Utilizing a Virtual Environment to Measure Cognitive Load using Eye Tracking Technology

Master of Science Thesis

Department of Psychology

submitted at the

Universität Konstanz

by

Ariane Rahn

1. Supervisor: Prof. Dr. Harald Reiterer (Human-Computer Interaction Group)
2. Supervisor: Prof. Dr. Ronald Huebner (Cognitive Psychology Group)

Konstanz, 2019

Abstract

A valid measurement of Cognitive Load (CL) is of great interest in several areas, such as Human-Computer Interaction (HCI), education, medical training, aviation simulations, and the military industry. *Task-Evoked Pupillary Responses* (TEPRs) are a widely used method to measure CL. Especially the pupil dilation seems to be an indicator of cognitive processing. But TEPRs underlie a luminance bias, which is an empirical obstacle in CL research. This pioneer work used VR Technology to explore a new approach to better address this issue. The *n-back* task was used with three levels to induce CL: Low CL (*0-back*), medium CL (*1-back*), and high CL (*2-back*). An integrated eye tracker provided eye-related parameters. Furthermore, the impact of CL on the emotional state (SAM), perceived stress (PASA) and subjective CL (NASA TLX) was assessed.

Findings indicate a rather small effect of CL on TEPRs since CL only partly increased significantly with increasing task difficulty. The novel IPA calculation did not render any significance. But self-reported CL and performance metrics were highly sensitive to task difficulty. The impact of CL on perceived stress and the emotional state (Valence, Arousal, and Dominance) was rather small and only partly significant: Results indicate that CL has an impact on stress and emotional response, particularly when a high level of CL is induced. Furthermore, a general pattern was found that confirms a successful manipulation of low CL (*0-back*) and high CL (*2-back*), but the *1-back* condition seems to induce rather low than medium CL. The rather small effect of CL on pupil size change could indicate a common overestimation of the pupil dilation's sensitivity to cognitive processing in the research field. Hence, future work is highly recommended to gain further insights about VR Technology as a promising novel approach in CL research.

Keywords: Cognitive Load, Pupil diameter, VR technology, Stress, Emotions

Table of Content

List of Tables

List of Figures

List of Abbreviations

ACTH	Adrenocorticotropic Hormone
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
ANS	Autonomous Nervous System
ARET	Augmented Reality
BCPD	Inter-trial change in pupil diameter
CL	Cognitive Load
CLT	Cognitive Load Theory
CNS	Central Nervous System
CPD	Intra-trial change in pupil diameter
CRH	Corticotropin-Releasing Factor/Hormone
DLPFC	Dorsolateral Prefrontal Cortex
ECL	Extraneous Cognitive Load
ERPs	Event-related Potentials
ET	Exposure Therapy
fMRI	Event-related functional magnetic resonance
fNIRS	Functional near-infrared spectroscopy
GCL	Germane Cognitive Load
GSR	Galvanic Skin Response
H	Hypothesis/ Hypotheses
HPA	Hypothalamic-Pituitary-Adrenal
HR	Heart Rate
HRV	Heart Rate Variability
ICA	Index of Cognitive Activity
ICL	Intrinsic Cognitive Load
IPA	Index of Pupillary Activity
ms	Millisecond(s)
OFC	Orbitofrontal Cortex
PASA	Primary Appraisal – Secondary Appraisal
PNS	Peripheral Nervous System
RQ	Research question(s)
s	Second(s)
sAA	Alpha-Amylase
SAM	Self-Assessment Manikin
SAM	Sympathetic Adrenal Medullary
SCL	Skin Conductance Level
SMT	Stress Management Training
TEPR	Task-Evoked Pupillary Response
TSST	Trier Social Stress Test
VR	Virtual Reality
VRET	Virtual Reality Exposure Therapy

Introduction

In our globalized and complex world, we are confronted with increasing mental demands in our daily life. Especially advancing technologies gain more and more entry in our private and occupational environment and require advanced technical skills.

But these technological advances also offer new possibilities in facilitating our daily life, presuming that users are not overloaded in handling them. Therefore, it is essential to develop technical solutions to perform tasks successfully without excessive demands.

Research suggests that Cognitive Load (CL) has a fundamental, direct relation to human performance, such as successful task completion: Human performance increases with increasing task difficulty until the individual limit is reached (Chen, Zhou, Wang, Yu, Arshad, Khawaji & Conway, 2016). After that, human performance declines (e.g., in the form of error rates) and causes additional negative outcomes such as stress and negative emotions (Chen et al., 2016). Thus, CL reflects a reliable determinant of the performance of human-computer interaction. This is why developing precise measures of CL is of high interest in Human-Computer Interaction (HCI) research with the vision to develop an intelligent system that responds appropriately to the user's individual CL.

On the field, *Usability* evaluations commonly include retrospective and subjective-based methods to measure CL due to its low-cost and easy implementation (Lin, Omata & Imamiya, 2005; Lin & Imamiya, 2006). Hence, developing objective and *real-time* methods (outside the laboratory setting) are needed to ensure a holistic *Usability* assessment.

Providing reliable measures of CL is not only of interest in Computer Science. Since high CL inhibits learning and performance, assessing CL plays a crucial role in education, training design, and performance evaluations in several areas. For instance, the popularity of *Multimedia Learning* or *E-Learning* has grown immensely over the last decades, and so the desire to enhance the remote learning experience (Martin, 2005). A *real-time* assessment of the learner's current CL could help to depart more and more from the classical top-down learning, not only in a *Multimedia Learning* setting.

Furthermore, assessing CL has been integrated in medical education (e.g. Andersen, Mikkelsen, Konge, Cayé-Thomasen & Sorensen, 2016), aviation simulations (e.g. Lini, Favier, Hourlier, Vallespir, Bey & Baracat, 2012), academic education (e.g. Cranford, Tiettmeyer, Chuprinko, Jordan & Grove, 2014), maritime industry (e.g. Wu, Miwa & Uchida), and car-driving performance (e.g. Palinko, Kun, Shyrokov & Heeman, 2010). Especially in high-risk sectors, such as aviation or military, where human failure can cause severe consequences, examining the human's CL successfully is essential.

There is also a common interest in integrating CL measurement in clinical diagnostic to assess mental impairments in certain fields, such as neuropsychology, rehabilitation strategies, long-term drug abuse, or Schizophrenia, and others (Chen & Epps, 2013). Since CL is defined as a result of limited memory capacity during a task, it can offer insights into the patient's current mental abilities and support choosing an appropriate treatment.

In summary, there is a broad interest in measuring CL with an objective and precise method in research and on the practice field. Even though research has gained considerable progress regarding several physiological and *real-time* approaches over the last decades, methodical obstacles (e.g., eliminating confounding variables) need to be overcome until more valid methods can find their way out of the laboratory setting.

Until finishing this work, we have not known of a published study examining the potential of Virtual Reality (VR) Technology to provide less biased eye-related data to assess CL. This pioneer work aims to contribute first findings about integrating a VR environment into CL research. Additionally, examining the impact of CL on stress and emotions should provide a better understanding of the psychophysical constructs and their relation to each other.

Theoretical Foundations

Cognitive Load as a Psychophysical Construct

There have been different attempts, both overlapping and sometimes divergent definitions of Cognitive Load (CL). Chen and colleagues (2016) see CL "… as a variable that attempts to quantify the extent of demands placed by a task on the mental resources we have at our disposal to process information." (p.4). Another popular definition was contributed by Paas and Merrienboer (1994a): "… it is a multi-dimensional construct representing the load imposed on the working memory during the performance of a cognitive task." (p. 353). These definitions imply an interaction between task and learner characteristics and measurable constructs like Cognitive Load (Paas, Tuovinen, Tabbers & van Gerven, 2003). Hence, CL is highly dynamic and task-related. Among others, Paas and colleagues (2003) see task format, task complexity, time pressure, and task instructions among others, as task features. According to the authors, learners' characteristics include expertise level, age, and spatial ability.

Additionally, there have been similar but slightly different terms of cognitive processing. Paas and colleagues (2003) further differentiate between CL, *mental load,* and *mental effort*. They understand *mental load* as the element of CL that occurs due to the interaction between task and learners' characteristics. According to Paas and Merrienboer (1994a), *mental load* can be estimated by the a priori knowledge about the tasks' and subjects' characteristics and therefore provides a forecast of the subject's CL. In contrast to that, *mental effort* describes the actual invested mental resources to meet the demands of a task; thus, it reflects the actual CL (Paas et al., 2003). Other CL definitions emphasize the neurophysiological character. Just, Carpenter and Miyake (2003) define CL as "how hard a cognitive system needs to work to perform a given task." This underlines the relation between CL and attentional or working memory resources needed to meet task demands (Vogels, Demberg & Kray, 2018).

Another widely used term is *workload*. Hart and Staveland (1988) define *workload* as followed: "…is a hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance." (p.140). In contrast to CL, *workload* can be understood as a more general cognitive processing to accomplish a certain output that is performance-related, not task-related (Hart & Staveland, 1988).

Our CL understanding is in line with the above presented definition by Chen and colleagues (2016).

Cognition & Learning

Working Memory Models. CL research is based on cognition and learning findings since CL is induced by reaching the working memory capacity (Vogels, Demberg & Kray, 2018). Therefore, the following section illustrates the most important theoretical and empirical foundations of memory and learning.

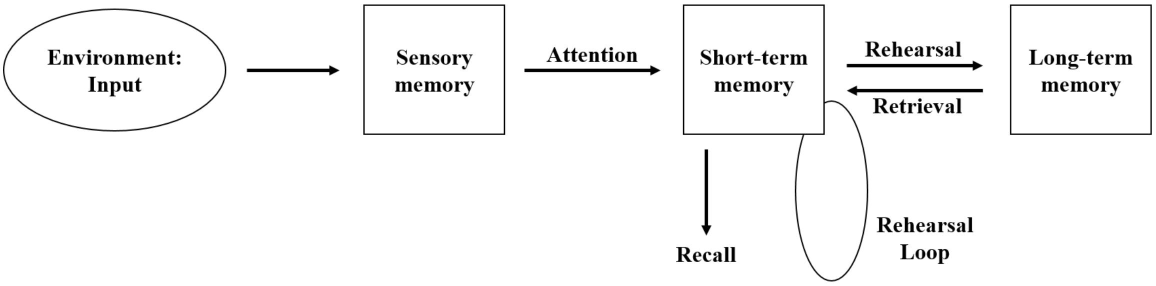


Figure 1. The *Multistore Model of Memory* by Atkinson & Shiffrin (1968).

One important theory was introduced by Atkinson and Shiffrin (1968): The *Multistore Model of Memory* imposes different process stages and explains how information is organized and stored in the human memory system.

