journal of Eye Movement Research, 4(2). doi:10.16910/jemr.4.2.2

Berthold, A., & Jameson, A. (1999). Interpreting symptoms of cognitive load in speech input. In Kay, J. (Ed.), UM99 user modeling. CISM International Centre for Mechanical Sciences (Courses and Lectures) (Vol. 407, pp. 235-244). Vienna: Springer.

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., Lim, T., Ninaus, M., Ribeiro, C., & Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. Computers & Education, 94, 178-192.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. Educational psychologist, 38(1), 53-61.

Brünken, R., Seufert, T., & Paas, F. (2010). Measuring cognitive load.

Camos, V., Portrat, S., Vergauwe, E., & Barrouillet, P. (2007). The cognitive cost of executive functions. Paper presented at the Joint Meeting of the EPS and the Psychonomic Society, Edinburgh (Great-Britain).

Capon, N., Farley, J. U., & Hulbert, J. M. (1994). Strategic planning and financial performance: more evidence. Journal of Management Studies, 31(1), 105-110.

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. Journal of experimental child psychology, 33(3), 386-404.

Chang, C.-C., Warden, C. A., Liang, C., & Lin, G.-Y. (2018). Effects of digital game-based learning on achievement, flow and overall cognitive load. Australasian Journal of Educational Technology, 34(4), 155-167.

Charabati, S., Bracco, D., Mathieu, P., & Hemmerling, T. (2009). Comparison of four different display designs of a novel anaesthetic monitoring system, the 'integrated monitor of anaesthesia (IMA™)'. British journal of anaesthesia, 103(5), 670-677.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. Communications in Statistics—simulation and Computation®, 39(4), 860-864.

Chen, S., Epps, J., Ruiz, N., & Chen, F. (2011). Eye activity as a measure of human mental effort in HCI. Paper presented at the Proceedings of the 16th international conference on Intelligent user interfaces.

Clifton Jr, C., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. Journal of Memory and Language, 86, 1-19.

Cope, M., Delpy, D., Reynolds, E., Wray, S., Wyatt, J., & Van der Zee, P. (1988). Methods of quantitating cerebral near infrared spectroscopy data. In Oxygen Transport to Tissue X (pp. 183-189): Springer.

Csikszentmihalyi, M. (1975). Beyond boredom and anxiety (Csikszentmihalyi, I., Graef, R., Holcomb, J. H., Hendin, J., & MacAloon, J. Eds. 1 ed.). San Francisco, London: Jossey-Bass Publishers.

Cummings, M. L., & Nehme, C. E. (2009). Modeling the impact of workload in network centric supervisory control settings. Paper presented at the 2nd Annual Sustaining Performance Under Stress Symposium.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. Journal of Memory and Language, 19(4), 450.

De Rivecourt, M., Kuperus, M., Post, W., & Mulder, L. (2008). Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. Ergonomics, 51(9), 1295-1319. doi:10.1080/00140130802120267

Derosière, G., Mandrick, K., Dray, G., Ward, T. E., & Perrey, S. (2013). NIRS-measured prefrontal cortex activity in neuroergonomics: strengths and weaknesses. Frontiers in human neuroscience, 7, 583.

Eggemeier, F. T., Shingledecker, C. A., & Crabtree, M. S. (1985). Workload measurement in system design and evaluation. In Proceedings of the Human Factors Society Annual Meeting (Vol. 29, pp. 215-219). Los Angeles, CA: SAGE Publications Sage CA.

Eggemeier, F. T., Wilson, G. F., Kramer, A. F., & Damos, D. L. (1991). Workload assessment in multi-task environments. In Damos, D. L. (Ed.), Multiple-task performance (pp. 207-216). London, Washington, DC: Taylor & Francis.

Ehlis, A.-C., Herrmann, M., Wagener, A., & Fallgatter, A. (2005). Multi-channel near-infrared spectroscopy detects specific inferior-frontal activation during incongruent Stroop trials. Biological psychology, 69(3), 315-331.

FakhrHosseini, S. M., & Jeon, M. (2019). How do angry drivers respond to emotional music? A comprehensive perspective on assessing emotion. Journal on multimodal user interfaces, 13(2), 137-150. doi:https://doi.org/10.1007/s12193-019-00300-3

Fallgatter, A., Ehlis, A., Wagener, A., Michel, T., & Herrmann, M. (2004). Near-infrared spectroscopy in psychiatry. Der Nervenarzt, 75(9), 911.

Fan, J., & Smith, A. P. (2017a). The impact of workload and fatigue on performance. In Longo, L. & Leva, M. (Eds.), Human Mental Workload: Models and Applications. H-WORKLOAD 2017. Communications in Computer and Information Science (Vol. 726, pp. 90-105). Cham: Springer.

Fan, J., & Smith, A. P. (2017b). The impact of workload and fatigue on performance. Paper presented at the International symposium on human mental workload: Models and applications.

Fishburn, F. A., Norr, M. E., Medvedev, A. V., & Vaidya, C. J. (2014). Sensitivity of fNIRS to cognitive state and load. Frontiers in human neuroscience, 8, 76.

Freire, M., Serrano-Laguna, Á., Manero, B., Martínez-Ortiz, I., Moreno-Ger, P., & Fernández-Manjón, B. (2016). Game learning analytics: learning analytics for serious games. In Spector, M. J., Lockee, B. B., & Childress, M. D. (Eds.), Learning, design, and technology (pp. 1-29). Cham: Springer.

Fukuda, K., Stern, J. A., Brown, T. B., & Russo, M. B. (2005). Cognition, blinks, eye-movements, and pupillary movements during performance of a running memory task. Aviation, space, and environmental medicine, 76(7), C75-C85. Retrieved from https://www.ingentaconnect.com/content/asma/asem/2005/00000076/A00107s1/art00014

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. Advances in Methods and Practices in Psychological Science, 2(2), 156-168.

Gao, X., Yan, H., & Sun, H.-j. (2015). Modulation of microsaccade rate by task difficulty revealed through between-and within-trial comparisons. Journal of vision, 15(3), 3-3. doi:doi.org/10.1167/15.3.3

Geng, X., & Yamada, M. (2020). An augmented reality learning system for Japanese compound verbs: study of learning performance and cognitive load. Smart Learning Environments, 7(1), 1-19.

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., & Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. Frontiers in neuroscience, 8, 385.

Gerjets, P. H., & Hesse, F. W. (2004). When are powerful learning environments effective? The role of learner activities and of students' conceptions of educational technology. International Journal of Educational Research, 41(6), 445-465.

Gopher, D., & Braune, R. (1984). On the psychophysics of workload: Why bother with subjective measures? Human factors, 26(5), 519-532.

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. Paper presented at the Proceedings of the 12th ACM international conference on Ubiquitous computing, Copenhagen, Denmark. https://doi.org/10.1145/1864349.1864395

Haerle, S. K., Daly, M. J., Chan, H. H., Vescan, A., Kucharczyk, W., & Irish, J. C. (2013). Virtual surgical planning in endoscopic skull base surgery. The Laryngoscope, 123(12), 2935-2939.

Hancock, G., Hancock, P., & Janelle, C. (2012). The impact of emotions and predominant emotion regulation technique on driving performance. Work, 41(Supplement 1), 3608-3611.

Hancock, P. (1989). The effect of performance failure and task demand on the perception of mental workload. Applied Ergonomics, 20(3), 197-205.

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 50, pp. 904-908). Los Angeles, CA: Sage Publications CA.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Advances in psychology (Vol. 52, pp. 139-183): Elsevier.

He, X., Wang, L., Gao, X., & Chen, Y. (2012). The eye activity measurement of mental workload based on basic flight task. Paper presented at the IEEE 10th International Conference on Industrial Informatics.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2014). Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS. Frontiers in human neuroscience, 7, 935.

Hernández-Sabaté, A., Albarracín, L., Calvo, D., & Gorgorió, N. (2016). EyeMath: Identifying mathematics problem solving processes in a RTS video game. In Bottino, R., Jeuring, J., & Veltkamp, R. (Eds.), International Conference on Games and Learning Alliance. GALA 2016. Lecture Notes in Computer Science (Vol. 10056, pp. 50-59). Cham: Springer.

Herold, F., Wiegel, P., Scholkmann, F., & Müller, N. G. (2018). Applications of functional near-infrared spectroscopy (fNIRS) neuroimaging in exercise–cognition science: a systematic, methodology-focused review. Journal of clinical medicine, 7(12), 466.

Herrmann, M. J., Walter, A., Schreppel, T., Ehlis, A. C., Pauli, P., Lesch, K. P., & Fallgatter, A. (2007). D4 receptor gene variation modulates activation of prefrontal cortex during working memory. European Journal of Neuroscience, 26(10), 2713-2718.

Hoge, R. D., Atkinson, J., Gill, B., Crelier, G. R., Marrett, S., & Pike, G. B. (1999). Linear coupling between cerebral blood flow and oxygen consumption in activated human cortex. Proceedings of the National Academy of Sciences, 96(16), 9403-9408.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 50(3), 346-363.

Hutton, S. B. (2019). Eye Tracking Methodology. In Klein, C. & Ettinger, U. (Eds.), Eye Movement Research: An Introduction to its Scientific Foundations and Applications (pp. 277-308). Cham: Springer International Publishing.

Ikehara, C. S., & Crosby, M. E. (2005). Assessing cognitive load with physiological sensors. In Proceedings of the 38th annual hawaii international conference on system sciences (pp. 295a-295a). New York: IEEE.

Jobsis, F. F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. Science, 198(4323), 1264-1267.

Johannsen, G. (1979). Workload and workload measurement. In Moray, N. (Ed.), Mental Workload (Vol. 8, pp. 3-11). Boston: Springer.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. Educational psychology review, 19(4), 509-539.

Kerr, B. (1973). Processing demands during mental operations. Memory & Cognition, 1(4), 401-412.

Kiili, K., Lindstedt, A., & Ninaus, M. (2018, May 21-23, 2018). Exploring characteristics of students' emotions, flow and motivation in a math game competition. Paper presented at the GamiFIN Conference, Pori, Finland.

Kim, S. G., Rostrup, E., Larsson, H. B., Ogawa, S., & Paulson, O. B. (1999). Determination of relative CMRO2 from CBF and BOLD changes: significant increase of oxygen consumption rate during visual stimulation. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 41(6), 1152-1161.

Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., & Ravaja, N. (2011). A review of the use of psychophysiological methods in game research. journal of gaming & virtual worlds, 3(3), 181-199.

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

Kohlmorgen, J., Dornhege, G., Braun, M., Blankertz, B., Müller, K.-R., Curio, G., Hagemann, K., Bruns, A., Schrauf, M., & Kincses, W. (2007). Improving human performance in a real operating environment through real-time mental workload detection. In Dornhege, G., Millan, J. d. R., Hinterverger, T., McFarland, D. J., & Müller, K.-R. (Eds.), Toward Brain-Computer Interfacing (Vol. 409422, pp. 409-422). Cambridge, Massachusetts, London, England: MIT Press.

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. In Damos, D. (Ed.), Multiple-task performance (pp. 279-328). London, Washington, DC: Taylor & Francis.

Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., & Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. PloS one, 13(9), e0203629.

Kret, M. E., & Sjak-Shie, E. E. (2019). Preprocessing pupil size data: Guidelines and code. Behavior research methods, 51(3), 1336-1342.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). Package “emmeans”: Estimated Marginal Means, aka Least-Squares Means. Retrieved from https://cran.r-project.org

Lépine, R., Bernardin, S., & Barrouillet, P. (2005). Attention switching and working memory spans. European Journal of Cognitive Psychology, 17(3), 329-345.