First, external stimuli are detected by our sense organs and transferred to the sensory memory (also called *sensory register*). It involves a separate store for each sense and is rather an automatic, unconscious registration of the environment. It can register huge amounts of information but only for less than a second (Goldstein & van Hooff, 2018). Sperling (1960) investigated how much information from sensory memory is retrievable. He presented 12 letters for 50 milliseconds (ms) to participants and asked them to recall as many as possible right after. The participants reported 4.5 letters on average, which decreased with increasing difficulty conditions (Sperling, 1960). He concluded that the sensory memory registers all or most of the information that decays within less than a second (Goldstein & van Hooff, 2018; Sperling, 1960).

Only if certain stimuli get attention, this information passes on to the short-term memory. The unattended information gets lost, whereas the transmitted information in the short-term memory can be held about 20 seconds (Peterson & Peterson, 1959). Goldstein and van Hooff (2018) call the short-term memory as “the window of the present” and emphasize its importance in our daily life. Moreover, if a person continually repeats the elements, it is possible to keep them in the short-term memory for even longer. To explain this phenomenon, Atkinson and Shiffrin (1968) introduced the control process *rehearsal* (*loop*). But when rehearsing elements in the short-term memory, it inhibits the entry of new information, because the short-term memory is limited and has a capacity of seven chunks, plus or minus two, that can be stored simultaneously (Miller, 1956). According to the author, chunks are independent items of information. By chunking elements, for example, remembering a telephone number by double digits, the limited capacity of the short-term memory can increase (Miller, 1959). Besides rehearsal, there are other control processes to keep information in the short-term memory for longer, for instance, relating the content to a stored element from the long-term memory (Goldstein and van Hooff, (2018).

The well-known *recency effect* has been replicated in many studies and is widely seen as an indication for short-term memory. The *recency effect* describes the phenomenon that the last elements of a sequence can be recalled better than information located in the middle of the sequence (Murdock, 1962). The recent elements are more likely to be stored in short-term memory and, therefore, can be still remembered (Goldstein and van Hooff, 2018; Murdock, 1962; Daniel & Katz, 2018). Additionally, the *recency effect* disappeared or decreased when the time between presenting stimuli and recall was extended, or a distractor task was added (Glanzer & Cunitz, 1966; Nakajima & Sato, 1989; Postman & Phillips 1965).

The close exchange between the short-term and long-term memory is constant, due to the continuous comparison between current information in the short-term memory and stored information in the long-term memory (*retrieval*). This way, certain elements can be remembered or linked to certain experiences (Aktinson & Shiffrin, 1968). According to the authors, the longer a piece of information is held in short-term memory, the more likely a transformation to the long-term memory is. There, the information is primarily stored by its meaning (semantic format). The transfer from the short-term memory to the long-term memory is also called *encoding* (Goldstein and van Hooff, 2018). The long-term memory has an unlimited capacity, and information can be accessed permanently. Bahrick and colleagues (1975) introduced the *very long-term memory* because, in their studies, they could show that people can remember information even after a long time. For instance, after 15 years, participants still had a 90% success rate in identifying certain names and faces.

Goldstein and van Hooff (2018) further specify different types of long-term memory. The authors divide these into explicit (conscious) and implicit (unconscious) memory. According to Tulving (1985), explicit memories contain episodic and semantic memory. Episodic memories are individual experiences (personal events) in the past that involves *mental time traveling* (Goldstein & van Hooff, 2018). In contrast to that, semantic memory involves accessing general knowledge about the environment that is not tied to individual experiences (Tulving, 1985). According to Goldstein and van Hooff (2018), this knowledge covers facts, vocabulary, numbers, and concepts. Tulving (1985) sees semantic memory as *knowing* without *mental time travel*. Due to certain brain damage cases where either the episodic or semantic memory was impaired, this differentiation also has neuropsychological support (Rosenbaum et al., 2005; De Renzi et al., 1987).

In contrast to explicit memory that we are aware of, our memory system also contains unconscious information. The procedural memory includes all the information about learned skills; therefore, it is also called *skill memory* (Goldstein & van Hooff, 2018). Another form of implicit memory is priming, where a first stimulus (*priming stimulus*) affects the reaction to another stimulus (*test stimulus*). For instance, if the priming stimulus is the word *tree,* then it is likely that a person will respond quicker to a later representation of the word *tree* than to another word. This so-called *repetition priming* can be an unconscious process; this is why it is considered to be an implicit memory (Goldstein & van Hooff, 2018). Goldstein and van Hooff (2018) also see *classical conditioning* as implicit memory. *Classical conditioning* can be an unconscious process where two stimuli are paired: a neutral stimulus occurs simultaneously with a conditioning stimulus that causes a certain reaction. After some repetitions, the prior neutral stimulus induces the same reaction without the representation of the conditioning stimulus. This way, a pet expects food (conditioning stimulus) when it hears a bell ringing (prior neutral stimulus) if both stimuli appeared simultaneously before.

As discussed earlier, it is more likely to remember recent list elements than the ones before (*recency effect*). The also well-known *primacy effect* shows that the first elements of a sequence are also more likely to be recalled than units in the middle (Murdock, 1962). According to Rundus (1971), this effect appears because participants could still rehearse the first elements and thus transfer it into the long-term memory as they still had available cognitive resources at the beginning. So the *primacy effect* can be seen as an indication for long-term memory existence (Rundus, 1971; Postman & Phillips, 1965). The so-called *serial* *position curve* combines the *primacy* and *recency effect* and shows that memory is better for presented stimuli at the beginning and the end than elements in the middle of a presented list (Murdock, 1962). A recent study conducted by Daniel and Katz (2018) could even show that both *primacy* and *recency effect* do not only appear when using visual or auditory stimuli. The authors could demonstrate that first and last presented tastes could be remembered better by participants than the ones in the middle. Interestingly, this also happened despite differences in liquid concentrations and delay lengths (Daniel & Katz, 2018).

Baddeley and Hitch (1974) introduced the term working memory instead of short-term memory. Even though these terms are usually used as synonyms, Baddeley, and Hitch (1974) intended to distinguish between the working memory and the short-term memory defined by Atkinson and Shiffrin (1968). Whereas the latter emphasizes the storage role of the short-term memory, Baddeley and Hitch (1974) amplified this concept by highlighting that the working memory also manipulates information during complex cognitive processes (i.e., remembering numbers while reading). According to the authors, the *Multistore Model of Memory* (Atkinson & Shiffrin, 1968) does not take the dynamic processes into account and, therefore, propose the more suitable term working memory instead of short-term memory. Questioning the function of the short-term memory got started as Baddeley and Hitch (1974) noticed that people could carry out two tasks simultaneously, which stands in contradiction with the model introduced by Aktinson and Shiffrin (1968). Baddeley and Hitch (1974) concluded that there must be autonomous components of the working memory that make multitasking possible.

Thus, Baddeley and Hitch (1974) introduced the autonomous components *Phonological Loop*, *visuo-spatial sketch pad,* and the *central executive* (or *control executive*). The assumption of having separate coding channels for verbal and nonverbal information is originally derived from Paivio's *Dual Coding Theory* (Paivio, 1986). The *Phonological Loop* consists of two sub elements that process verbal material and language: The *Phonological store* that holds information within a limited capacity of a few seconds and the *Articulary Control* (or *Articulatory Rehearsal Process*), which is activated when a person repeats information (i.e., phone number) and that way is kept in the *Phonological store* (Baddeley & Hitch, 1974). Some empirical findings support the existence of the *Phonological Loop*.

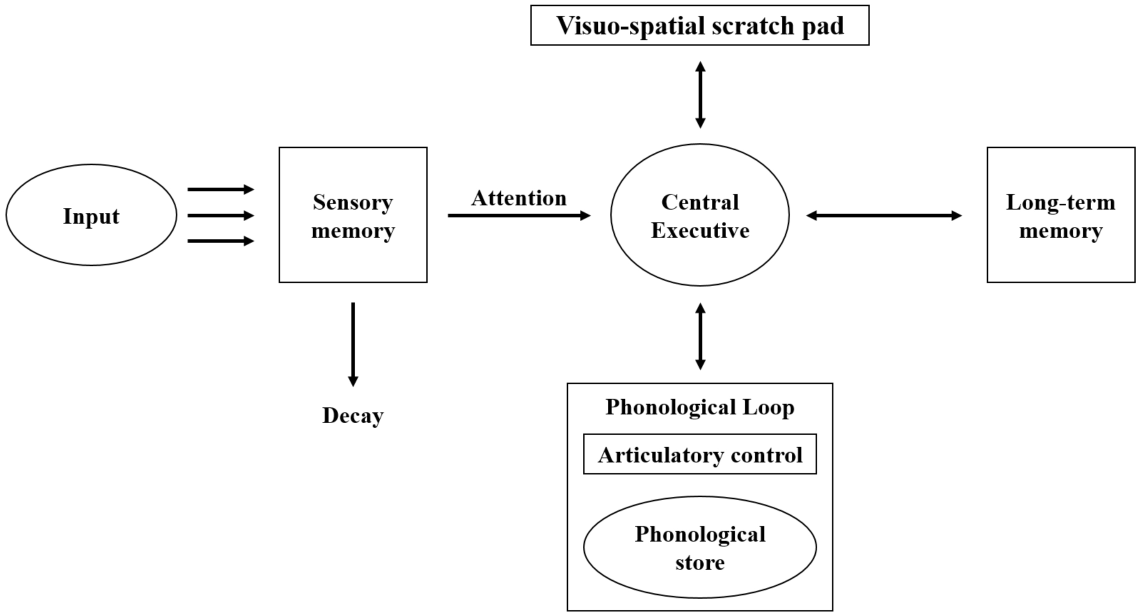


Figure 2. The extended version of the *Multistore Model of Memory*. Adapted from Baddeley & Hitch (1974).

Conrad (1964) demonstrated with the *phonological similarity effect* that people tend to mix up letters or words that sound familiar instead of familiar-looking letters or words. Additionally, the *word length effect* shows that we are better at remembering short than long words (Baddeley, Lewis & Vallar, 1984). The authors' argument is that this phenomenon occurs because it takes more time to rehearse and recall longer words in the *Phonological Loop* than short ones.

Another phenomenon shows that if a person is prevented from rehearsing due to a task of repeating an irrelevant sound, he or she remembers less (Baddeley, Lewis & Vallar, 1984; Murray, 1968). According to the authors, this so-called *articulatory suppression* reduces memory because speaking interferes with the rehearsal process and hence is also seen as evidence in favor of the working memory, introduced by Baddeley and Hitch (1964).

The *visuo-spatial sketch pad* holds visual and spatial information. This helps us to orient ourselves in the environment and allows visual imagery (Goldstein & van Hooff, 2018). *Mental Rotation* is widely seen as an example of how the *visuo-spatial sketch pad* operates. Shepard and Metzler (1971) conducted one of the first experiments testing *Mental Rotation*. The authors presented rotated three-dimensional objects, and subjects had to decide whether they showed the same object or not. The more they rotated the objects, the longer the subjects needed to decide (Shepard & Metzler, 1971). According to the authors, this phenomenon occurs because participants solve the task by rotating the objects in their mind (visual representation).

Furthermore, neuroscience suggests that visual and spatial information is processed in separate systems: Klauer and Zhao (2004) could demonstrate in a *dual-task* experiment that a visual short-term memory task was stronger impaired by a secondary visual task than by a spatial exercise and vice versa. This finding is in accordance with the widely accepted differentiation between the dorsal (“where/how”) and ventral (“what”) stream in neuroscience research. The *two-streams hypothesis* suggests that visual information is processed along two routes in the brain: The dorsal stream (from the primary visual cortex to the parietal lobe) processes information regarding spatial vision and controlled action whereas the ventral stream (from the primary visual cortex to the temporal lobe) is responsible for the identification of objects (Goodale & Milner, 1992; Teixeira Ferreira, Ceccaldi, Giusiano & Poncet, 1998; Wang et al., 1999; Ungerleider, Courtney & Haxby; 1998).

The *central executive* (or *control executive*) represents the center of the working memory. It pulls information from the long-term memory and manages the division of capacities between the *Phonological Loop* and *visuo-spatial sketch pad*. Therefore, Goldstein and van Hoff (2018) call it *traffic controller*. There has been some empirical research on the *central executive*. Vogel and colleagues (2005) could demonstrate that individuals who were better at filtering out irrelevant information, by using so-called *event-related potentials* (ERPs), had larger working memory capacities (Vogel, McCollough, & Machizawa; 2005). This "filter efficiency" (which can reflect the *central executive*) shows that the working memory capacity not only depends on how many items (cf. Miller, 1956) but how efficiently irrelevant information can be filtered out (Vogel et al., 2005).

In 2009, Baddeley and colleagues added a fourth component to their model due to some neuropsychological findings that show memory effects that last too long for the working memory and too short for the long-term memory. The additional *episodic buffer* provides extra capacity and communicates with the long-term memory just as the *Phonological Loop* and *visuo-spatial sketch pad*. But it has to be mentioned that this fourth component is still based on theoretical assumptions, and empirical evidence is needed to confirm the modified model (Baddeley at al., 2009).

Cognitive Load Theory. The Cognitive Load Theory (CLT) introduced by Sweller and colleagues (1998) is a learning and instruction theory and usually forms the theoretical basis for CL research. The theory itself is based on research on human cognitive architecture (Paas & Sweller, 2014). In line with several CL definitions, working memory capacity plays an important role in learning and knowledge acquisition. According to Sweller and colleagues (1998), CL is the result of three elements that have an impact on the working memory capacity (*Figure 3*). *Intrinsic Cognitive Load* (ICL) equates the perceived complexity of the content. Regarding ICL, the authors introduce the *element interactivity,* which describes the dependence between different content elements (Kalyuga, 2011; Sweller et al., 1998). That way, high *element interactivity* (elements cannot be learned independently) increases ICL, because the elements have to be held in the working memory simultaneously. ICL is not dependent on the learning environment, i.e., instructions, but varies among learners due to intra-individual differences in prior knowledge and cognition (Sweller et al., 1998). If the learner has prior knowledge about the content, he or she might combine certain elements, or some are already represented in the long-term memory. This results in lower *element interactivity;* thus, the person probably has a lower ICL. Research could demonstrate that by comparing expert chess players with novices. The experts could remember greater numbers of figures from real game situations than the beginners due to their prior experience and therefore, a higher ability to form larger chunks (De Groot, 1965; Chase & Simon, 1973). Kalyuga (2011) adds that even though ICL depends on the learners' expertise, it is possible to reduce ICL by simplifying the task as reducing the presented number of elements.

The *Extraneous Cognitive Load* (ECL) emerges from the cognitive effort that is needed to process the visualization and structure of the content. A high ECL can be induced by suboptimal designs that include redundant or unnecessary elements. For instance, if a design includes a lot of text elements and the learner needs to read all this information before working on the task, the person probably has a high ECL due to the reading activity, irrespective of the perceived complexity of the task (ICL). This evokes additional demands on working memory (Kalyuga, 2011). The author names the *split-attention eff*ect (when distributed attention is needed to process dependent information simultaneously) and *redundancy effect* (the same information is presented through different modalities and creates unnecessary loads) as examples for high ECL. Interestingly, a study trying to differentiate between the CLT elements conducted by Sweller (1994) shows a successful manipulation (and thus evidence) of ECL but only when the ICL was high, not low.

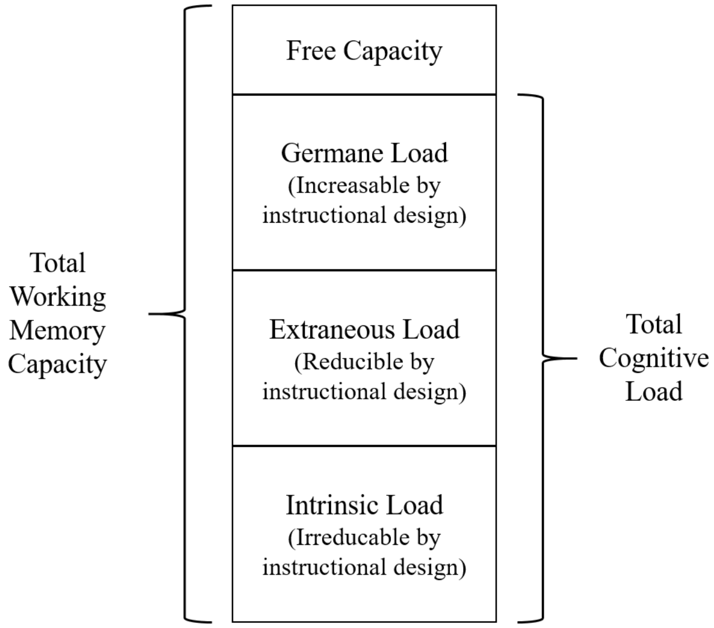

Free Capacity
Germane Load
(Increasable by instructional design)
Extraneous Load
(Reducible by instructional design)
Intrinsic Load
(Irreducable by instructional design)
Total Working Memory Capacity
Total Cognitive Load

Figure 3. Triarchic model of *Cognitive Load Theory*. Adapted from Moreno & Park (2010).

The third CL element is the beneficial *Germane Cognitive Load* (GCL) that emerges by conducting exercises to consolidate the content. The GCL aims to process the content in a deeper way by transferring it into long-term memory. In contrast to ICL, ECL and GCL can be influenced by instructions, design, or exercises, for instance. To get the maximum learning effect, one has to minimize the ECL by an optimal representation of the content. This way, there are more working memory capacities available for the GCL to effectively consolidate the learned content.

In summary, one should try to minimize ECL while ICL should be properly managed (neither too easy nor difficult) to have sufficient capacities that can be put into the GCL to enhance long-term storage. According to the CLT, this way, one can ensure the optimum learning experience (Kalyuga, 2011). Hence, ICL and GCL represent learning facilitators, whereas ECL reflects a learning inhibitor (Kalyuga, 2011).

Kalyuga (2011) suggests a modification of the CLT introduced by Sweller and colleagues (1998). He postulates that the differentiation between the ICL and GCL is based on theoretical assumptions and emphasizes the challenge to isolate both constructs empirically. Whereas research has found robust effects reducing ECL and ICL, there exist way less established techniques to manipulate the GCL (Kalyuga, 2011). The few findings supporting the existence of GCL can also be explained by an increased ICL or by varying definitions of CL constructs. For example, several studies are using similar subjective ratings of learning difficulty to measure different CLT elements (DeLeeuw and Mayer, 2008; Schwonke et al., 2011; Gerjets et al., 2009, Cierniak et al., 2009). According to the author, this lack of consistent research may reflect redundancy and overlapping definitions within the CLT framework (Kalyuga, 2011). The author claims that a differentiation between ICL and ECL is sufficient and might be more transparent regarding the limitations of the theory (Kalyuga, 2011). Due to this controversial background, we consider the CLT as the theoretical basis for our study, but we claim only to measure the general CL within our experiment.

Mayer's Cognitive Theory of Multimedia Learning. In 2005, Mayer introduced his *Cognitive Theory of Multimedia Learning*. If not stated otherwise, his work published in *The Cambridge Handbook of Multimedia Learning* (2005) forms the basis for the following section.

Mayer (2005) included effects identified by CLT and well-established memory research (Sweller, Ayres & Kalyuga, 2011). Mayer (2005) defines *Multimedia Learning* as creating mental representations through text and pictures. Pictures can be static or dynamic in forms of videos. His theory is based on three principles: First, the information processing system has two channels for handling visual/pictorial and auditory/verbal information, which corresponds with the *visuospatial sketch pad* and *Phonological Loop* introduced by Baddeley and Hitch (1974). Consistently with widely-accepted memory models, he also emphasizes the limited capacity of both channels.

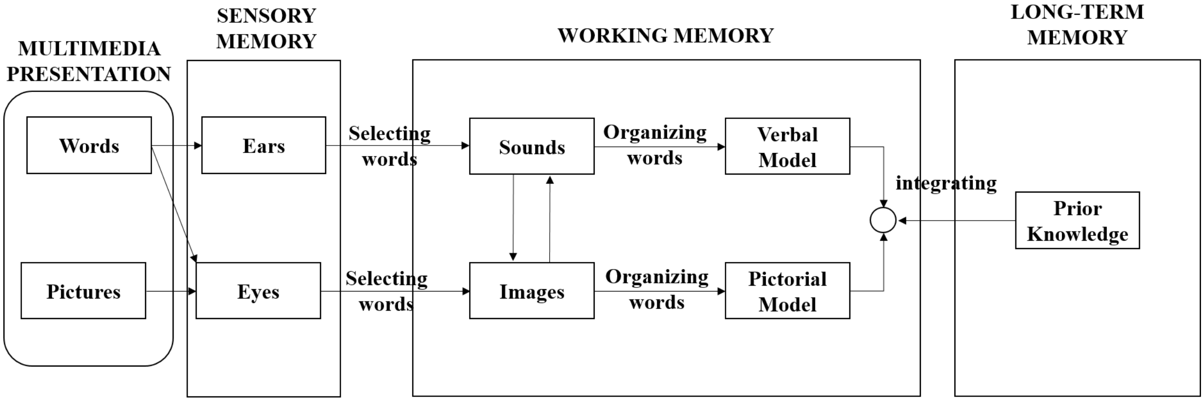


Figure 4. The *Theory of Multimedia Learning.* Adapted from Mayer (2005).

Additionally, he states that learning requires the activation and coordination of several cognitive processes. Mayer (2005) defines these cognitive processes that are involved in *multimedia learning*. It starts with selecting important content (sensory memory) from presented words or pictures to transfer it to the working memory. Then, the selected information has to be mentally organized into a coherent cognitive structure in the working memory. Finally, the organized content has to be compared and integrated with relevant prior knowledge from long-term memory. Similar to the CLT, he specifies three demands on the limited capacity during learning: *extraneous processing* (ECL), *essential processing* (ICL), and *generative processing* (GCL). According to Mayer (2005), the instructional goal is to ensure appropriate cognitive processing during learning without overcharging the learner.