Li, C., Gong, H., Gan, Z., & Luo, Q. (2005). Monitoring of prefrontal cortex activation during verbal n-back task with 24-channel functional NIRS imager. Paper presented at the Optics in Health Care and Biomedical Optics: Diagnostics and Treatment II.

Liefooghe, B., Barrouillet, P., Vandierendonck, A., & Camos, V. (2008). Working memory costs of task switching. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(3), 478.

Lim, Y. M., Ayesh, A., & Stacey, M. (2015). Using mouse and keyboard dynamics to detect cognitive stress during mental arithmetic. In Arai, K., Kapoor, S., & Bhatia, R. (Eds.), Intelligent Systems in Science and Information 2014. SAI 2014. Studies in Computational Intelligence (Vol. 591, pp. 335-350). Cham: Springer.

Linton, P., Jahns, D., & Chatelier, P. (1978). Operator workload assessment model: An evaluation of a VF/VA-V/STOL system. AGARD Methods to Assess Workloads 12 p(SEE N 78-31745 22-54).

Liu, R., Walker, E., Friedman, L., Arrington, C. M., & Solovey, E. T. (2020). fNIRS-based classification of mind-wandering with personalized window selection for multimodal learning interfaces. Journal on multimodal user interfaces, 1-16. doi:https://doi.org/10.1007/s12193-020-00325-z

Magerko, B., Stensrud, B. S., & Holt, L. S. (2006). Bringing the schoolhouse inside the box-a tool for engaging, individualized training. Retrieved from https://apps.dtic.mil/sti/pdfs/ADA481593.pdf

Magnusdottir, E. H., Borsky, M., Meier, M., Johannsdottir, K., & Gudnason, J. (2017). Monitoring cognitive workload using vocal tract and voice source features. Periodica Polytechnica Electrical Engineering and Computer Science, 61(4), 297-304.

Makowski, D., & Lüdecke, D. (2019). The report package for R: Ensuring the use of best practices for results reporting. CRAN. Retrieved from https://github.com/easystats/report

Makowski, D., Lüdecke, D., & Ben-Schachar, M. (2020). Automated reporting as a practical tool to improve reproducibility and methodological best practices adoption. J. Open Source Softw, 5, 2815.

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. Behavior research methods, 50(1), 94-106.

Meshkati, N. (1988). Toward development of a cohesive model of workload. In Advances in Psychology (Vol. 52, pp. 305-314): Elsevier.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. Annual review of neuroscience, 24(1), 167-202.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review, 63(2), 81.

Montani, F., Vandenberghe, C., Khedhaouria, A., & Courcy, F. (2020). Examining the inverted U-shaped relationship between workload and innovative work behavior: The role of work engagement and mindfulness. Human Relations, 73(1), 59-93.

Muratet, M., Torguet, P., & Jessel, J.-P. (2009). Learning programming with an RTS based Serious Game. In.

Nakayama, M., Takahashi, K., & Shimizu, Y. (2002). The act of task difficulty and eye-movement frequency for the'Oculo-motor indices'. In Proceedings of the 2002 symposium on Eye tracking research & applications (pp. 37-42): ACM Digital Library.

Nebel, S., & Ninaus, M. (2019). New perspectives on game-based assessment with process data and physiological signals. In Ifenthaler, D. & Kim, Y. (Eds.), Game-Based Assessment Revisited. Advances in Game-Based Learning (pp. 141-161). Cham: Springer.

Niederhauser, D. S., Reynolds, R. E., Salmen, D. J., & Skolmoski, P. (2000). The influence of cognitive load on learning from hypertext. Journal of educational computing research, 23(3), 237-255.

Nihashi, T., Ishigaki, T., Satake, H., Ito, S., Kaii, O., Mori, Y., Shimamoto, K., Fukushima, H., Suzuki, K., & Umakoshi, H. (2019). Monitoring of fatigue in radiologists during prolonged image interpretation using fNIRS. Japanese journal of radiology, 37(6), 437-448.

Ninaus, M., Kober, S. E., Friedrich, E. V., Dunwell, I., De Freitas, S., Arnab, S., Ott, M., Kravcik, M., Lim, T., Louchart, S., Bellotti, F., Hannemann, A., Thin, A. G., Berta, R., Wood, G., & Neuper, C. (2014). Neurophysiological methods for monitoring brain activity in serious games and virtual environments: a review. International Journal of Technology Enhanced Learning, 6(1), 78-103. doi:https://doi.org/10.1504/IJTEL.2014.060022

Ninaus, M., Witte, M., Kober, S. E., Friedrich, E. V., Kurzmann, J., Hartsuiker, E., Neuper, C., & Wood, G. (2013). Neurofeedback and serious games. In T. M. Connolly, E., Boyle, T., Hainey, G., Baxter, P., & Moreno-ger (Eds.), Psychology, Pedagogy, and Assessment in Serious Games (Vol. i, pp. 82-110). USA: IGI Global.

O'Donnell, R., & Eggemeier, F. (1986). Workload assessment methodology. Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance. KR Boff, L. Kaufman and JP Thomas. In: John Wiley and Sons, Inc.

Orru, G., & Longo, L. (2019). The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and Germane loads: a review. In Longo, L. & Leva, M. (Eds.), Human Mental Workload: Models and Applications. H-WORKLOAD 2018. Communications in Computer and Information Science (pp. 23-48). Cham: Springer.

Oviatt, S. (2006). Human-centered design meets cognitive load theory: designing interfaces that help people think. Paper presented at the Proceedings of the 14th ACM international conference on Multimedia.

Paas, F. G., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. Educational psychology review, 6(4), 351-371.

Parasuraman, R., & Rizzo, M. (2006). Introduction to Neuroergonomics. In.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. Journal of educational psychology, 102(3), 531.

Portrat, S. (2008). Working memory and executive functions: The Time-Based Resource-Sharin account Université de Bourgogne. Dijon.

Pouget, P. (2019). Introduction to the Study of Eye Movements. In Klein, C. & Ettinger, U. (Eds.), Eye Movement Research: An Introduction to its Scientific Foundations and Applications (pp. 3-10). Cham: Springer International Publishing.

Promotion Software GmbH. (1999). World of Emergency. Retrieved from https://www.world-of-emergency.com/?lang=en

R Core Team. (2020). R: A Language and Environment for Statistical Computing. Retrieved from https://www.R-project.org/

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological bulletin, 124(3), 372.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In Advances in psychology (Vol. 52, pp. 185-218): Elsevier.

Richards, K. C., Enderlin, C. A., Beck, C., McSweeney, J. C., Jones, T. C., & Roberson, P. K. (2007). Tailored biobehavioral interventions: a literature review and synthesis. Research and theory for nursing practice, 21(4), 271-285.

Ruck, H., & Stottan, T. (2014). Interactive Steering Wheel for Optimal Operation. ATZ worldwide, 116(5), 40-45.

Ruiz, N., Liu, G., Yin, B., Farrow, D., & Chen, F. (2010). Teaching athletes cognitive skills: detecting cognitive load in speech input. Proceedings of HCI 2010 24, 484-488.

Saddler, B., Moran, S., Graham, S., & Harris, K. R. (2004). Preventing writing difficulties: The effects of planning strategy instruction on the writing performance of struggling writers. Exceptionality, 12(1), 3-17.

Salomon, G. (1984). Television is" easy" and print is" tough": The differential investment of mental effort in learning as a function of perceptions and attributions. Journal of educational psychology, 76(4), 647.

Scerbo, M. W. (1996). Theoretical perspectives on adaptive automation. In Parasuraman, R. & Mouloua, M. (Eds.), Automation and Human Performance: Theory and Applications (pp. 37-64): CRC Press.

Sevcenko, N., Ninaus, M., Wortha, F., Moeller, K., & Gerjets, P. (2021). Measuring cognitive load using in-game metrics of a serious simulation game. Frontiers in Psychology, 12, 906. doi:https://doi.org/10.3389/fpsyg.2021.572437

Sheridan, T. B., & Simpson, R. (1979). Toward the definition and measurement of the mental workload of transport pilots. Retrieved from https://dspace.mit.edu/handle/1721.1/67913

Shute, V. J., & Kim, Y. J. (2014). Formative and stealth assessment. In Spector, J., Merrill, M., Elen, J., & Bishop, M. (Eds.), Handbook of research on educational communications and technology (pp. 311-321). New York, NY: Springer New York.

Siegenthaler, E., Costela, F. M., McCamy, M. B., Di Stasi, L. L., Otero-Millan, J., Sonderegger, A., Groner, R., Macknik, S., & Martinez-Conde, S. (2014). Task difficulty in mental arithmetic affects microsaccadic rates and magnitudes. European Journal of Neuroscience, 39(2), 287-294. doi:10.1111/ejn.12395

Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses. Psychophysiology, 45(5), 679-687. doi:10.1111/j.1469-8986.2008.00681.x

Simons, A., Wohlgenannt, I., Weinmann, M., & Fleischer, S. (2020). Good gamers, good managers? A proof-of-concept study with Sid Meier's Civilization. Review of Managerial Science, 1-34. doi:10.1007/s11846-020-00378-0

Smith-Jackson, T. L., & Klein, K. W. (2009). Open-plan offices: Task performance and mental workload. Journal of Environmental Psychology, 29(2), 279-289.

Smith, E. E., & Jonides, J. (1997). Working memory: A view from neuroimaging. Cognitive psychology, 33(1), 5-42.

Spronck, P., Ponsen, M., Sprinkhuizen-Kuyper, I., & Postma, E. (2006). Adaptive game AI with dynamic scripting. Machine Learning, 63(3), 217-248.

Statistisches Bundesamt. (2017). Verkehrsunfälle. Unfälle von Güterkraftfahrzeugen im Straßenverkehr. Retrieved from https://www.destatis.de/DE/Publikationen/Thematisch/TransportVerkehr/Verkehrsunfaelle/UnfaelleGueterkraftfahrzeuge5462410167004.pdf?__blob=publicationFile

Strangman, G., Boas, D. A., & Sutton, J. P. (2002). Non-invasive neuroimaging using near-infrared light. Biological psychiatry, 52(7), 679-693.

Strangman, G., Goldstein, R., Rauch, S. L., & Stein, J. (2006). Near-infrared spectroscopy and imaging for investigating stroke rehabilitation: test-retest reliability and review of the literature. Archives of physical medicine and rehabilitation, 87(12), 12-19.

Strangman, G. E., Li, Z., & Zhang, Q. (2013). Depth sensitivity and source-detector separations for near infrared spectroscopy based on the Colin27 brain template. PloS one, 8(8), e66319.

Susi, T., Johannesson, M., & Backlund, P. (2007). Serious games: An overview. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-1279

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. Educational psychology review, 10(3), 251-296.

Temple, J. G., Dember, W. N., Warm, J. S., Jones, K. S., & LaGrange, C. M. (1997). The effects of caffeine on performance and stress in an abbreviated vigilance task. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Thier, P. (2006). Die funktionelle Architektur des präfrontalen Kortex. In Neuropsychologie (pp. 471-478): Springer.

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T.-P. (2001). Eye activity correlates of workload during a visuospatial memory task. Human factors, 43(1), 111-121. doi:10.1518/001872001775992570

Van Rossum, G., & Drake, F. L. (2009). PYTHON 2.6 Reference Manual.

Veltman, J., & Jansen, C. (2005). The role of operator state assessment in adaptive automation. Retrieved from https://apps.dtic.mil/sti/citations/ADA455055

Vygotsky, L. S. (1980). Mind in society: The development of higher psychological processes (Cole, M., John-Steiner, V., Scribner, S., & Souberman, E. Eds.). Cambridge, London: Harvard university press.

Walter, C., Rosenstiel, W., Bogdan, M., Gerjets, P., & Spüler, M. (2017). Online EEG-Based Workload Adaptation of an Arithmetic Learning Environment. Frontiers in human neuroscience, 11(286). doi:10.3389/fnhum.2017.00286

Wang, Y.-R., & Gibson Jr, G. E. (2010). A study of preproject planning and project success using ANNs and regression models. Automation in Construction, 19(3), 341-346.

Watters, P. A., Martin, F., & Schreter, Z. (1997). Caffeine and cognitive performance: The nonlinear Yerkes–Dodson law. Human Psychopharmacology: Clinical and Experimental, 12(3), 249-257.

Welford, A. (1978). Mental work-load as a function of demand, capacity, strategy and skill. Ergonomics, 21(3), 151-167.

Xu, X., Deng, Z.-Y., Huang, Q., Zhang, W.-X., Qi, C.-z., & Huang, J.-A. (2017). Prefrontal cortex-mediated executive function as assessed by Stroop task performance associates with weight loss among overweight and obese adolescents and young adults. Behavioural brain research, 321, 240-248.

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.

Yuksel, B. F., Oleson, K. B., Harrison, L., Peck, E. M., Afergan, D., Chang, R., & Jacob, R. J. (2016). Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In Proceedings of the 2016 CHI conference on human factors in computing systems (pp. 5372-5384).

Yurko, Y. Y., Scerbo, M. W., Prabhu, A. S., Acker, C. E., & Stefanidis, D. (2010). Higher mental workload is associated with poorer laparoscopic performance as measured by the NASA-TLX tool. Simulation in healthcare, 5(5), 267-271. doi:10.1097/SIH.0b013e3181e3f329

Zhou, T., Cha, J. S., Gonzalez, G., Wachs, J. P., Sundaram, C. P., & Yu, D. (2020). Multimodal Physiological Signals for Workload Prediction in Robot-assisted Surgery. ACM Transactions on Human-Robot Interaction (THRI), 9(2), 1-26.

Zohaib, M. (2018). Dynamic difficulty adjustment (dda) in computer games: A review. Advances in Human-Computer Interaction, 2018.

Zook, A. E., & Riedl, M. O. (2012). A temporal data-driven player model for dynamic difficulty adjustment. Paper presented at the Eighth Artificial Intelligence and Interactive Digital Entertainment Conference.

5 General Discussion

5.1 Summary and discussion of general findings

Recent developments continue to move towards increasing complexity in HMIs, making the operators' task more complex due to the increasing number of available options and actions (Stapel et al., 2019). In this context, determining the operators' cognitive load seems crucial for the timely prediction of an emerging performance breakdown, whereby such information might be useful for a timely adaptation of the HMI interface with the aim of keeping the operator within the optimal activation range, thus achieving the best possible performance. Keeping operators' performance at an optimal level becomes especially important in time-critical situations, imposing time pressure. The main objective of this dissertation was to provide a contribution to the development of adaptive HMI systems for real-world applications capable of recognizing and adapting to the operators' cognitive need. Accordingly, a novel method for measuring cognitive load in situations involving resource management under time pressure was elaborated, presented, and evaluated in the process. Thereby, the following requirements were set: (1) to obtain a method that might be easily generalizable to similar situations, it should be based on an appropriate theoretical framework; (2) to be adequate for use in adaptive HMI systems, it should be able to track variations in cognitive load over time while not impairing performance in the main task; (3) and finally, to be suitable for real-world and commercial developments, the measurement techniques used in the method should be as discrete and unobtrusive as possible, with the computational algorithms being as simple to execute as possible to enable its usage in rather small devices with low computing power.

In a sequence of three studies, the established theoretical model of WM was applied to a realistic HMI setting, represented by a serious game. In the first study, three variations of the *TADD* metric were derived from the TBRS model of Barrouillet et al. (2004) and evaluated using behavioral data. The second and the third studies aimed to further evaluate and investigate the concept with (neuro-) physiological measurement methods. In the following, the aforementioned findings are discussed in more detail.

It is worth noting that the first study demonstrated the validity of the level construction of the customized version of the Emergency (Promotion Software GmbH,

1999) serious game. This version of the game was used throughout the project to induce different levels of cognitive load in participants; therefore, a valid level construction was an important prerequisite for all further analyses. Although all of the three behavioral metrics proposed in the first study (i.e. *initial TADD*, *mean TADD*, and *normalized gaming time*) were found to be significantly associated with cognitive load, the *initial TADD* metric appeared to be the most promising for use in adaptive scenarios given that it was calculated very early in the course of the level and thus qualified for near-real-time adaptation. The validity of the metric was supported in statistical analyses, particularly when used under severe time constraints. This was interpreted as a possible indication of an existing floor effect, indicating that this metric may be best suitable in situations eliciting phases of maximum cognitive load. Another important result was that the beginning of the entire interaction appeared crucial for the final outcome. This finding is consistent with empirical evidence, suggesting that better planning in early stages of various tasks is associated with better performance (Capon, Farley, & Hulbert, 1994; Liu, Walker, Friedman, Arrington, & Solovey, 2020; Saddler, Moran, Graham, & Harris, 2004) and thus might represent a valid global phenomenon in different contexts. At the same time, the question arises whether this structure is typical for situations involving resource management under time pressure and, if not, how deviations from this structural pattern should be dealt with when applying the proposed method.

The second study (see Table 2) further supported this finding by demonstrating that cortical hemodynamics assessed by fNIRS during the first minutes of the interaction were more informative in detecting cognitive load than during later time periods. This result supports the idea that a theoretical knowledge of task structure and an appropriate model of the cognitive activities ongoing during the task processing might be used advantageously to identify cognitive load, whereas focusing on specific theoretically-derived time intervals might be superior to averaging over random time slots, which carries the risk of overlooking existing difference features. Furthermore, against expectations, it was found that hemodynamic activation during the *initial burst* phase was significantly associated with the difficulty level. This result was crucial because the important assumption made in the definition of the *initial TADD* metric was that all participants would be working at their cognitive limits during the *initial burst* and thus no association between cognitive load and difficulty or the final performance was expected. At the same time, this finding might help to understand why the *initial TADD* metric showed better predictive power when applied under stronger time pressure, as it

emerged in the first study. Differences observed in cognitive load during the *initial burst* suggested that the induced time pressure was not sufficient to force all participants to their limits, thus confirming the aforementioned hypothesis of existing floor effect. Finally, it was demonstrated that participants who showed stronger hemodynamic cortical activation within the *initial idle* phase perceived less cognitive load at the end of the level, which might indicate that more successful players tended to use this time for active monitoring and planning (P. Hancock, 1989), whereas more challenged participants might use this time to relax (Nihashi et al., 2019).

Table 2. *Graphical summary of significant effects obtained in study 2.*

			NASA-TLX		
	level difficulty	perfor-mance	mental demand	time demand	effort
initial burst					
DLPFC-left	+				
DLPFC-right	+	-		+	
initial idle (t0-t20)					
DLPFC-left			-	-	-
DLPFC-right					-
t20-t40					
DLPFC-left					
DLPFC-right	-		-		
t40-t60					
DLPFC-left					
DLPFC-right					

Note: Significant results are marked with green '+' for positive and red '-' for negative effects. Level difficulty: percentage of participants who did not complete the level. Performance (related to participant): binary indicator of whether the level was completed successfully or not.

The results from the third study (see Table 3) were largely consistent with previous findings. Eye-tracking features acquired during the *initial burst* phase were significantly associated with the level difficulty, whereas it was found that more successful participants tended to use the *initial idle* period for more active visual

exploration compared to participants who failed the level. Fixations frequency, saccadic rate, and pupil diameter appeared to be well suited to predict task difficulty during the *initial burst* phase. Fixation rate was the best indicator to predict performance during the *initial burst*. If an estimate of subjectively perceived cognitive load is required, the microsaccadic rate recorded during the *initial action block* might be a good option.

Table 3. *Graphical summary of significant effects obtained in study 3.*

			NASA-TLX		
	level difficulty	perfor-mance	mental demand	time demand	effort
initial burst					
fixations	-	+			
blinks		-			
saccades	-	+			
pupil diameter	+				
microsaccades		+	-	-	-
initial idle					
fixations		+			
blinks					
saccades		+			
pupil diameter					
microsaccades		+		-	-

Note: Significant results are marked with green '+' for positive and red '-' for negative effects. *Level difficulty*: percentage of participants who did not complete the level. *Performance* (related to participant): binary indicator of whether the level was completed successfully or not.

Altogether the three studies of this dissertation have indicated that the TBRS model can be used for assessing cognitive load in a realistic setting with inherent time pressure (as proposed in Section 1.4). The three studies have provided first empirical evidence of the value of the presented approach and opened a new perspective for the development of theory-based metrics for measuring cognitive load in time pressure situations. In particular, three main outcomes can be emphasized. First, in the course of the realistic time-critical interaction simulated by the Emergency serious game, a predicted occurrence of periods of high and low activity (*burst* and *idle*) was detected. Second, the relation of these time periods (*TADD*) was significantly associated with level

difficulty, performance, and subjective cognitive load (NASA-TLX). Third, fNIRS and eye-tracking data collected during the initial burst and initial idle periods were significantly associated with level difficulty, performance, and subjective cognitive load (NASA-TLX). Fourth, the *initial TADD* metric appears to be better suited for use in high-pressure scenarios, whereas fNIRS and eye-tracking data collected during the *initial burst* may be best applicable in more relaxed situations.

To provide the full picture of the project, Table 4 represents an overview of all metrics and characteristics analyzed over the course of the three studies. Depending on the complexity of the scenario and the resulting time pressure, different results were obtained regarding the *initial TADD* metric. Although other features were not tested separately for different scenarios, this suggests that different measures and features may be more appropriate under different time pressure conditions, which should be explored in future studies. For this reason, although this distinction is only relevant for the *initial TADD* metric, the lower horizontal header of Table 4 distinguishes between different time pressure conditions, with the Crash train representing a high time pressure scenario and Fire representing a low time pressure scenario. The vertical header is divided into three measurement methods used in the project: Behavioral analysis, fNIRS, and eye-tracking. For each method, the associated time intervals used for the calculation of the respective features/metrics are listed in the right part of the vertical header. All presented metrics were collected at the beginning of each level during three time periods: *initial burst* (approx. the first 40 sec), *initial idle* (approx. the following 10 sec), and *initial action block* (consisting of *initial burst* and *initial idle* together, approx. the first 50 sec). Finally, the body of Table 4 presents significant associations between the analyzed metrics/features and the difficulty of each level within the scenario, the performance (whether or not the level was successfully completed), and the subjective estimation of cognitive load (NASA-TLX) after completion of the respective level. The significance level is indicated by different fonts.