He also investigated *multimedia learning* and established the *modality effect* that indicates that better learning takes place when using text and pictures than only pictures (Ayres, 2015; Mayer, 2005). Even though several studies support the *modality effect*, Mayer and Pilegard (2014) found out that too many modalities inhibit learning if the content is redundant (*redundancy effect*).

Even though the *Cognitive Theory of Multimedia Learning* has been widely accepted, Ayres (2015) criticizes that research supporting Mayer's theory has not included enough variation of different multimedia factor conditions, hence it is not clear if certain multimedia environments add more or less value than others (Ayres, 2015; Nye, Graesser & Hu, 2014). Moreover, Ayres (2015) argues that the supporting studies were conducted in a laboratory setting, hence not representative of learning in real life. See Ayres' review of *multimedia learning* (2015) for more details on *Mayer's Cognitive Theory of Multimedia Learning*.

Summary. There has been extensive memory research conducted over the last decades. This led to a strong rise of modified models that explain mechanisms of the whole memory system more and more. Even though there are different views on certain elements and their structure, the fundamental assumption of a short and long-term memory is indisputable. Especially the limited capacity of short-term memory is confirmed by several independent studies, which forms the basis for CL research. Despite new technologies to gain deeper neurophysiological insights, there are still methodical obstacles to be overcome to confirm or refuse widely-accepted memory models.

Measurement of Cognitive Load

There have been several approaches to measure CL. Generally, one can differentiate between objective and subjective instruments. Objective methods have the advantage of a *real-time* measurement, but due to the psychophysical data, these methods underlie confounding variables, for instance, uncontrollable light conditions. This is not the case for self-report questionnaires that are easily implemented, but one has to keep in mind that this data is retrospective and based on the individual's perception (Martin, 2015; Lin, Li, Wu, & Tang, 2013). Therefore, a combination of both is recommendable until there are better instruments developed and established (Martin, 2015).

Chen and colleagues (2016) specified four main domains regarding CL measurement in HCI and other domains: Subjective (self-report), performance, physiological, and behavioral methods. In this work, we adopt the categorization introduced by Chen and colleagues (2016), but it has to be noted that the borders between the categories are permeable; some methods comply with the requirements of more than one category. In the following, the most popular measurement instruments for each domain are illustrated.

Subjective Measures. Subjective methods rely on self-reported evaluations and are widely used in CL research. These metrics reflect the subjects' perception of CL and require an immediate self-assessment after a task (Chen et al., 2016). Further, the authors divide subjective measures into unidimensional and multidimensional scales. An example of a unidimensional scale is the *mental effort rating scale* by Paas (1992) that consists of one item with nine gradations, where subjects reported their invested mental effort.

Multidimensional scales include several components of CL. One of the most used instruments is the NASA task load index (TLX) questionnaire (Hart & Staveland, 1988) and, therefore, is discussed more in detail in this work. The authors included six dimensions (*Performance, Mental Demand, Frustration, Effort, Physical Demand,* and *Temporal Demand*) that form the basis for the TLX score (Hart & Staveland, 1988). The *Performance* dimension determines how well the person performed. *Effort* describes the individual mental cost to achieve this performance. *Mental Demand* assesses how easy or demanding the task was. *Frustration* covers the affective impact, thus how irritated, stressed, or content and relaxed the subject feels after the task. *Physical Demand* determines how much physical activity was required. Lastly, the *Temporal Demand* describes CL induced by time pressure during the task. After completing the questionnaire, subjects have to weight which member of each paired combination of the six dimensions is more related to their workload definitions. Each subscale rating (with 20 gradations each) is then multiplied by the chosen weight and divided by 15 to get the final overall TLX score between 0 and 100 (Hart & Staveland, 1988; Hart, 2006). The authors point out that this weighting component increased the questionnaire sensitivity and decreased between-rater variability. Nevertheless, the most common modification over the last decades has been leaving out the weighting process to simplify its use (Hart, 2006): The so-called RAW TLX identifies the overall score by averaging or adding up the ratings without the individual weighting. Hart (2006) states that there have been findings indicating that the RAW TLX was either more, less, or equally sensitive in comparison to the overall TLX and leaves the decision to the researcher which method to pick. Another popular modification is to leave out or add dimensions because there are irrelevant or relevant to the chosen task (Hart, 2006). The author supports this approach but emphasizes the importance of reviewing retesting reliability, sensitivity, and validity.

The TLX score correlates with error rates in complex socio-technical domains and has been used in a wide range of different research areas (Colligan, Potts, Finn, & Sinkin, 2015; Grigg, Garrett & Benson, 2012; Hart, 2006). Stapel, Mullakkal-Babu, and Riender (2019) used the NASA TLX to determine drivers' mental workload, and results show that the perceived workload increased with traffic complexity. William (2017) could show that multimedia instructions that used visual cues reduced the NASA TLX, hence subjective CL. Another study, conducted by De la Torre, Ramallo, and Cervantes (2016), used a modified version of the NASA TLX to assess mental demand during drone flight simulation tasks.

Like Hart (2006) suggests, De la Torre, Ramallo, and Cervantes (2016) tested statistic quality criteria successfully before implementing the modified NASA TLX. The authors identified the subscale *Mental Demand* as the best indicator for workload during the training tasks and saw it as a potential measurement to improve remote pilots' skills and training (De la Torre, Ramallo & Cervantes (2016). The NASA TLX also finds its relevance in the maritime industry. Kim, Yang, Lee, Yang, and Hong (2007) examined the effect of alcohol intake on workload (perception) during a ship navigation simulation. Increased alcohol intake impaired the performance, which also resulted in an increased perceived mental workload, assessed by the NASA TLX (Kim at al., 2007). The NASA TLX is also used in medical research. Felton, Williams, Vanderheiden, and Radwin (2012) conducted a study assessing the mental workload using a brain-computer interface (interface control without extremities) for physically impaired people. The authors conclude that NASA TLX is an effective tool to compare the workload between different groups and tasks (Felton at al., 2012).

Due to the simple, non-intrusive application and evaluation, subjective measures are usually included in CL research. Questionnaires like the NASA TLX are well-established and can be used in several areas. But they are also some issues regarding subjective methods one has to consider when integrating them into the study design. In general, using multidimensional scales instead of a unidimensional scale is recommendable. Unidimensional scales are criticized because CL is only measured by one item of perceived difficulty; nevertheless, it is helpful to interpret it as an indicator of an overall CL (Debue & van de Leemput, 2014). Further, it is important to mention the retrospective character of subjective measures since the data is collected after the CL manipulation. This missing *real-time* collection implicates that participants have to remember CL, which can lead to biases. Another general issue regarding questionnaires is the sensitivity to several confounding variables. For instance, *social desirability* is a well-known phenomenon showing that the desire for social acceptance influences participants' answers (e.g., Wheeler, Gregg & Singh, 2019). Additionally, there have been identified certain response patterns of participants. For instance, the tendency to use more "neutral" item gradations in the center than extreme values (*error of central tendency*) or the impact of item order (*order effect*) that can lead to different responses (e.g., Yu, Albaum & Swenson, 2003; Cochran, 2018).

Hence, researchers have to keep in mind that subjective measures only assess the individual perceived CL that can include biases. Nevertheless, it is a simple method, and depending on the research question, subjective evaluations can add more or less value than other types of measures.

Performance Measures. Chen and colleagues (2016) define performance measures as measures that can explain individual variations occurring during a task. Performance Measures are based on the assumption that learning, thus performance, is measurably inhibited when the working memory capacity is overloaded and hence, an indicator for CL (Paas & Merrienboer, 1994a). One of the most established methods is the *dual-task paradigm*. Here the subject has to accomplish two tasks simultaneously. The *Primary Task* represents the main task, and the measured performance of a *Secondary Task* equates the available working memory capacity while solving the main task. Hence, a good performance on the *Secondary Task* means that the main task does not induce much CL, and there still are enough cognitive capacities to perform well on the *Secondary Task*. Conversely, if the main task is already causing a high CL, the performance of the *Secondary Task* will decrease due to a lack of cognitive resources. According to Khawaja (2010), task completion time, speed or accuracy, error rates or false starts are examples for dependent variables in combination with *dual-task* manipulations.

There have been different approaches to implement *dual-task* settings. Vogels, Demberg, and Kray (2018) included two *Primary Task*s (either language comprehension or driving simulation task) and combined it with a memory task as a *Secondary Task*. They compared these tasks as single and dual tasks. The pupillary response indicated that the CL increased much more when adding a *Secondary Task* than completing the tasks independently (Vogels, Demberg, and Kray, 2018). Park and Brünken (2015) tested foot tapping as a novel *Secondary Task* while learning (*Primary Task*). The authors see rhythm precision as a new potential way of measuring CL since it decreased with increasing learning difficulty (Park & Brünken, 2015). Another study used the *dual-task paradigm* to test training effects in a virtual reality (VR) surgical simulation. Rasmussen, Konge, Mikkelsen, Sorensen, and Andersen (2015) let participants perform an advanced medical procedure (*Primary Task*) and a visual monitoring task (*Secondary Task*) in the VR environment. During the medical simulation, the *Secondary Task* precision decreased significantly; unfortunately, training did not affect its precision (Rasmussen et al., 2015). Further, Karatekin, Couperus, and Marcus (2004) used auditory stimuli for the primary memory task. Participants had to listen to digit sequences and remember them as complete as they could. In the dual condition, subjects simultaneously had to respond quickly to a small symbol randomly appearing on the screen (Karatekin, Couperus & Marcus, 2004). Hence, the authors used performance measures (accuracy and reaction time) to determine CL.

Even though the *dual-task paradigm* is applied globally, there has been a methodical critique among researchers. O'Donnell and Eggemeier (1986) claim that the method can be used to detect medium and high CL but is not sensitive to low cognitive effort. Fisk, Derrick, and Schneider (1986) further argued that many chosen *Secondary Task*s evoke learning effects, resulting in an "automatized" performance. For this reason, the authors define the criterion that the *Secondary Task* should require effortful processing throughout the experiment. Moreover, Fisk and colleagues (1986) state that people have the ability to trade-off their performance due to a controlled distribution of cognitive capacities within tasks (this would conform to Baddeley and Hitch, 1974). In contrast to the *dual-task paradigm,* subjects can derive resources from the *Primary Task* (Fisk et al., 1986). As a result, the authors suggest taking both single and dual-task *Primary Task* performance into account. Thirdly, Fisk, and colleagues (1986) criticize that many experimental designs include widely used *Secondary Task*s without considering their fit for the *Primary Task*. Therefore, both tasks must demand the same mental resources; for instance, two visual tasks (Fisk et al., 1986; Martin, 2015). Martin (2015) additionally points out that it is relevant to choose an appropriate difficulty level for the *Secondary Task* because if too difficult, it may become the *Primary Task*, and if too easy, both tasks are accomplished successfully and hence, not sensitive to low CL.