Table 4. *Summary of three studies of the dissertation with preliminary suggestions of which measurement features might be better suitable for which situation.*

Measure ↓ time period	level difficulty	performance	subjective CL
	Time pressure (scenario complexity)		

		low	high	low	high	low	high
Behavioral	initial action block 50 sec	x	x		**initial TADD (-)**	**initial TADD (+)**	
fNIRS	initial burst 40 sec	*DLPFC-left (+)* *DLPFC-right (+)*		*DLPFC-right (-)*		*DLPFC-right (+)*	
	initial idle 10 sec					*DLPFC-left (-)* DLPFC-right (-)	
Eye-tracking	initial burst 40 sec	fixations (-), *saccades (-)*, **pupil diameter (+)**		**fixations (+)**, *blinks (-)*, *saccades (+)*, *microsaccades (+)*		**microsaccades (-)**	
	initial idle 10 sec			fixations (+), **saccades (+)**, **microsaccades (+)**		**microsaccades (-)**	

Note: The significance level is marked as follows. 0.5: italic, 0.1: standard, 0.001: bold; '+': indicates positive association, '-': indicates negative association; 'x': indicates associations that have not been tested; CL: cognitive load.

According to the discriminant analyses with Leave-One-Subject-Out Cross-Validation, the initial TADD metric demonstrated a 64.53% accuracy in predicting performance in the more complex Train Crash scenario, where the models' outcome was significantly above the score of the random model and at the same time did not significantly differ from the prediction based on NASA-TLX scores (for this reason an additional model was build based on NASA-TLX scores). At the same time, this pattern was not found for less complex Fire scenario, where the prediction of performance based on the *initial TADD* did not significantly differ from models predicting randomly permuted outcomes. This result indicates that the assumption that all participants would play at their maximum speed and thus achieve the same level of cognitive load regardless of the level of difficulty and performance might be incorrect. It is possible that participants acted more relaxed in the less time pressure Fire scenario because they could afford to win the level even if they were not working at their maximum speed, which affected the predictive power of the *initial TADD* metric. The results of the second

and third studies support this suspicion, showing correlations between level difficulty and cortical activation and eye-tracking features, thus suggesting that participants were not working at their cognitive limits at this time. At the same time, increased cognitive activation during this time as assessed by fNIRS and eye-tracking features was associated with decreased performance. This suggests that acting too fast may also be counterproductive, causing subjects to seemingly overload their cognitive resources, leading to deterioration in performance. Another important piece of information, which can be seen in Table 4, is that eye-tracking features showed stronger associations with level difficulty, performance, and subjective cognitive load than fNIRS features, while at the same time showing the same pattern as the fNIRS associations, suggesting that eye-tracking could be used in place of fNIRS in further studies.

To summarize all information presented in Table 4 and the last section. The results of this dissertation indicate that (1) the *initial TADD* metric might be better suited for more complex scenarios, inducing more time pressure, which should be investigated in future studies; (2) Eye-tracking features during *initial burst* and *initial idle* periods might be better used to access cognitive load in less demanding scenarios, although also this assumption needs detailed investigation in the future; (3) fixation count, saccadic and microsaccadic rates and pupil diameter showed stronger associations with level difficulty, performance, and subjective cognitive load (NASA-TLX) compared to blink count and should therefore be preferred in further investigations; (4) FNIRS measurements might be replaced by eye-tracking measurements because eye-tracking features showed stronger associations with level difficulty, performance, and subjective cognitive load while maintaining the same pattern; (5) Adaptive systems based on the proposed metrics/features could prevent performance degradation by detecting user actions that are too fast and prompting them to slow down. In the next section, the value and scientific contribution of the proposed approach will be discussed (Section 5.2), followed by the strength and limitations (Section 5.3) as well as implications for future research (Section 5.4).

5.2 Scientific contribution

As described earlier (Sections 1.1, 1.2), interaction systems that are able to adapt to the cognitive load are increasingly becoming the focus of modern research and seem very promising in terms of optimizing performance in a variety of realistic contexts. At the

same time, adaptation to cognitive load appears to be quite feasible, although a reliable measure for online and near-real-time assessment of cognitive load remains under development. Furthermore, it is apparent that although a variety of measurement methods can be used to assess cognitive load, they all have their strengths and limitations and the perfect single measurement method does not exist. Considering this, multimodal approaches, especially the combination of behavioral and (neuro-) physiological measurements, seem to be a potentially good solution for online and near-real-time assessment of cognitive load in adaptive systems. The research area is still very nascent and no common theoretical framework has been established yet, whereas most attempts in this regard are often based on data-driven probabilistic performance evaluations, which often require user- and situation-specific calibration and cannot be easily transferred to other contexts.

Mostly research on the assessment of cognitive load concentrates on investigating certain measures that are known to be associated with cognitive load from the empirical evidence (e.g., eye-tracking features such as pupil diameter of blink count) and relating them to performance/self-reported rating or applying machine learning classifications for different levels of cognitive load. Typically, no attempt is made to analyze acquired measures in the light of an established theoretical approach, although such an attempt might be useful in providing a general foundation for developing new metrics and making it easier to generalize the developed approaches to similar settings.

Such an attempt was made by Sweller (1988), who developed a computational model of WM using PRISM language (Langley & Neches, 1981). Based on this model, the cognitive load during the processing of a given task could be estimated based on task characteristics. However, this approach did not include individual and immediate feedback from learner and might therefore be suitable for preparing instructional material for the classroom, but did not qualify for online assessment of cognitive load. Five years later, Paas and Van Merriënboer (1993) presented a combined measurement method for assessing the efficiency of instructional material. According to the “efficiency view” (Ahern & Beatty, 1979), the learner is more efficient if his performance is higher than could be expected from his cognitive load and/or his cognitive load is lower than could be expected from his performance. Based on this idea, researchers developed an approach for measuring the efficiency of instructional materials using combined measures of cognitive load and task performance. Thereby, cognitive load was measured using a simple self-report nine-point Likert scale (Paas, 1992) and

performance quantified by the number of correct answers. Thus, this method might also not be used for online assessment of learner efficiency because, as described earlier, both self-report and performance-based measures are generally not available until task completion.

In this dissertation, I have attempted to provide a theoretical basis for the measurement of cognitive load, linking the theoretical background to the direct assessment of the users' cognitive load. An overview of the most influential theoretical models of working memory was provided (Section 1.3), and it became apparent that depending on how the known limitations of WM, and thus cognitive load, are explained, all models can be roughly divided into two main classes: emphasizing role of memory vs. attention as a limited resource. The best-known WM model by Baddeley and Hitch (1974) emphasizes the role of memory structures. However, in this work, however, the attention-oriented approach was pursued and the TBRS model of Barrouillet et al. (2004) was selected because it particularly emphasizes the role of time in WM functioning and therefore seems to be particularly suitable for assessing cognitive load in time-critical situations. An approach for its application to a realistic HMI was presented (Section 1.4).

What can we derive from the results of the three studies in this dissertation in terms of theoretical perspective? In fact, the suggestion of calculating the *TADD* metric appeared useful, namely periods of burst and idle were found in the game course. Difficulty level, which can also be interpreted as time pressure because it was constructed by presenting more tasks in the same given the amount of time available, was significantly positively associated with the metric, i.e. under increased time pressure, cognitive load increased, as quantified by the *initial TADD* metric. Thus, it was shown that time pressure leads to an increase in cognitive load, although the difficulty of the individual tasks did not change, which supports Barrouillets' hypothesis.

Moreover, winners tended to complete their tasks more quickly, which resulted in lower *initial TADD* scores, i.e. lower cognitive load. This observation is consistent with evidence (Csikszentmihalyi, 1975; Vygotsky, 1980; Yerkes & Dodson, 1908) and underlines that excessively high levels of cognitive load are associated with reduced performance.

But at this point, another question arises. Why were some participants more successful than others? The most obvious answer would be because some of them had

more experience. However, in the context of these analyzes, no significant effect of experience level on performance was found after accounting for the effect for difficulty level and scenario (see Chapter 2). The results suggest that performance appears to be related to participants' cognitive actions during the *initial idle* time period. While this association during the *initial burst* is straightforward and is strongly influenced by the imposed time pressure – we see increasing activity in both measurement domains associated with level difficulty, whereas cognitive overload is associated with decreased performance (as expected based on ample evidence, see Chapter 1), during the *initial idle* periods, participants had the choice of either passively waiting for the next burst or actively researching, observing, and planning their future strategy. Following the theoretical account what results could be expected for *initial idle* period based on two main approaches described in Section 1.3? Based on the hypothesis of limited memory capacity, we would expect the success of winners to be due to the greater capacity of their WM storage systems compared to participants who lose the level. In this case, they may be able to remember and process more objects and interactions at the same time, so that they could develop a more comprehensive picture of the gameplay, act more sufficiently, and consequently be more likely to win the level. In this case, based to the evidence showing that DLPFC activity is related to memorization and WL (Narayanan et al., 2005; D. J. Veltman, Rombouts, & Dolan, 2003), we would expect increased activity in DLPFC areas during *initial idle* period. Alternatively, based on the hypothesis of limited attention we might expect the winners to be able to switch their attention more quickly as compared to losers. Due to the fact that visual focus was found to be associated with attention (Kruschke, Kappenman, & Hetrick, 2005; Rehder & Hoffman, 2005), we would thus expect increased activity of eye-tracking features association with attention focus during this time. The results from the analyses of fNIRS and eye-tracking data provide evidence for the limited attention hypothesis. While cortical activation in the left DLPFC as well as pupil diameter and blink count (eye-tracking features related to non-visual cognitive load) during the *initial idle* phase showed no association with performance (see Tables 1, 4), at the same time better players showed more fixations, saccades, and microsaccades (eye-tracking features associated with attentional focus, see Tables 2, 4) compared to the participants who tended to fail the level.

In summary, this dissertation supports the applicability of the TBRS model as a theoretical basis for measuring cognitive load in realistic time pressure situations and suggests that time pressure and attention play a role in the development of cognitive

load. Furthermore, it also provides an indication of which features are more likely to be useful for assessing cognitive load (see Table 4).

5.3 Strengths and limitations

This dissertation provides first empirical evidence for a theory-based multimodal approach for assessing cognitive load in situations involving time pressure. In this section, the strengths and limitations of the presented concept are discussed..

5.3.1 Theory-based approach

One of the important strengths of the presented approach is that all proposed metrics were derived based on predictions of cognitive load made by the TBRS model, which is a well-established theoretical model of WM. While data-driven bottom-up approaches do not use theoretical knowledge about the task structure or the undergoing cognitive processes thus need to calibrate each situation and each new user individually, a theory-inspired approach provides the foundation for the simplified generalizability of the method to similar situations and potentially facilitates adaptation to different types of situations, if needed. For example, building on the TBRS model allows deriving several different ways to quantify the *TADD* metric. As one example, the mean value for a sequence of detected *TADDs* for some time period can be used to estimate cognitive load during this period. In the first study, the *mean TADD* metric was calculated as the mean of all *TADD* relations detected per level, but the results showed that *initial TADD* was superior to *mean TADD* in terms of predictive power, indicating that the beginning of the level appeared to be crucial for the final outcome. Theoretically, it might become necessary to adapt this method for other types of emergency situations, in which this would not be the case. For example, the critical interaction period (analog to the *initial action block*, which includes *initial burst* and *initial idle*) might potentially occur later, or not at all. In this case the interaction patterns would remain relatively stable throughout the interaction process. Due to the theoretical foundation of this approach, in this case the adjustment of the measurement method could be made simply by adjusting investigated time periods.