Besides *dual-task* approaches, several studies used single tasks to induce CL. Since CLT is based on the assumption that working memory is limited, researchers have also used working memory tasks to manipulate CL. Guastello and colleagues (2015) chose the *n-back* working memory task with both auditory and visual stimuli simultaneously to induce workload and fatigue, because according to the authors, this task imposes heavy cognitive processing on participants. When conducting the *n-back task*, subjects have to compare the present letter with a letter *n* steps back. The more steps back, the more CL is induced (Guastello et al., 2015). Interestingly, there are different understandings among researchers about how many steps of the *n-back* task reflect low or high CL. For instance, Zuo, Salami, Yang, Yang, Sui, and Jiang (2019) define the *2-back* condition as high CL, whereas Moore, Eccleston, and Keogh (2017) used the *2-back* condition as low CL manipulation and added the *3-back* task to induce high CL. Besides the *n-back task*, Moore, Eccleston, and Keogh (2017) integrated two more working memory tests: An attentional switching and divided attention task. Attentional switching tasks include switching between tasks that impairs performance more than task repetition (Moore, Eccleston & Keogh; 2017). First, participants had to memorize a list with either two (low CL) or five (high CL) letters. After that, a random letter was shown one at a time on the screen. If the letter was black, the participants had to decide whether it was on the list or not. If the letter was green or red, they had to "decide" whether the letter was green or red. Performance measures (reaction time and accuracy) were used to estimate CL. As predicted, higher CL reduced performance on the task (Moore, Eccleston & Keogh; 2015). When performing the divided attention task, participants had to press a button when two randomly located digits were either a 0 or 5. In two further conditions, they had to either memorize a list of three or seven items before starting the digit task. Results show that when remembering the list the reaction time was significantly higher when performing the digit task before (Moore, Eccleston & Keogh; 2015). Wilson and Russel (2003) assessed CL via EEG Technology and induced mental effort by applying two difficulty conditions of the *Multi-Attribute Task Battery* (MATB). The MATB consists of different simultaneous tasks inspired by real aircraft challenges (Wilson & Russel, 2003). Another simpler approach is presented by a study conducted by Pecchinenda and Petrucci (2016). They manipulated CL by either counting backward by seven (high CL) or counting forwards by two (low CL). Chong, Mills, Dailey, Lane, Smith, and Lee (2010) also defined counting backward by seven and generating words with the same first letter as their cognitive tasks and analyzed whether the CL manipulation had an impact on physical balance control. Data suggest that only the subtraction task impaired balance control as a result of assumed shared cognitive capacities (Chong et al., 2010).

Performance measures, especially the *dual-task paradigm*, are popular methods to asses CL. In general, there are studies where only performance indicators are used to determine mental effort directly (e.g., Moore, Eccleston & Keogh, 2015) or studies using performance tasks to only induce CL but apply additional measures to assess the workload (e.g. Karatekin, Couperus & Marcus; 2004; Wilson & Russel, 2003). The latter is recommendable because raw performance indicators can be the same between subjects but could be achieved by different mental effort (Khawaja, 2010). Chen and colleagues (2013) also integrated performance indicators in their study but emphasized that these measures alone are no evidence for invested mental capacity. Thus, it is suggested to use performance tasks as independent variables, and further methods (e.g., physical or subjective) as dependent variables. Another important aspect when including performance tasks in CL research is to address the mentioned methodical issues, for instance, the necessary "sense" match in *dual-task* approaches, so the same working memory networks are addressed (Fisk et al., 1986; Martin, 2015).

Physiological Measures. The physical approach as a CL measurement is based on cognitive processes that have an impact on human physiology, amongst other cardiovascular responses that have been found to be sensitive to task difficulty (Kramer, 1991; Carroll, Turner & Prasad, 1986). Some studies have used heart rate (HR) / pulse and heart rate variability (HRV) to measure CL (Mulder, 1992; Kennedy & Scholey, 2000; Nickel & Nachreiner, 2000). For instance, Turner and Carroll (1985) could demonstrate an increased HR during a mental arithmetic task. In opposition to this, other studies found HRV to be intrusive, invalid, and insensitive to fluctuations in cognitive load (e.g., Nickel & Nachreiner, 2003, Paas & Van Merrienboer, 1994b).

Besides cardiovascular parameters, there have been attempts to measure CL with the brain's electrical activity (Chen et al., 2016). The continuous Electroencephalography (EEG) signal consists of oscillations in several frequencies that are assumed to correspond with information representation and transfer within the neuronal network (Antonenko, Paas, Grabner & van Gog, 2010). Other studies also show that EEG Technology, usually changes in alpha and theta wave rhythms, is sensitive to task difficulty manipulations (Gevins & Smith, 2003; Klimesch, Schack & Sauseng 2005; Huang et al., 2013; Zhao & Yao, 2017).

Wilson and Russell (2003) report an 85% success rate of correctly classifying high mental workload using EEG Technology. Also, there have been attempts to measure CL using event-related functional magnetic resonance (fMRI) that detects the amount of hemodynamic activity (oxygen saturation and blood flow) in neuronal regions (Martin, 2015). Zuo and colleagues (2019) used the fMRI technique to detect neuronal networks that are activated during CL. Results suggest that the frontoparietal executive control network, dorsal attention network, and salience network are activated during high CL (Zuo et al., 2019).

Functional near-infrared spectroscopy (fNIRS) represents an alternative to the sensitive but bulky and highly expensive EEG and (f)MRI Technologies (Martin, 2015; Fishburn, Norr, Medvedev & Vaidya, 2014). Fishburn and colleagues (2014) could show that fNIRS Technology is sensitive to CL using the *n-back* working memory task and thus, see this neuroimaging technique as an alternative to fMRI use.

Further, *Galvanic Skin Response* (GSR) or *Skin Conductance Level (SCL)* is a measure of the conductivity of human skin and can imply changes in the human sympathetic nervous system (Shi, Choi, Ruiz, Chen & Taib, 2007). The authors could demonstrate that the mean GSR increased as the induced CL increased. Nourbakhsh and colleagues (2012) could show successfully that the mean and accumulative GSR signal could distinct between CL levels using text reading and arithmetic tasks.

One of the most used physical measures are eye-related parameters for CL. The most established psychophysical parameters among eye-tracking studies include fixation and saccade patterns, blink rates, and pupil dilation during cognitive processing (Zagermann, Pfeil & Reiterer; 2016). Fixations describe the eye focusing on a particular point and last about 100 to 1000 ms (Chandra, Sharma, Malhotra, Jha & Mittal, 2015). Fixations happen when the eye appears to be relatively stable processing information. Commonly used metrics are fixation duration, fixations per area of interest, number of fixations overall, fixation spatial density and repeated fixations (Chandra et al., 2015; Chen & Epps; 2013). Saccades are quick movements between two fixations, usually lasting about 20 to 35 ms. Fixations are often used to indicate the difficulty of particular information, whereas saccades imply difficulty in locating target stimuli (Chen, Epps & Chen, 2013). The authors confirm similar research results that fixation duration can be used as an indicator of perceptual load.

Regarding a study dealing with a target identification memory task, the authors found out that fixation frequency and saccadic extent reflected changes in task difficulty (van Orden, Limbert & Makeig, 2001). All in all, one has to take into account that these eye movement metrics are not applicable for all common CL manipulations, especially when the task requires the eye to focus on a certain point.

Blink activity is another well-known indicator for CL. According to Chen and Epps (2013), this measurement can be seen as a behavioral or psychophysical response because blinks can be both voluntary (behavioral) and endogenous (psychophysical). The latter is the case for most blinks that occur two to four times per minute in a normal state (Irwin & Thomas, 2010). For that reason, blinking activity is categorized here as a physical measurement. Most research focuses on the blink rate because blink amplitude and duration do not seem to be reliable indicators for CL (Tanaka & Yamaoka, 1993).

Concerning CL, several studies found out that the blink rate decreases with increasing task difficulty (Irwin & Thomas, 2010; Ledger, 2013). Irwin and Thomas (2010) argue that blinking is reduced to maximize stimulus perception during cognitive processing. On the contrarily, there are also research findings that suggest an increased blinking rate with increased task demand (Chen & Epps, 2013; Tanaka & Yamaoka; 1993). Moreover, Stern, Walrath, and Goldstein (1984) reported a blink boost at the start and end of a cognitive process. Van Orden and colleagues (2001) suggest that these conflicting results are due to different tasks used in these studies. They assume that by task-induced saccadic eye movements (e.g., visual search tasks), there is also a need for an increased blinking rate and blink durations to accomplish the task (van Orden, Limbert & Makeig, 2001). More in detail, Chen, Epps, and Chen (2013) point out that the blinking rate is inhibited in tracking tasks and increases in conversational and arithmetic tasks. More comparable research is needed to capture the mechanisms behind endogenous blink activity to further interpret the mixed results.

Pupil dilation or change in pupil size is one of the most used methods to measure CL. The adaptive function of the pupil's diameter is to regulate the light that enters the eye and to control the depth of the visual field (Beatty & Lucero, 2000; Chen & Epps, 2013; Kramer, 1990). The pupil's diameter can vary from two to eight mm and is controlled by antagonistic muscles in the iris that contain the muscle groups *dilator pupillae* and *sphincter pupillae* (Kramer, 1990): The former causes a retraction of the iris, which leads to an increased pupil size and the latter expands the iris, hence reduces the pupil size.

Studies conducted over the last decades have shown the robust link between pupillary changes and perceptual (Chen, Epps & Chen, 2013; Beatty; 1988), cognitive (Chen, Epps & Chen; 2013) and response related demands (Richer & Beatty, 1985; Ikehara & Crosby, 2005). So it is not surprising that several studies indicate that an increase of the pupil diameter indicates increasing cognitive processing (e.g., Chen & Epps, 2013; Krejtz, Duchowski, Niedzielska, Biele & Krejtz; 2018; Pomplum & Sunkara, 2003). This so-called *Task-Evoked Pupillary Response* (TEPR) has been observed among cognitive tasks including arithmetic, driving simulation, memory, and visual search tasks (Wang, Li, Wang & Chen; 2013). Krejtz and colleagues (2018) further distinct between inter-trial change in pupil diameter (BCPD) and intra-trial change in pupil diameter (CPD): The former uses the average pupil diameter during a baseline trial whereas the CPD computes a baseline measurement made at the beginning of each trial. In their experiment, both could significantly discriminate between task difficulties, whereas the BCPD provided a clearer interpretation and showed a better effect size than CPD (Krejtz et al., 2018). Moreover, some studies could observe a decrease in pupil dilation towards the end of a task (and rises when beginning a new task), which Porta and colleagues interpret as a sign of tiredness (Porta, Ricotti & Perez; 2013; Iqbal, Adamczyk, Zheng & Bailey; 2005).

Even though pupil dilation is one of the most used parameters to detect CL, there are some confounds that one has to take account of when conducting research. As explained above, the light and near reflex are the main functions of the accommodation of the pupil (Chen & Epps, 2013; Kramer, 1990). Other study results suggest that the light and near reflex evoke even bigger changes in pupil size than mental processing (Kramer, 1990; Pomplum & Sunkara, 2003). Additionally, the pupil diameter underlies irregular changes, which are called *pupillary hippus* or *pupil unrest* that happen independently of illumination or eye movements (Beatty & Lucero, 2000; Stark & Campbell, 1958). This complicates CL research because researchers have to control the ambient light and take it as a potential confound into account when interpreting results. In 2002, Marshall (2002) proposed an approach to address this issue. In contrast to the common baseline-related statistics, the *Index of Cognitive Activity* (ICA) measures the rate of change and not the difference between averaged pupil diameters in a resting state and under cognitive demands (Duchowski et al., 2018; Marshall, 2002). The ICA score ranges from 0 to 20 Hz, whereas a low value represents little mental effort and high values suggest high mental effort (Duchowski et al., 2018). Vogels, Demberg, and Kray (2018) compared the ICA score and overall pupil size as potential indices for CL that was induced by a *dual* or *single-task paradigm*. Interestingly, the ICA was only sensitive for CL in *single-task* conditions and decreased during the more difficult *dual-task* setting, whereas the pupil size increased with increased difficulty in both *single* and *dual-task conditions*

(Vogels, Demberg & Kray, 2018). The authors interpret that both parameters are legitimate indices for CL but reflect different neuronal processes in *dual-task* settings.

The ICA claims to discriminate between the light reflex and dilation during cognitive processing. Unfortunately, there is no detailed description of the procedure publicly accessible, and the ICA can only be implemented by purchasing the software that does not reveal any detailed information neither. Hence, the ICA remains without independent verification (Duchowski et al., 2018). That is why Duchowski and colleagues (2018) introduced a similar but fully accessible alternative, called the *Index of Pupillary Activity* (IPA). Both the ICA and IPA locate peaks in the wavelet signal that are then "de-noised" via hard thresholding followed by calculating the frequency (per second) of abrupt discontinuities detected in the signal (Duchowski et al., 2018; Marshall, 2002). See the published paper by Duchowski and colleagues (2018) for more detailed information regarding the IPA approach. The paper also includes a replicated experiment that confirms the IPA's sensitivity to task difficulty. Since this promising approach is relatively novel, the IPA has not been replicated through independent studies yet.

Besides the light reflex, Duchowksi and colleagues (2018) emphasize the issue with the fixed camera angle of the eye tracker: When the eye is rotating, the pupil looks like an ellipse which can result in a reduced record of the pupil dilation up to 12% (Mathur, Gehrmann & Atchison, 2013). Hence, this off-axis distortion has to be taken into account when measuring and interpreting changes in pupil diameter.

Pupil dilation, as a *Task-Evoked Pupillary Response*, offers many advantages compared to other psychophysical parameters. Due to the advancing technology, it is an easy and non-invasive approach, which is also objective because it is an unconscious physiological response. Even though the relationship between pupil dilation and cognitive processing has been replicated in several studies, one has to consider the confound factors that accompany eye movement measurement and take it into account when interpreting the outcomes since biases can significantly affect them.

Behavioral Measures. Since behavioral measures are not included in our study and less used in CL research than subjective or physiological measures, this topic will not be discussed in detail.

According to Chen and colleagues (2016), these response-related behavioral methods are defined as "… those that can be extracted from any user activity that is predominantly related to deliberate/voluntary task completion" (p. 19). He lists examples such as linguistic and grammatical features, eye gaze (because it is under voluntary control), mouse and keyboard usage and gait patterns (Chen et al., 2016). Frosina and colleagues (2018) compared different non-verbal behavior patterns between two conditions where the participants either had a true or false alibi. During the cognitive interview for suspects (to induce CL), they could observe that subjects in the false alibi condition significantly used fewer hand gestures (Frosina et al., 2018). Khawaja (2010) tested several speech and linguistic measures to examine their sensitivity to CL manipulations. In high CL conditions, he could observe that people tend to use more and longer pauses, longer response times, spoke longer, and in longer sentences, and also the use of plural personal pronouns increased whereas the use of singular pronouns decreased (Khawaja, 2010). He summarizes that under high CL, language becomes more complex and difficult to comprehend. Ikehara and Crosby (2005) included several measures in their study to compare their sensitivity to CL. Among other results, the authors see usage patterns of the computer mouse as a potential indicator to measure CL but recommend a combination of different behavioral parameters.

Like physiological indicators, behavioral patterns are objective and capture CL in *real-time*. In contrast to well-established physiological parameters, behavioral measures have a voluntary character. Even though they have won less recognition than other types of measures, they have the advantage that they can be collected implicitly and usually without additional equipment. This is why they can be applied easily in research, and more importantly, they could be more easily implemented in interactive systems to adapt to the user's CL dynamically. Nevertheless, more research is necessary to strengthen the mentioned links to CL.

Individual Differences. In general, CL can vary among participants due to interpersonal differences. One important aspect is the available working memory capacity since CL is conceptually linked to it. More in detail, people differ in their ability to control attention, which represents the *central executive* in widely-accepted working memory models (Baddeley and Hitch, 1974). See Chapter *Working Memory Models*. for more details. Unsworth (2009) examined the link between working memory capacity and free recall measures and concluded that people with high working memory capacity focused their attention more effectively. Delaney and Sahakyan (2007) instructed their participants to intentionally forget words from the first list and remember words from a second list. Interestingly, participants with high working memory capacity also remembered (via free call) fewer words from the first list than participants with low working memory capacity. The authors suggest that people with high working memory capacities are more context-dependent, hence are more able to effectively control mental processes (Delaney & Sahakyan, 2007).

When talking about working memory capacity, the individual motivation is often mentioned, too. Moreno (2010) suggests that motivation predicts the amount of invested cognitive effort in the task. In their review, Dai and Sternberg (2004) summarize that motivation has an important impact on attentional and cognitive processes in both laboratory and educational environments despite equivalent cognitive skills. Regarding classroom learning, Pintrich (2003) points out that motivational factors mediate learning by increasing or decreasing cognitive engagement. Gareau and Gaudreau (2017) differentiated between implicit and explicit autonomous (intrinsic) motivation and could show that only explicit autonomous motivation significantly predicted academic achievements when controlling for performance in the past (which many studies miss out). More important, there have been several studies confirming that motivation (often manipulated by monetary incentives) enhances working memory capacity (e.g., Gilbert & Fiez, 2004; Sanada, Kimura & Hasegawa, 2013). This effect could also be seen on the neurophysiological level. Szatkowska, Bogorodzki, Wolak, Marchewka, and Szeszkowski (2008) used fMRI Technology and the *n-back* task to measure working memory capacity. Motivation was manipulated by promising a monetary reward in the experimental condition. Results suggest that the right lateral OFC (*orbitofrontal cortex*) and left DLPFC (*dorsolateral prefrontal cortex*) play an important role in understanding the motivational influence on working memory (Szatkowska et al., 2008).

Another confounding factor when collecting TEPRs is a potential age effect. Lobato-Rincon, del Carmen, Bonnin-Arias, Chamorro-Gutierrez, Murciano-Cespedosa, and Roda (2014) investigated age differences in pupillary responses due to distinct light wavelengths. Older participants (46 – 78 years) showed a more delayed response to white light (Lobato-Rincon et al., 2014). Van Gerven, Paas, van Merrienboer and Schmidt (2004) could also observe an age effect when measuring CL with pupil dilation: Older participants' (*M*=68.6 years) pupillary response was not sensitive to cognitive processing.

Controlling for individual differences still is a fundamental issue in human-centered research. Fortunately, statistical approaches try to control these factors. Also in this area, further research is necessary to understand the underlying mechanisms behind these common confounding variables.

Summary. CL and related constructs have been the focus of several studies of different research areas. Even though there are slightly distinct definitions of CL established by researchers, the foundations usually are in accordance with each other: CL is the amount of invested cognitive resources to achieve a certain objective and is mainly modulated by the individual's working memory capacity. But as stated precisely by Hart (2006), "The many definitions that exist in the psychological literature are a testament to the complexity of the construct as are the growing number of causes, consequences, and symptoms that have been identified." (p. 904).

There have been many attempts in Cognition research to understand the underlying mechanisms behind CL over the last decades. Especially empirical work related to memory processes helped to gain a deeper understanding of how CL occurs. Even though researchers contribute important findings supporting crucial theories such as the *Multistore Model of Memor*y by Aktinson and Shiffrin (1968), there are still unsolved questions coming up, also demonstrating the complexity of neuronal networks. This challenge also reflects the state-of-art regarding the CLT. CLT forms the basis for CL understanding and research, but confirming its validity empirically leads to methodical hindrances. Despite advanced technology, neuroscience still has its methodical issues to confirm these Cognition and Learning theories. For instance, CLT is in accordance with acknowledged memory findings, but the differentiation between three separate load elements remains unconfirmed (Kalyuga, 2011). Therefore, it is important when dealing with certain (psychophysiological) models to look into their theoretical foundations and considering it when interpreting results.

There are distinct approaches to assess CL in a laboratory setting. Including subjective and/or objective measures seem to be the most popular procedure among researchers. As discussed, each approach has its methodical advantages and disadvantages. This is why many studies use combinations of different measures to assess CL more holistically. Especially physiological metrics such as the pupillary response or EEG parameters are on the rise since they offer a *real-time* and objective assessment. But these promising approaches still have to overcome methodical obstacles. For instance, several empirical findings suggest a relation between pupil dilation and CL, but light incidence represents a severe confounding factor. Hence, improving these encouraging objective methods should be a priority in research. Furthermore, the chosen measure(s) should be in line with the task; for instance, a visual search task assumingly evokes more saccades than fixations. It also depends on the aim of the study whether subjective or objective indicators are of more interest. All in all, researchers should put time into the experimental design and investigate the pros and cons of potential CL manipulations and indicators before conducting the study. This way, it is more likely to gain a real insight into CL mechanisms and contribute important research findings.

Stress as a Psychophysical Construct

Traditional and Modern Stress Theories. One of the first theories about stress was introduced by Walter Cannon (1914). He defines the biological stress reaction as the adaption of an organism to external threats. The release of energy results from a "fight-or-flight-reaction" (Cannon, 1914). This way, Cannon (1914) gives stress the purpose of the organism's survival: The biological preparation to flee or fight. Furthermore, he introduced the term homeostasis, which describes the process of maintaining internal stability while adjusting to external conditions (Brown & Fee, 2002).

According to today's understanding, Walter Cannon already conducted stress research, but it was Hans Selye (1950) who officially introduced the term "stress". He defined stress as an unspecific neuroendocrine physical reaction to stressors (Selye, 1970). Hans Selye differentiated between individual stressors as stress elicitors and an equal, unspecific biological stress reaction (Szabo, Tache & Somogyi, 2012). For instance, a spider (stressor) can cause a stress reaction for some people but not for others. Furthermore, Selye and Fortier (1984) classified the stress reaction into different phases: In the first stage of an acute stress reaction ("alarm reaction") caused by individual stressors, the body releases biochemical substances to adapt to the external environment, which stops as soon as stressors disappear.