Nevertheless, it is probably not possible to generalize the proposed method to all potential HMI interactions. Presumably, it may well be used in settings with inherent time limits, particularly where participants are exposed to new or unusual situations and have to manage a large number of tasks and resources under time pressure. Examples of such situations may comprise different types of emergencies such as a critical situation on the road or in the sky, or a sudden critical situation during a surgery, but also less dramatic situations such as working on an assembly line or managing an event.

In this dissertation, the approach was validated on two different types of emergencies, i.e. the management of a fire and a train accident, providing promising initial evidence of the transferability between similar emergency situations. Nevertheless, further research will be needed in the future to evaluate whether this method is applicable to other scenarios that may differ in structure from the investigated situations. Moreover, more research will be needed to substantiate the predictive power of proposed metrics and possibly to these metrics appropriately to different situations.

Another important point in this context is that while using theory as the basis for the measurement method may provide advantages in terms of transferability, each new situation needs to be studied at least briefly in advance to select the most appropriate metrics or metric combinations. In this way, depending on the task and situation, this approach could be more profitable than calibrating a new neural network, but it would still not provide a fully transferable solution. One possible way to overcome this limitation could be to develop multimodal methods that are able to identify the situation and select appropriate metrics from a predefined set (see Section 5.3). However, even such multimodal approaches would likely first be calibrated to a given set of situations.

5.3.2 Multimodality

To accomplish the overarching research project on the differential value of different cognitive load measures, this dissertation implemented a multimodal measurement approach, which is the second major strength of this work. As shown in the introduction (Section 1.2), each of the existing measurement methods has its strengths and limitations, whereas there is no perfect single measure for cognitive load. For this reason, to ensure strong validity of the research, this work concentrated on three different methods that potentially qualify for online assessment: analysis of behavioral

data, fNIRS, and eye-tracking. Efforts were made to select methods that would not interfere with the performance of the primary task, be sensitive to variations in cognitive load over time, and at the same time be implemented as discretely and unobtrusively as possible. Importantly, the measurement methods were selected to complement each other to provide the comprehensive picture of the proposed approach. For example, behavioral data, which are very practical and easy to collect, do not provide accurate information about the users' cognitive state, which can be better estimated using neurophysiological methods, whereas eye-tracking can provide additional information about the visual focus of attention, which can be useful in interpreting cortical activation (cf. Scharinger, 2018; Scharinger et al., 2015; Scharinger et al., 2020).

However, despite the aforementioned advantages, the implementation of multimodal approaches comes with certain limitations. Strong attention was paid to render the interaction scenario as naturalistic as possible. This was one of the reasons for using the eye-tracking method, which was installed as a remote system positioned in an absolutely discreet way, without using glasses or chin rests and without restricting participants in any way apart from fluent calibration at the beginning of the session. Furthermore, the fNIRS was chosen partly due to its feasibility for naturalistic scenarios. fNIRS is the most naturalistic neuro-imaging method available, allowing much more freedom of movement for participants than other brain-imaging methods (such as EEG or fMRI). The wires of the fNIRS cap were sufficiently long for the participants to move their heads freely, and no gel had to be applied, as is required when using EEG. Nevertheless, using this equipment meant that participants had to wear this measurement equipment and sit relatively still in a quiet room under constant light conditions, implying some limitation of ecological validity.

At the same time, the measuring equipment did not disturb participants from operating the game, and therefore I hope that the interaction with the game did not lose its naturalistic and realistic character, especially considering the strong immersive character of the game used. This hope is sustained by the data collected in a small pilot study addressing a different age group. In this pilot experiment, about ten people were asked to play the game, with only in-game analyses recorded in the background. Although these participants were from different age spans and did not wear measurement equipment, they showed very similar behavior pattern.

5.3.3 Suitability for adaptive systems

Because cognitive load can rapidly change during the processing of a task, the measurement procedure suitable for online adaptation should be able to respond sensitively to these variations without causing external disturbances to performance on the primary task (Orru & Longo, 2019). This work is based on this principle, starting with the selection of measurement methods and ending with the fact that investigated time periods (*initial burst*, *initial idle*), which form the proposed *initial TADD* metric, occur at the beginning of the interaction. In this way, the probable outcome of the entire interaction can be predicted at an early stage, which enables the timely adaptation of the HMI system. Thus, the proposed method is suitable for near-real-time adaptation in the HMI environment involving time pressure.

At the same time, this approach comes with an important limitation. The *initial TADD* metric can be calculated relatively early, but not before the end of the action block, which took an average 50 seconds in the investigated scenarios. Even if one tries to gain some time and uses *initial burst* as a basis for calculation with neurophysiological or eye-tracking data (see Table 3), it takes on average 40 seconds until *initial burst* is completed. Hence, in the case of a critical situation when the system is supposed to react within milliseconds, e.g., in a car accident, this method would be too slow. In this way, the method appears useful for assessing cognitive load within *action blocks* of several minutes, such as in detecting cognitive states in managers during coping with an emergency, organizing an event, or assessing the cognitive load of assembly line workers. Furthermore, this method could be applied in the educational context in developing pedagogical adaptive simulations or serious games.

Another important limitation relates to the application of the fNIRS method in realistic commercial settings. I decided to use fNIRS as the least obtrusive direct neuroimaging method to gain a deeper insight into cortical activation during the *initial burst* and *initial idle* periods. This decision yielded good results that helped to better understand what cognitive processes might happen during these time periods. However, given that this method requires relatively complex and precise placement of the optodes on the users' head, it appears impractical for realistic application. Nonetheless, the fNIRS method might be used for research purposes when investigating potential new situations and metrics to act as a direct validation method. At the same time, the results of the second and third studies were largely consistent, suggesting that in this

application fNIRS could be replaced by eye-tracking. Therefore, for realistic commercial development, eye-tracking in combination with behavioral measurements seems to be a promising solution.

Although the eye-tracking method is also not completely free of limitations, it brings a major advantage. Assuming that such an adaptive system would be installed in a vehicle, only a few drivers would be willing to wear additional devices or caps to operate the system, whereas eye-tracking can be installed completely without body contact and would thus be inconspicuous to the user. However, it must be taken into account that eye-tracking systems usually still need initial calibration, makeup and glasses (especially sunglasses) can dramatically impair their work, and pupil detection is also lost if the user turns his head away too much.

5.3.4 Methodology

All studies in this dissertation were conducted based on a large data set obtained in the course of one extensive experiment. This fact might represent an important limitation in terms of generalizability. Although the theoretical foundation of the method, the results of the pilot experiment (see below) and the fact that two different emergency scenarios were used to induce cognitive load indicate that the method is likely to be transferrable between situations and samples, this must be further investigated in subsequent experiments, including different samples and situations.

The 47 participants who took part in the experiment were mainly students with an average age of about 25 years. In the context of this work, this sample seems permissible because basic cognitive mechanisms and indicators of cognitive load were studied, so it seems unlikely that a broader sampling procedure with more diversity in terms of social or socioeconomic parameters would substantially affect the pattern of results. Moreover, the pilot experiment revealed that the behavioral patterns were very similar among 60-year-olds from the middle-high socioeconomic layer, whereas at the same time some of the older participants had problems operating the mouse, and needed more time to understand the logic of the game. Based on these usability

problems and the fact that older participants are not the typical target group for playing digital games, it was decided to limit studies of this dissertation with a more representative sample of younger individuals. Nonetheless, this point represents a limitation. Because evidence indicates that cognitive abilities are affected by age, in future studies it might be worth investigating the distribution of ages in the sample, as well as expanding the number of participants and examined situations to improve the external validity of the results obtained.

Another aspect that should be considered and elaborated in the future is that although the presented studies were designed as realistic HMI in the context of playing a digital game, i.e. the participants were sitting in a quiet room in front of a laptop representing the game scene, this situation does not represent a realistic situation in the sense of a real emergency. Indeed, this step is important because evidence indicates that results obtained in simulations often differ from those obtained in corresponding real-life settings (e.g., Wilson et al., 1987). In this way, the next step would be to take the method outside the laboratory.

5.4 Implications for future research

In this section, key implications and corresponding ideas for future research are discussed. Finally, the practical relevance of the proposed method is outlined.

5.4.1 Combination of behavioral and physiological measurements

In general, the results of the second and third studies indicated that both cortical hemodynamic activation and eye-tracking features behavior during the *initial burst* phase were significantly positively associated with level difficulty and negatively associated with performance. This result was unexpected and disproved the important foundation for the *initial TADD* metric that all participants would be working at their cognitive limit at this time. At the same time, these findings help to explain why in the first study the *initial TADD* metric performed better when applied to a more challenging scenario. Possibly the time pressure induced by the Emergency serious game was not sufficient to push all participants to their limits. This raises a potential research question for future research. It would be interesting to investigate how time pressure affects the associations between task difficulty and performance, *initial TADD*, and (neuro-) physiological activation during the *initial burst*.

On the one hand, this indicates that the assumption made was correct and that strong time pressure during HMI is a necessary condition for the application of the *initial TADD* metric. At this point, another question may arise: what happens when the induced time pressure is too high? Is the metric still applicable in this case? The ideal setting for the application of the metric is when the participants reach their maximum processing speed. Further increasing the speed would lead to a decrease in the *initial TADD* relation, which would be interpreted as increased cognitive load, while at the same time increasing the in error rate resulting from the overloaded cognitive state according to the "inverted-U" relation between cognitive load and performance as described by Yerkes and Dodson (1908). In this case, the metric would correctly reflect the cognitive overload, which might be used as a signal for necessary adaptation to reduce the cognitive load and prevent errors. Thus, excessive time pressure does not seem to affect the validity of the metric, which nevertheless should be investigated in further studies. Consequently, the *initial TADD* metric appears to be well suited for situations with severe time pressure, while more relaxed situations might limit its operational capability.

On the other hand, because (neuro-) physiological activation during the *initial burst* and *initial idle* periods seems to show a stronger association with cognitive load specifically in more relaxed scenarios (which is another important aspect to be investigated in future studies), it seems reasonable to combine the behavioral *initial TADD* metric with the (neuro-) physiological data collected during the *initial burst* and *initial idle* periods.

Here, it should be noted that no assumptions were made regarding cognitive processes in the *initial idle* phase, and these are irrelevant to the operation of the *initial TADD* metric. Nevertheless, the results of the second study in particular showed a significant positive correlation between visual search behavior during this time and performance, with this pattern found consistently across different eye-tracking features. Thus, it seems reasonable to use these associations to predict final performance during HMI, although future research should investigate whether this association also depends on induced time pressure.

In summary, for the development of multimodal measurement systems, it is important to consider that the *initial TADD* metric should be primarily used under severe time constraints, whereas in milder settings cognitive activation during the *initial burst* appears to be positively associated with task difficulty and increased visual activity

during the *initial idle* is likely to predict better performance. In this way, it may be possible to close this gap and develop a system capable of predicting cognitive load for a wide range of situations with varying degrees of time pressure.

5.4.2 Detection of cognitive underload

In this dissertation, the second "horn" of the "inverted U" of Yerkes and Dodson (1908) was investigated. Accordingly, the presented approach was tested in a demanding environment under time pressure, which caused cognitive overload in some participants, and it was found that the proposed approach seems suitable to detect states of cognitive overload and related performance degradation. However, cognitive underload is also associated with performance deterioration and should therefore be detected by an adaptive system. For example, in the case of autonomous driving, where the driver is still present and needs to monitor the system, this approach might also be suitable to define states of cognitive underload and fatigue. In this case, a combination of behavioral data with eye-tracking seems particularly appropriate because, in addition to the eye-tracking features described above, it can also be used to estimate fatigue caused by eyes that are closed for too long (e.g., Eriksson & Papanikotopoulos, 1997). This represents an important future direction for further research on this method.

5.4.3 Development of new metrics

Another direction to consider in this context is to develop new behavioral or physiological metrics based on the proposed principle. Accordingly, for instance, in the first study, the *mean TADD* metric was proposed, which was calculated as the mean for all *TADD* relations detected per level. This metric has not been further investigated in subsequent studies. However, its application could be promising in other HMI contexts that might be structurally different from the interaction studied in this dissertation. The *mean TADD* metric can be beneficial when multiple adjustments are required per session. For example, the *mean TADD* can be calculated every x minutes, and based on the result of this calculation an appropriate adjustment to the user interface or training environment can be made. It is also conceivable that other metrics that were not

investigated in this dissertation might be developed and adjusted to different types of contexts.

5.4.4 Practical relevance

As discussed above, the proposed method might be used for developing standalone systems capable of predicting cognitive load for a diversity of situations with varying degrees of time pressure. Whereas the use of more complex psychophysiological measures would come with additional computational and procurement costs, systems that operate on simple unobtrusive metrics may be made more easily accessible to the general public. More complex systems relying on resources such as neural networks are computationally rather expensive and might require substantial computing power. By contrast, simpler models for cognitive load estimation may be easily run in the background without significant consumption of computing resources and thus they would be suitable for real-life applications.

Additionally, complex multimodal measurement systems that operate with sophisticated algorithms and integrate data from physiological and behavioral sources in research laboratories may benefit from the development of simpler metrics as these may be added to these more complex algorithms quite easily, thereby leading to improved classification accuracy in the future. The substitution of more complex probabilistic algorithms through simpler but reliable metrics (whenever possible) might lead to the simplifications of complex models, while at the same time expanding their availability and usage, specifically if using theory-based metrics.

5.5 Conclusion and outlook

This dissertation presents a novel approach to measure cognitive load based on behavioral and (neuro)-physiological data during HMIs in situations involving the management of resources under time pressure. To potentially achieve the easy generalizability of the approach to similar contexts, it was developed relying on the established theoretical model of WM and cognitive load, namely the TBRS model (Barrouillet et al., 2004). Importantly, the method is potentially suitable for use in near-real-time adaptive systems given that the proposed metrics can be computed relatively

early during the interaction. At the same time, the unobtrusiveness and simplicity of the proposed measurements appear promising for use in lightweight standalone systems or to extend comprehensive multimodal solutions. The results obtained in a series of three studies within the scope of this dissertation appeared stable and consistent, presenting a comprehensive picture of the method and cognitive activities that take place during investigated time periods. It was consistently demonstrated that theoretically specified time periods at the beginning of a situation involving management of resources under time pressure play a critical role and can be used to predict the outcome of the entire interaction. This observation was thoroughly verified using different measurement domains including behavioral measurements, cortical activation measurements, and eye-tracking data. The approach seems promising for developing realistic HMI systems capable of adapting to the cognitive load of their operators in near-real-time. In summary, this dissertation supports the applicability of the TBRS model as a theoretical basis for measuring cognitive load in realistic time pressure situations and suggests that time pressure and attention play a role in the development of cognitive load. Furthermore, it also provides an indication of which features are more likely to be useful for assessing cognitive load under time pressure. Further research is needed to complete the development of the method and provide a ready-to-use concept for practical application.

6 References

Aasted, C. M., Yücel, M. A., Cooper, R. J., Dubb, J., Tsuzuki, D., Becerra, L., Petkov, M. P., Borsook, D., Dan, I., & Boas, D. A. (2015). Anatomical guidance for functional near-infrared spectroscopy: AtlasViewer tutorial. Neurophotonics, 2(2), 020801.

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. Science, 205(4412), 1289-1292.

Ahmad, R. F., Malik, A. S., Kamel, N., & Reza, F. (2016). Machine learning approach for classifying the cognitive states of the human brain with functional magnetic resonance imaging (fMRI). Paper presented at the 2016 6th International Conference on Intelligent and Advanced Systems (ICIAS).

Alexander, R. G., & Martinez-Conde, S. (2019). Fixational Eye Movements. In Klein, C. & Ettinger, U. (Eds.), Eye Movement Research: An Introduction to its Scientific Foundations and Applications (pp. 73-115). Cham: Springer International Publishing.

Andreassi, J. L. (2013). Psychophysiology: Human behavior & physiological response (4 ed.). USA: Lawrence Erlbaub Associates.

Andreessen, L. M., Gerjets, P., Meurers, D., & Zander, T. O. (2021). Toward neuroadaptive support technologies for improving digital reading: a passive BCI-based assessment of mental workload imposed by text difficulty and presentation speed during reading. User Modeling and User-Adapted Interaction, 31(1), 75-104.

Appel, T., Sevcenko, N., Wortha, F., Tsarava, K., Moeller, K., Ninaus, M., Kasneci, E., & Gerjets, P. (2019). Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures. In Gao, W., Meng, H. M. L., Turk, M., Fussell, S. R., Schuller, B., Song, Y., & Yu, K. (Eds.), 2019 International Conference on Multimodal Interaction (pp. 154-163). New York United States: Association for Computing Machinery.

Aricò, P., Borghini, G., Di Flumeri, G., Colosimo, A., Bonelli, S., Golfetti, A., Pozzi, S., Imbert, J.-P., Granger, G., & Benhacene, R. (2016). Adaptive automation triggered by EEG-based mental workload index: a passive brain-computer interface application in realistic air traffic control environment. Frontiers in human neuroscience, 10, 539.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Psychology of learning and motivation (Vol. 2, pp. 89-195): Elsevier.

Ayaz, H., Izzetoglu, M., Bunce, S., Heiman-Patterson, T., & Onaral, B. (2007). Detecting cognitive activity related hemodynamic signal for brain computer interface using functional near infrared spectroscopy. Paper presented at the Neural Engineering, 2007. CNE'07. 3rd International IEEE/EMBS Conference on.

Babiloni, F. (2019). Mental Workload Monitoring: New Perspectives from Neuroscience. In Longo, L. & Leva, M. (Eds.), Human Mental Workload: Models and Applications. H-WORKLOAD 2019. Communications in Computer and Information Science (Vol. 1107, pp. 3-19). Cham: Springer.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. Journal of Experimental Psychology: General, 133(1), 83.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(3), 570.

Barrouillet, P., & Camos, V. (2015). Working memory: Loss and reconstruction (Roediger, H., Pomerantz, J., Baddeley, A. D., Bruce, V., & Grainger, J. Eds.). New York: Psychology Press.

Benedetto, S., Pedrotti, M., & Bridgeman, B. (2011). Microsaccades and exploratory saccades in a naturalistic environment. Journal of Eye Movement Research, 4(2). doi:10.16910/jemr.4.2.2

Berthold, A., & Jameson, A. (1999). Interpreting symptoms of cognitive load in speech input. In Kay, J. (Ed.), UM99 user modeling. CISM International Centre for Mechanical Sciences (Courses and Lectures) (Vol. 407, pp. 235-244). Vienna: Springer.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. Educational psychologist, 38(1), 53-61.

Brünken, R., Seufert, T., & Paas, F. (2010). Measuring cognitive load.

Buchwald, M., Kupiński, S., Bykowski, A., Marcinkowska, J., Ratajczyk, D., & Jukiewicz, M. (2019). Electrodermal activity as a measure of cognitive load: a methodological approach. Paper presented at the 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA).

Baddeley, A. (2000). The episodic buffer: a new component of working memory? Trends in cognitive sciences, 4(11), 417-423.

Baddeley, A. (2012). Working memory: theories, models, and controversies. Annual review of psychology, 63, 1-29.

Baddeley, A., & Hitch, G. (1974). Working memory. In Psychology of learning and motivation (Vol. 8, pp. 47-89): Elsevier.

Baddeley, A. D., & Logie, R. (1999). Working Memory: The Multiple-Component Model. In Miyake, A. & Shah, P. (Eds.), Models of working memory: Mechanisms of active maintenance and executive control (pp. 62-101). New York: Cambridge University Press. (Reprinted from: 2003, 2004, 2007).

Camos, V., Portrat, S., Vergauwe, E., & Barrouillet, P. (2007). The cognitive cost of executive functions. Paper presented at the Joint Meeting of the EPS and the Psychonomic Society, Edinburgh (Great-Britain).

Capon, N., Farley, J. U., & Hulbert, J. M. (1994). Strategic planning and financial performance: more evidence. Journal of Management Studies, 31(1), 105-110.

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. Journal of experimental child psychology, 33(3), 386-404.

Charabati, S., Bracco, D., Mathieu, P., & Hemmerling, T. (2009). Comparison of four different display designs of a novel anaesthetic monitoring system, the 'integrated monitor of anaesthesia (IMA™)'. British journal of anaesthesia, 103(5), 670-677.

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., & Conway, D. (2016). Robust multimodal cognitive load measurement: Springer.

Chen, S., Epps, J., Ruiz, N., & Chen, F. (2011). Eye activity as a measure of human mental effort in HCI. Paper presented at the Proceedings of the 16th international conference on Intelligent user interfaces.

Clifton Jr, C., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. Journal of Memory and Language, 86, 1-19.

Cope, M., Delpy, D., Reynolds, E., Wray, S., Wyatt, J., & Van der Zee, P. (1988). Methods of quantitating cerebral near infrared spectroscopy data. In Oxygen Transport to Tissue X (pp. 183-189): Springer.

Cowan, N. (1998). Attention and memory: An integrated framework: Oxford University Press.

Cowan, N. (1999). An Embedded-Processes Model of Working Memory. In Miyake, A. & Shah, P. (Eds.), Models of working memory: Mechanisms of active maintenance and executive control (pp. 62-101). New York: Cambridge University Press. (Reprinted from: 2003, 2004, 2007).

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? Current directions in psychological science, 19(1), 51-57.

Croskerry, P., & Sinclair, D. (2001). Emergency medicine: A practice prone to error? Canadian Journal of Emergency Medicine, 3(4), 271-276.

Csikszentmihalyi, M. (1975). Beyond boredom and anxiety (Csikszentmihalyi, I., Graef, R., Holcomb, J. H., Hendin, J., & MacAloon, J. Eds. 1 ed.). San Francisco, London: Jossey-Bass Publishers.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. Journal of Memory and Language, 19(4), 450.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. Psychonomic bulletin & review, 3(4), 422-433.

De Rivecourt, M., Kuperus, M., Post, W., & Mulder, L. (2008). Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. Ergonomics, 51(9), 1295-1319. doi:10.1080/00140130802120267

Derosière, G., Mandrick, K., Dray, G., Ward, T. E., & Perrey, S. (2013). NIRS-measured prefrontal cortex activity in neuroergonomics: strengths and weaknesses. Frontiers in human neuroscience, 7, 583.

Dias, R. D., Ngo-Howard, M. C., Boskovski, M. T., Zenati, M. A., & Yule, S. J. (2018). Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload. Journal of British Surgery, 105(5), 491-501.

Eggemeier, F. T., Shingledecker, C. A., & Crabtree, M. S. (1985). Workload measurement in system design and evaluation. In Proceedings of the Human Factors Society Annual Meeting (Vol. 29, pp. 215-219). Los Angeles, CA: SAGE Publications Sage CA.