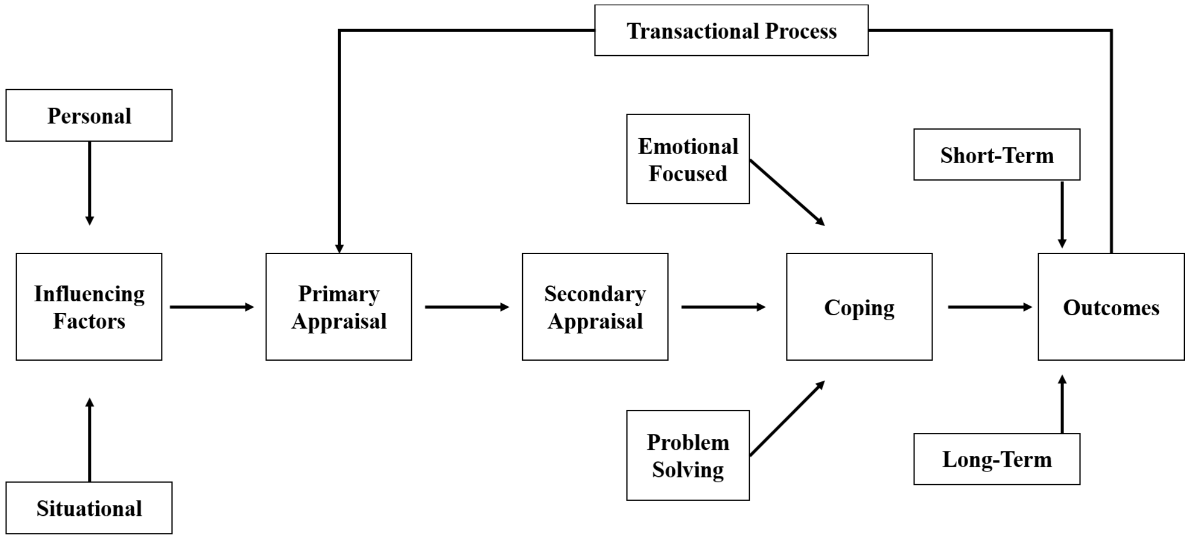


Figure 5. Transactional Stress Model (Lazarus & Folkman, 1984). Adapted from Schuster, Hammit and Moore (2003).

In case of a lasting stressor, the body develops a temporal stress tolerance ("resistance phase") which cannot be maintained permanently and results in an "exhaustion phase" which can cause serious impairments (Selye & Fortier, 1950).

The widely-accepted stress theory of Lazarus and Folkman (1984), the *Transactional Stress Model*, focuses on mental evaluation as the basis of stress triggers. It integrates the cognitive aspect of situational and resource appraisal and provides a crucial contribution to modern stress research (Lazarus & Folkman, 1984). According to the authors, stress (or the opposite: relaxation) is the result of two cognitive evaluation processes. First, the threat and its impact on the person's well-being is mentally evaluated (*Primary Appraisal*). If a stimulus is not identified as a threat, the situation will not cause stress for the person. Of course, there are personal (e.g., how the individual perceives the environment) and situational factors (e.g., novelty or duration) that also influence if a stimulus is perceived as a stressor (Lazarus & Folkman, 1984; Schuster, Hammit & Moore; 2003). If a stimulus is identified as a threat, the person will estimate available coping resources (*Secondary Appraisal*) to manage the threat successfully (Lazarus & Folkman, 1984). Hence, the person is not stressed in case of a threatening situation if he or she is convinced to have enough skills (or other resources) to handle the stressor. According to the *Transactional Stress Model*, the physical stress reaction is only triggered when there is a perceived threat (*Primary Appraisal*), and the person does not believe to be able to cope with it (*Secondary Appraisal*). Moreover, Lazarus and Folkman (1984) illustrate several levels of the stress reaction. According to them, stress causes a physical, cognitive, emotional, and behavioral reaction that all impact the re-evaluation of the coping reaction and hence, prospective dealing with stressors (Lazarus & Folkman, 1984). Research has identified two main behavioral reactions (or coping strategies): *Emotion-focused coping* tries to change the relation to the situation in form of avoidance, distancing, and optimism, whereas *problem-solving coping* tries to change actively the situation itself

(Schuster, Hammit & Moore, 2003). Either way, the applied coping strategy can have short-term (e.g., positive or negative feelings) and long-term (e.g., social functioning) consequences (Lazarus & Folkman, 1984; Schuster, Hammit & Moore, 2003).

Types of Stress. When illustrating the (physical) impact of stress, it is important to differentiate between types of stress that mainly differ in the duration of the stress reaction. In the following, the categorization from Miller, Smith, and Rothstein (1994) is adopted.

Acute Stress is the most common form of stress. It can be perceived as "exciting" for a short time, but it is exhausting for the organism in the long run. *Acute Stress* can also be triggered by athletic activities. Miller, Smith, and Rothstein (1994) list several psychological and physical symptoms of *Acute Stress*: If the stress reaction lasts over a certain time, it can manifest itself in negative emotions like anger or fear. On the physical level, lasting *Acute Stress* can cause headaches, muscle hardening, and indigestion among others (more in detail in Chapter *Physical Stress Reaction and its Consequences*.). Due to its short-term load, *Acute Stress* does not result in permanent physical and psychological impairment. Most people are exposed to *Acute Stress* regularly and are able to cope adequately (Miller, Smith & Rothstein, 1994).

The authors see *Episodic Acute Stress* as a regular exposition to stress. This constant stress perception often arises from highly developed personal demands and perceived social pressure. This type of stress can increase jumpiness and mental instability (Miller, Smith & Rothstein; 1994) According to the authors, these symptoms can be misinterpreted as hostile behavior, which can result in a negative impact on the social and professional environment. Another subtype of *Episodic Acute Stress* expresses itself through constant worry and an increase of pessimistic attitudes. Affected people mainly show crankiness, anxiety and depressive mood and less anger or hostility (Miller, Smith & Rothstein, 1994). On the physical level, *Episodic Acute Stress* is associated with chronical headache and migraine, high blood pressure, chest pain and coronal diseases (Miller, Smith & Rothstein, 1994). Furthermore, the authors point out that treatment generally requires professional support because *Episodic Acute Stress* can have a severe impact on the person's environment and perception. A lack of discernment and external blaming can make an intervention more difficult (Miller, Smith & Rothstein, 1994).

Chronic Stress reflects the opposite of *Acute Stress*. *Chronic Stress* is defined by long-term exposure to stressors that evoke a permanent stress reaction, which has severe consequences for "body, soul and life" (Miller, Smith & Rothstein, 1994). *Chronic Stress* can be triggered by threats on existential needs, dysfunctional personal relations or high pressure and demands within the work environment. Especially hopelessness on a foreseeable improvement of the situation is seen as one of the main causes of *Chronic Stress* stated by the authors. Often, this is mirrored in a negative worldview (Miller, Smith & Rothstein, 1994). Further, the authors note that traumatic events can increase the probability of developing *Chronic Stress*. The consequences of *Chronic Stress* are far-reaching and severe. *Chronic Stress* can result in violence towards the person itself (suicide) or others. It also causes a higher probability of cardiovascular diseases (Li, Zhang, Loerbroks, Angerer & Siegrist, 2015; Miller, Smith & Rothstein, 1994; Vitaliano, Scanlan, Zhang, Savage & Hirsch, 2002). In Chapter *Physical Stress Reaction and its Consequences*., the physical consequences of (chronic) stress are illustrated more in detail.

Physical Stress Reaction and its Consequences. Until today, stress researchers have identified several biochemical processes that induce or come along with the organism's "alarm reaction". In his review, Chrousos (2009) summarizes the elementary functions of the *central nervous system* (CNS) and *peripheral nervous system* (PNS). The CNS suppresses fatigue by increased arousal. At the same time, alertness, attention, vigilance, and aggressiveness also increase. Further, vegetative functions are inhibited: reproduction (libido) mechanisms, growth-stimulating processes, digestion and counter-regulatory feedback loops (Chrousos, 2009). The peripheral stress reactions contain an increased oxygen saturation in the blood, and increased blood circulation, reduced salivation, a dilation of bronchial tubes, an increased tension of skeletal muscles, enhanced reflexes, increased blood pressure (hypertension), quicker heartbeat (tachycardia), provision of energy (e.g. glycogenolysis), increased metabolism, short-term enhancement of the immune system and also counter-regulatory feedback loops that inhibit inflammation reactions temporary (Chrousos, 2009; Kaluza, 2012).

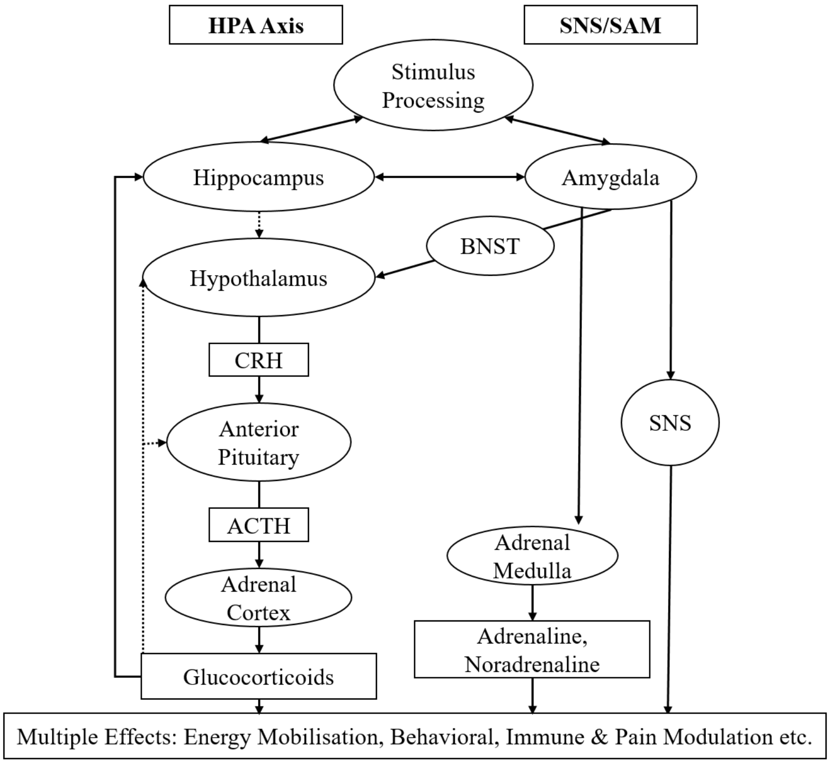


Figure 6. Major pathways of the two axes of the stress response illustrated by Murison (2016).