Eggemeier, F. T., Wilson, G. F., Kramer, A. F., & Damos, D. L. (1991). Workload assessment in multi-task environments. In Damos, D. L. (Ed.), Multiple-task performance (pp. 207-216). London, Washington, DC: Taylor & Francis.

Ehlis, A.-C., Herrmann, M., Wagener, A., & Fallgatter, A. (2005). Multi-channel near-infrared spectroscopy detects specific inferior-frontal activation during incongruent Stroop trials. Biological psychology, 69(3), 315-331.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. Models of working memory: Mechanisms of active maintenance and executive control, 4, 102-134.

Eriksson, M., & Papanikotopoulos, N. P. (1997). Eye-tracking for detection of driver fatigue. Paper presented at the Proceedings of Conference on Intelligent Transportation Systems.

Fallgatter, A., Ehlis, A., Wagener, A., Michel, T., & Herrmann, M. (2004). Near-infrared spectroscopy in psychiatry. Der Nervenarzt, 75(9), 911.

Fan, J., & Smith, A. P. (2017). The impact of workload and fatigue on performance. In Longo, L. & Leva, M. (Eds.), Human Mental Workload: Models and Applications. H-WORKLOAD 2017. Communications in Computer and Information Science (Vol. 726, pp. 90-105). Cham: Springer.

Fishburn, F. A., Norr, M. E., Medvedev, A. V., & Vaidya, C. J. (2014). Sensitivity of fNIRS to cognitive state and load. Frontiers in human neuroscience, 8, 76.

Fontaine, G., Cossette, S., Maheu-Cadotte, M.-A., Mailhot, T., Deschênes, M.-F., Mathieu-Dupuis, G., Côté, J., Gagnon, M.-P., & Dubé, V. (2019). Efficacy of adaptive e-learning for health professionals and students: a systematic review and meta-analysis. BMJ open, 9(8), e025252.

Fowler, A., Nesbitt, K., & Canossa, A. (2019). Identifying Cognitive Load in a Computer Game: An exploratory study of young children. Paper presented at the 2019 IEEE Conference on Games (CoG).

Fukuda, K., Stern, J. A., Brown, T. B., & Russo, M. B. (2005). Cognition, blinks, eye-movements, and pupillary movements during performance of a running memory task. Aviation, space, and environmental medicine, 76(7), C75-C85. Retrieved from https://www.ingentaconnect.com/content/asma/asem/2005/00000076/A00107s1/art00014

Gao, X., Yan, H., & Sun, H.-j. (2015). Modulation of microsaccade rate by task difficulty revealed through between-and within-trial comparisons. Journal of vision, 15(3), 3-3. doi:doi.org/10.1167/15.3.3

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., & Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. Frontiers in neuroscience, 8, 385.

Gopher, D., & Braune, R. (1984). On the psychophysics of workload: Why bother with subjective measures? Human factors, 26(5), 519-532.

Grissmann, S., Faller, J., Scharinger, C., Spüler, M., & Gerjets, P. (2017). Electroencephalography based analysis of working memory load and affective valence in an n-back task with emotional stimuli. Frontiers in human neuroscience, 11, 616.

Grissmann, S., Spüler, M., Faller, J., Krumpe, T., Zander, T. O., Kelava, A., Scharinger, C., & Gerjets, P. (2017). Context sensitivity of EEG-based workload classification under different affective valence. IEEE Transactions on Affective Computing, 11(2), 327-334.

Hancock, G., Hancock, P., & Janelle, C. (2012). The impact of emotions and predominant emotion regulation technique on driving performance. Work, 41(Supplement 1), 3608-3611.

Hancock, P. (1989). The effect of performance failure and task demand on the perception of mental workload. Applied Ergonomics, 20(3), 197-205.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Advances in psychology (Vol. 52, pp. 139-183): Elsevier.

He, X., Wang, L., Gao, X., & Chen, Y. (2012). The eye activity measurement of mental workload based on basic flight task. Paper presented at the IEEE 10th International Conference on Industrial Informatics.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2014). Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS. Frontiers in human neuroscience, 7, 935.

Herold, F., Wiegel, P., Scholkmann, F., & Müller, N. G. (2018). Applications of functional near-infrared spectroscopy (fNIRS) neuroimaging in exercise–cognition science: a systematic, methodology-focused review. Journal of clinical medicine, 7(12), 466.

Herrmann, M. J., Walter, A., Schreppel, T., Ehlis, A. C., Pauli, P., Lesch, K. P., & Fallgatter, A. (2007). D4 receptor gene variation modulates activation of prefrontal cortex during working memory. European Journal of Neuroscience, 26(10), 2713-2718.

Hoge, R. D., Atkinson, J., Gill, B., Crelier, G. R., Marrett, S., & Pike, G. B. (1999). Linear coupling between cerebral blood flow and oxygen consumption in activated human cortex. Proceedings of the National Academy of Sciences, 96(16), 9403-9408.

Hutton, S. B. (2019). Eye Tracking Methodology. In Klein, C. & Ettinger, U. (Eds.), Eye Movement Research: An Introduction to its Scientific Foundations and Applications (pp. 277-308). Cham: Springer International Publishing.

Ikehara, C. S., & Crosby, M. E. (2005). Assessing cognitive load with physiological sensors. In Proceedings of the 38th annual hawaii international conference on system sciences (pp. 295a-295a). New York: IEEE.

Jobsis, F. F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. Science, 198(4323), 1264-1267.

Johannsen, G. (1979). Workload and workload measurement. In Moray, N. (Ed.), Mental Workload (Vol. 8, pp. 3-11). Boston: Springer.

Kahneman, D. (1973). Attention and effort (Vol. 1063): Prentice-Hall Englewood Cliffs, NJ.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. Educational psychology review, 19(4), 509-539.

Kerr, B. (1973). Processing demands during mental operations. Memory & Cognition, 1(4), 401-412.

Khawaja, M. A., Chen, F., Owen, C., & Hickey, G. (2009). Cognitive Load Measurement from User's Linguistic Speech Features for Adaptive Interaction Design, Berlin, Heidelberg.

Kiili, K., Lindstedt, A., & Ninaus, M. (2018, May 21-23, 2018). Exploring characteristics of students' emotions, flow and motivation in a math game competition. Paper presented at the GamiFIN Conference, Pori, Finland.

Kim, S. G., Rostrup, E., Larsson, H. B., Ogawa, S., & Paulson, O. B. (1999). Determination of relative CMRO2 from CBF and BOLD changes: significant increase of oxygen consumption rate during visual stimulation. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 41(6), 1152-1161.

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

Kohlmorgen, J., Dornhege, G., Braun, M., Blankertz, B., Müller, K.-R., Curio, G., Hagemann, K., Bruns, A., Schrauf, M., & Kincses, W. (2007). Improving human performance in a real operating environment through real-time mental workload detection. In Dornhege, G., Millan, J. d. R., Hinterverger, T., McFarland, D. J., & Müller, K.-R. (Eds.), Toward Brain-Computer Interfacing (Vol. 409422, pp. 409-422). Cambridge, Massachusetts, London, England: MIT Press.

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(5), 830.

Langley, P., & Neches, R. (1981). PRISM user's manual. Technical report. In. University of Pittsburgh: Department of Psychology, Carnegie-Mellon University and Learning Research and Development Center

Lépine, R., Bernardin, S., & Barrouillet, P. (2005). Attention switching and working memory spans. European Journal of Cognitive Psychology, 17(3), 329-345.

Li, C., Gong, H., Gan, Z., & Luo, Q. (2005). Monitoring of prefrontal cortex activation during verbal n-back task with 24-channel functional NIRS imager. Paper presented at the Optics in Health Care and Biomedical Optics: Diagnostics and Treatment II.

Liang, Y., Liang, W., Qu, J., & Yang, J. (2018). Experimental study on EEG with different cognitive load. Paper presented at the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC).

Liefooghe, B., Barrouillet, P., Vandierendonck, A., & Camos, V. (2008). Working memory costs of task switching. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(3), 478.

Lim, Y. M., Ayesh, A., & Stacey, M. (2015). Using mouse and keyboard dynamics to detect cognitive stress during mental arithmetic. In Arai, K., Kapoor, S., & Bhatia, R. (Eds.), Intelligent Systems in Science and Information 2014. SAI 2014. Studies in Computational Intelligence (Vol. 591, pp. 335-350). Cham: Springer.

Lin, B., & Wu, C. (2010). Mathematical modeling of the human cognitive system in two serial processing stages with its applications in adaptive workload-management systems. IEEE Transactions on intelligent transportation systems, 12(1), 221-231.

Linton, P., Jahns, D., & Chatelier, P. (1978). Operator workload assessment model: An evaluation of a VF/VA-V/STOL system. AGARD Methods to Assess Workloads 12 p(SEE N 78-31745 22-54).

Liu, R., Walker, E., Friedman, L., Arrington, C. M., & Solovey, E. T. (2020). fNIRS-based classification of mind-wandering with personalized window selection for multimodal learning interfaces. Journal on multimodal user interfaces, 1-16. doi:https://doi.org/10.1007/s12193-020-00325-z

Magerko, B., Stensrud, B. S., & Holt, L. S. (2006). Bringing the schoolhouse inside the box-a tool for engaging, individualized training. Retrieved from https://apps.dtic.mil/sti/pdfs/ADA481593.pdf

Magnusdottir, E. H., Borsky, M., Meier, M., Johannsdottir, K., & Gudnason, J. (2017). Monitoring cognitive workload using vocal tract and voice source features. Periodica Polytechnica Electrical Engineering and Computer Science, 61(4), 297-304.

Malmberg, K. J., Raaijmakers, J. G., & Shiffrin, R. M. (2019). 50 years of research sparked by Atkinson and Shiffrin (1968). Memory & Cognition, 47(4), 561-574.

McDuff, D., Gontarek, S., & Picard, R. (2014). Remote measurement of cognitive stress via heart rate variability. Paper presented at the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Meshkati, N. (1988). Toward development of a cohesive model of workload. In Advances in Psychology (Vol. 52, pp. 305-314): Elsevier.

Michon, J. A. (1993). Generic intelligent driver support: CRC Press.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. Annual review of neuroscience, 24(1), 167-202.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review, 63(2), 81.

Mirbabaie, M., & Fromm, J. (2019). Reducing the cognitive load of decision-makers in emergency management through augmented reality.

Nakayama, M., Takahashi, K., & Shimizu, Y. (2002). The act of task difficulty and eye-movement frequency for the'Oculo-motor indices'. In Proceedings of the 2002 symposium on Eye tracking research & applications (pp. 37-42): ACM Digital Library.

Narayanan, N. S., Prabhakaran, V., Bunge, S. A., Christoff, K., Fine, E. M., & Gabrieli, J. D. (2005). The role of the prefrontal cortex in the maintenance of verbal working memory: an event-related FMRI analysis. Neuropsychology, 19(2), 223.

Nebel, S., & Ninaus, M. (2019). New perspectives on game-based assessment with process data and physiological signals. In Ifenthaler, D. & Kim, Y. (Eds.), Game-Based Assessment Revisited. Advances in Game-Based Learning (pp. 141-161). Cham: Springer.

Nihashi, T., Ishigaki, T., Satake, H., Ito, S., Kaii, O., Mori, Y., Shimamoto, K., Fukushima, H., Suzuki, K., & Umakoshi, H. (2019). Monitoring of fatigue in radiologists during prolonged image interpretation using fNIRS. Japanese journal of radiology, 37(6), 437-448.

Ninaus, M., & Nebel, S. (2021). A Systematic Literature Review of Analytics for Adaptivity Within Educational Video Games. Front. Educ. 5: 611072. doi: 10.3389/feduc.

Ninaus, M., Witte, M., Kober, S. E., Friedrich, E. V., Kurzmann, J., Hartsuiker, E., Neuper, C., & Wood, G. (2013). Neurofeedback and serious games. In T. M. Connolly, E., Boyle, T., Hainey, G., Baxter, P., & Moreno-ger (Eds.), Psychology, Pedagogy, and Assessment in Serious Games (Vol. i, pp. 82-110). USA: IGI Global.

O'Donnell, R., & Eggemeier, F. (1986). Workload assessment methodology. Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance. KR Boff, L. Kaufman and JP Thomas. In: John Wiley and Sons, Inc.

Orru, G., & Longo, L. (2019). The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and Germane loads: a review. In Longo, L. & Leva, M. (Eds.), Human Mental Workload: Models and Applications. H-WORKLOAD 2018. Communications in Computer and Information Science (pp. 23-48). Cham: Springer.

Oviatt, S. (2006). Human-centered design meets cognitive load theory: designing interfaces that help people think. Paper presented at the Proceedings of the 14th ACM international conference on Multimedia.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. Journal of educational psychology, 84(4), 429.

Paas, F. G., & Van Merriënboer, J. J. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. Human factors, 35(4), 737-743.

Paas, F. G., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. Educational psychology review, 6(4), 351-371.

Parasuraman, R., & Rizzo, M. (2006). Introduction to Neuroergonomics. In.

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.

Portrat, S. (2008). Working memory and executive functions: The Time-Based Resource-Sharin account Université de Bourgogne. Dijon.

Pouget, P. (2019). Introduction to the Study of Eye Movements. In Klein, C. & Ettinger, U. (Eds.), Eye Movement Research: An Introduction to its Scientific Foundations and Applications (pp. 3-10). Cham: Springer International Publishing.

Promotion Software GmbH. (1999). World of Emergency. Retrieved from https://www.world-of-emergency.com/?lang=en

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological bulletin, 124(3), 372.

Rehder, B., & Hoffman, A. B. (2005). Thirty-something categorization results explained: selective attention, eyetracking, and models of category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(5), 811.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In Advances in psychology (Vol. 52, pp. 185-218): Elsevier.

Riener, A., & Noldi, J. (2015). Cognitive load estimation in the car: Practical experience from lab and on-road tests. Paper presented at the Adjunct Proceedings of Automotive UI 2015, Workshop Practical Experiences in Measuring and Modeling Drivers and Driver-Vehicle Interactions.

Rosanne, O., Albuquerque, I., Cassani, R., Gagnon, J.-F., Tremblay, S., & Falk, T. H. (2021). Adaptive filtering for improved eeg-based mental workload assessment of ambulant users. Frontiers in neuroscience, 15, 341.

Roth, G., Schulte, A., Schmitt, F., & Brand, Y. (2019). Transparency for a Workload-Adaptive Cognitive Agent in a Manned–Unmanned Teaming Application. IEEE Transactions on Human-Machine Systems, 50(3), 225-233.

Ruck, H., & Stottan, T. (2014). Interactive Steering Wheel for Optimal Operation. ATZ worldwide, 116(5), 40-45.

Ruiz, N., Liu, G., Yin, B., Farrow, D., & Chen, F. (2010). Teaching athletes cognitive skills: detecting cognitive load in speech input. Proceedings of HCI 2010, 484-488.

Saddler, B., Moran, S., Graham, S., & Harris, K. R. (2004). Preventing writing difficulties: The effects of planning strategy instruction on the writing performance of struggling writers. Exceptionality, 12(1), 3-17.

Sarkar, P., Ross, K., Ruberto, A. J., Rodenbura, D., Hungler, P., & Etemad, A. (2019). Classification of cognitive load and expertise for adaptive simulation using deep multitask learning. Paper presented at the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII).

Scerbo, M. W. (1996). Theoretical perspectives on adaptive automation. In Parasuraman, R. & Mouloua, M. (Eds.), Automation and Human Performance: Theory and Applications (pp. 37-64): CRC Press.

Scharinger, C. (2018). Fixation-Related EEG Frequency Band Power Analysis: A Promising Methodology for Studying Instructional Design Effects of Multimedia Learning Material. Frontline Learning Research, 6(3), 56-71.

Scharinger, C., Kammerer, Y., & Gerjets, P. (2015). Pupil dilation and EEG alpha frequency band power reveal load on executive functions for link-selection processes during text reading. PloS one, 10(6), e0130608.

Scharinger, C., Schüler, A., & Gerjets, P. (2020). Using eye-tracking and EEG to study the mental processing demands during learning of text-picture combinations. International Journal of Psychophysiology, 158, 201-214.

Sheridan, T. B., & Simpson, R. (1979). Toward the definition and measurement of the mental workload of transport pilots. Retrieved from https://dspace.mit.edu/handle/1721.1/67913

Siegenthaler, E., Costela, F. M., McCamy, M. B., Di Stasi, L. L., Otero- Millan, J., Sonderegger, A., Groner, R., Macknik, S., & Martinez- Conde, S. (2014). Task difficulty in mental arithmetic affects microsaccadic rates and magnitudes. European Journal of Neuroscience, 39(2), 287-294. doi:10.1111/ejn.12395

Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses. Psychophysiology, 45(5), 679-687. doi:10.1111/j.1469-8986.2008.00681.x

Smith-Jackson, T. L., & Klein, K. W. (2009). Open-plan offices: Task performance and mental workload. Journal of Environmental Psychology, 29(2), 279-289.

Smith, E. E., & Jonides, J. (1997). Working memory: A view from neuroimaging. Cognitive psychology, 33(1), 5-42.

Spronck, P., Ponsen, M., Sprinkhuizen-Kuyper, I., & Postma, E. (2006). Adaptive game AI with dynamic scripting. Machine Learning, 63(3), 217-248.

Spüler, M., Walter, C., Rosenstiel, W., Gerjets, P., Moeller, K., & Klein, E. (2016). EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning. ZDM, 48(3), 267-278.

Stapel, J., Mullakkal-Babu, F. A., & Happee, R. (2019). Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving. Transportation research part F: traffic psychology and behaviour, 60, 590-605.

Statistisches Bundesamt. (2017). Verkehrsunfälle. Unfälle von Güterkraftfahrzeugen im Straßenverkehr. Retrieved from https://www.destatis.de/DE/Publikationen/Thematisch/TransportVerkehr/Verkehrsunfaelle/UnfaelleGueterkraftfahrzeuge5462410167004.pdf?__blob=publicationFile

Strangman, G., Boas, D. A., & Sutton, J. P. (2002). Non-invasive neuroimaging using near-infrared light. Biological psychiatry, 52(7), 679-693.

Strangman, G., Goldstein, R., Rauch, S. L., & Stein, J. (2006). Near-infrared spectroscopy and imaging for investigating stroke rehabilitation: test-retest reliability and review of the literature. Archives of physical medicine and rehabilitation, 87(12), 12-19.

Strangman, G. E., Li, Z., & Zhang, Q. (2013). Depth sensitivity and source-detector separations for near infrared spectroscopy based on the Colin27 brain template. PloS one, 8(8), e66319.

Strenzke, R., Uhrmann, J., Benzler, A., Maiwald, F., Rauschert, A., & Schulte, A. (2011). Managing cockpit crew excess task load in military manned-unmanned teaming missions by dual-mode cognitive automation approaches. Paper presented at the AIAA guidance, navigation, and control conference.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257-285.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. Educational psychology review, 10(3), 251-296.

Thier, P. (2006). Die funktionelle Architektur des präfrontalen Kortex. In Neuropsychologie (pp. 471-478): Springer.

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T.-P. (2001). Eye activity correlates of workload during a visuospatial memory task. Human factors, 43(1), 111-121. doi:10.1518/001872001775992570

Vella, K. M., Hall, A. K., van Merrienboer, J. J., Hopman, W. M., & Szulewski, A. An exploratory investigation of the measurement of cognitive load on shift: application of cognitive load theory in emergency medicine. AEM Education and Training, e10634.

Veltman, D. J., Rombouts, S. A., & Dolan, R. J. (2003). Maintenance versus manipulation in verbal working memory revisited: an fMRI study. Neuroimage, 18(2), 247-256.

Veltman, J., & Gaillard, A. (1996). Physiological indices of workload in a simulated flight task. Biological psychology, 42(3), 323-342.

Veltman, J., & Jansen, C. (2005). The role of operator state assessment in adaptive automation. Retrieved from https://apps.dtic.mil/sti/citations/ADA455055

Vygotsky, L. S. (1980). Mind in society: The development of higher psychological processes (Cole, M., John-Steiner, V., Scribner, S., & Souberman, E. Eds.). Cambridge, London: Harvard university press.

Walter, C., Rosenstiel, W., Bogdan, M., Gerjets, P., & Spüler, M. (2017). Online EEG-Based Workload Adaptation of an Arithmetic Learning Environment. Frontiers in human neuroscience, 11(286). doi:10.3389/fnhum.2017.00286

Welford, A. (1978). Mental work-load as a function of demand, capacity, strategy and skill. Ergonomics, 21(3), 151-167.

Wilson, G. F., Purvis, B., Skelly, J., Fullenkamp, P., & Davis, I. (1987). Physiological data used to measure pilot workload in actual flight and simulator conditions. Paper presented at the Proceedings of the Human Factors Society Annual Meeting.

Wilson, G. F., & Russell, C. A. (2007). Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. Human factors, 49(6), 1005-1018.

Wu, C., Tsimhoni, O., & Liu, Y. (2008). Development of an adaptive workload management system using the queueing network-model human processor (QN-MHP). IEEE Transactions on intelligent transportation systems, 9(3), 463-475.

Xu, X., Deng, Z.-Y., Huang, Q., Zhang, W.-X., Qi, C.-z., & Huang, J.-A. (2017). Prefrontal cortex-mediated executive function as assessed by Stroop task performance associates with weight loss among overweight and obese adolescents and young adults. Behavioural brain research, 321, 240-248.

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.

Yuksel, B. F., Oleson, K. B., Harrison, L., Peck, E. M., Afergan, D., Chang, R., & Jacob, R. J. (2016). Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In Proceedings of the 2016 CHI conference on human factors in computing systems (pp. 5372-5384).

Yurko, Y. Y., Scerbo, M. W., Prabhu, A. S., Acker, C. E., & Stefanidis, D. (2010). Higher mental workload is associated with poorer laparoscopic performance as measured by the NASA-TLX tool. Simulation in healthcare, 5(5), 267-271. doi:10.1097/SIH.0b013e3181e3f329

Zhou, T., Cha, J. S., Gonzalez, G., Wachs, J. P., Sundaram, C. P., & Yu, D. (2020). Multimodal Physiological Signals for Workload Prediction in Robot-assisted Surgery. ACM Transactions on Human-Robot Interaction (THRI), 9(2), 1-26.

Zook, A. E., & Riedl, M. O. (2012). A temporal data-driven player model for dynamic difficulty adjustment. Paper presented at the Eighth Artificial Intelligence and Interactive Digital Entertainment Conference.