Additionally, the neuroendocrine stress reaction has been intensively investigated. Research suggests that there are mainly two stress axes activated. Under *Acute Stress*, the *Sympathetic Adrenal Medullary* (SAM) Axis is activated immediately and enables the following reaction cascade: When a stimulus is being identified as a stressor, catecholamines (particularly noradrenaline) are released from the sympathetic nerve into several tissues and blood. Because of this output, the adrenal medulla is being stimulated and further releases the stress hormones noradrenaline and adrenaline. As soon as the stress situation wears off, the sympathetic nervous system is down-regulated, and released catecholamines are degraded within minutes (Esler et al., 1979).

When a stress situation lasts for longer, the second axis *Hypothalamic-Pituitary-Adrenal* (HPA) is being activated. The Hypothalamus produces CRH (*Corticotropin-Releasing Hormone*) that induces the secretion of ACTH (*Adrenocorticotropic Hormone*) in the pituitary. CRH and ACTH enable the production of cortisol in the adrenal cortex (Kudielka, Hellhammer, Kirschbaum, Harmon-Jones & Winkielman, 2007). The release of cortisol has a counter-regulatory function on the axis' activation, since cortisol inhibits the production of CRF and ACTH, hence stops its own release over time (Golenhofen, 1997). The hormone cortisol influences about 20% of the human gene expression and therefore, has an elemental impact on several homeostatic processes (Chrousos, 2009).

All in all, the first axis gets activated very fast, and due to the adrenaline release prepares the body to cope with the stressor. The second axis gets activated after 20 to 30 minutes and prepares the body for a longer-lasting stress reaction through hormonal (cortisol) secretions (Chrousos, 2009; Kudielka et al., 2007).

While physiological processes and their measurement of the stress reaction have been extensively examined, psychological aspects have gained less attention in research (Gaab, Rohleder, Nater & Ehlert, 2005). As illustrated in *Traditional and Modern Stress Theories*., the *Transactional Stress Model* of Lazarus and Folkman (1984) points out that the cognitive appraisal precedes the physiological stress reaction. Several studies suggest that a stress evaluation significantly influences the cortisol production as a reaction on (psycho-social) stress (Gaab et al., 2005; Juster, Perna, Marin, Sindi & Lupien, 2012). Gaab and colleagues (2005) could demonstrate with a regression analysis that 35% of the variance of the cortisol production could be explained by the anticipatory stress appraisal. Further, empirical findings suggest that the anticipatory stress appraisal significantly affects changes in coagulation factors and inflammatory activity of monocytes (Wirtz et al., 2006; Wirtz et al., 2007). Once again, research demonstrates the impact of stress (appraisal) on the immune system.

Studying the effects of *Stress Management Trainings* (SMT) also lead to crucial conclusions. Studies show that the anticipatory stress appraisal functions as a mediator of training effects: A reduction of perceived stress caused a reduction of cortisol secretion (Hammerfald et al., 2006; Gaab et al., 2003; Storch, Gaab, Kuettel, Stuessi & Fend, 2007). Studies examining the effect of the anticipatory stress appraisal underlines the crucial role of cognitive evaluations as a precedent of the physiological stress reaction. These findings are in line with the *Transactional Stress Model* of Lazarus and Folkman (1984) and therefore support the model.

The complex neuroendocrine stress reaction shows that the human body can cope adequately with external stressors. Chrousos (2009) sees the (acute) stress reaction as crucial to develop a sense of positive states such as well-being, accomplishments, and functional social interactions. Contrarily, *Chronic Stress* is empirically linked to severe long-term consequences, such as impaired growing processes and can be (partly) responsible for behavioral, endocrine, metabolic, cardiovascular, autoimmune and allergic disorders (Chrousos, 2009; McEwen, 1998). Research conducted by Jergovic and colleagues (2014) provides findings that show that *Chronic Stress* causes telomere (chromosome's ends) shortening that is associated with a worse cell division and, therefore, biological aging (Levy, Allsopp, Futcher, Greider & Harley, 1992). Further, empirical findings suggest that people affected by posttraumatic stress disorder show an increase in DNA strand breaks, which could be reduced by psychotherapy (Morath et al., 2014). Furthermore, Nater and colleagues (2009) observed a genetic expression change induced by *Acute Stress* in healthy males. Another study showed that acute psychosocial stress increased the production of pro-inflammatory cytokines that are associated with a higher inflammatory reaction (Kuebler et al., 2015). Besides, Kuebler, Wirtz, Sakai, Stemmer & Ehlert (2013) show that *Acute Stress* provokes an impaired wound healing. These studies demonstrate the severe impact of stress on the human body, especially the immune system.

Stress Modulation & Measurement. Over the last decades, the *Trier Social Stress Test* (TSST) has been the prevailing method to induce (psycho-social) stress in experimental settings (Kirschbaum, Pirke & Hellhammer, 1993; Kudielka et al., 2007). The typical TSST procedure presented by Kudielka and colleagues (2007) contains three elements: A brief time (3 min) for the participant to prepare a speech (5 min) pretending a job interview in front of a "selection committee" followed by a mental arithmetic task (5 minutes) without preparation time. This procedure was designed to measure outcomes of the HPA axis response mainly; if including other measurements, modifications could be required (Kudielka et al., 2007). Several studies show that the TSST triggers a significant increase in saliva production and heart rate (Bohringer, Schwabe, Richter & Schachinger, 2008; Kirschbaum et al., 1993). Interestingly, Kirschbaum and colleagues (1993) could observe a gender effect: The TSST induced almost the double production of cortisol in male participants in comparison to female subjects. Further, a study conducted by Kudielka and colleagues (2004) suggests that the cortisol production induced by the TSST occurs independently of the day time. This is important because the normal cortisol production underlies a circadian rhythm: cortisol reaches its maximum in the morning and decreases during the day (Edwards, Clow, Evans & Hucklebridge, 2001). As explained in Chapter *Physical Stress Reaction and its Consequences*., cortisol is the outcome of the delayed HPA axis response. Correspondingly, the measured maximum of cortisol is measured 10 – 20 minutes after the TSST's end (Kudielka et al., 2007).

As illustrated in Chapter *Physical Stress Reaction and its Consequences*. (Nor) Adrenaline and Cortisol are secreted during the endocrine stress reaction and, therefore, well-established stress parameters in research. In particular, the cortisol measurement by saliva is widely recommended due to its non-invasive character (King, 2002). Dickerson and Kemeny (2004) conducted a meta-analysis about stressors and their cortisol responses and found out that the cortisol and ACTH secretion in saliva and blood plasma were the highest and needed the longest "recovery phase" when tasks had an uncontrollable and social-evaluative character. Besides a significant increase of the cortisol production and heart rate as a stress response, Rohleder and colleagues (2004) could also observe an increase in the activity of the saliva enzyme *alpha-amylase* (sAA). According to the authors, sAA has proven itself as a non-invasive and reliable sympathetic activity parameter. This is why sAA is used as a stress biomarker to an increasing degree (Rohleder, Nater, Wolf, Ehlert & Kirschbaum, 2004).

Galvanic Skin Response or (GSR) or also called *Skin Conductance Level (SCL),* is also a common method to determine a participant's stress level. Perala and Sterling (2007) tested the GSR to measure stress in soldiers. Their study results demonstrate that GSR is an acceptable method to assess stress objectively (Perala & Sterling, 2007). The authors see GSR's non-invasive, objective and rapidness as advantages over other common-used methods. A study conducted by Kadziolka, Pierdomenico, and Miller (2016) examined whether mindfulness as a personality trait has a stress-protective effect. Their findings indicate that “natively mindful” people were less likely to show a GSR increase in response to a stressful task (Kadziolka, Pierdomenico & Miller, 2016).

Besides the objective measures of the stress reaction’s outcomes, there have been several attempts to measure the subjective evaluation of perceived stress. As illustrated in Chapter *Physical Stress Reaction and its Consequences*., stress appraisal seems to play a crucial role in triggering the stress reaction. The *Primary Appraisal – Secondary Appraisal* (PASA) questionnaire introduced by Gaab (2009) is a reliable and valid method to measure the stress appraisal based on the *Transactional Stress Theory* from Lazarus and Folkman (1984). Accordingly, the PASA includes two subscales: Firstly, the degree of perceived threat (*Primary Appraisal*) and, secondly, the degree of perceived coping skills (*Secondary Appraisal*). The overall stress index is calculated by subtracting both subscales. Each subscale contains eight items with a six-point scale (from “totally wrong” to “totally right”). Using the PASA requires a certain situation to which the perceived stress is evaluated (Gaab, 2009). The PASA meets the psychometric quality criteria (Gaab, 2009). Since we included the PASA questionnaire in our study, it is explained more in detail in Chapter *Dependent Variables.*

Stress and Cognitive Load. Some researchers have recognized that CL has an impact on the user's stress and arousal level, which is why some studies use typical stress and arousal indicators to (indirectly) measure CL. Ikehara and Cosby (2005) see stress (among others) as a potential cognitive indicator for CL and emphasize that GSR (*Galvanic Skin Response*) can significantly detect stress or arousal induced by task difficulty since GSR is widely accepted as a measurement of the sympathetic nervous system. Hence, GSR is used to either detect the stress or arousal level (e.g., Joshi, Kiran & Sah, 2016) or cognitive processes (e.g., Shi, Choi, Ruiz, Chen & Taib, 2007; Nourbakhsh et al., 2012). Shi and colleagues (2007) point out that GSR has its origin in psychology to measure stress, but it also gets more and more popular among HCI researchers. The authors measure the user's stress and arousal level via GSR and observed that the GSR signal significantly increased when the task-induced CL increased. Therefore, Shi and colleagues (2007) see GSR as an objective indicator of the user's CL in *real-time*. Contrarily to findings, Conway, Dick, Li, Wang, and Chen (2013) came to another conclusion. They used GSR as the index of CL and compared a "stressed" group with a control condition. Results suggest that mean GSR values only differentiate between CL levels in the control group, inducing stress blurred the relation between CL and GSR signals (Conway et al., 2013). This finding supports the view that GSR mainly determines the stress level and therefore, measures CL only indirectly.

Other studies also suggest that arousal (e.g., stress) has an impact on cognitive processing. Brünken (2003) argues that affective states (measured by typical stress indicators such as GSR, body temperature and HR) have an impact on CL. Interestingly, he sees the self-reported stress level as a subjective but direct (causal) link to CL and physiological measures such as HR or pupil dilation as indirect causal links to CL (Brünken, 2003). But Brünken (2003) also emphasizes the other way around that, especially a high CL, may lead to a change in the stress and emotional state of the individual.

An example of the impact of stress on cognitive processes is the better-examined relation between stress and performance tasks, even though there exist inconsistent findings. Already in 1995, McEwen and Sapolsky suggest that the Glucocorticoid secretion as part of the physical acute stress response can enhance memory performance through oxygen and glucose delivery to the brain, whereas excessive stress exposure disrupts it. Kennedy and Scholey (2000) could also observe that glucose consumption leads to better performance in arithmetic tasks. Contrarily to these findings, Fraser and colleagues (2014) suggest that negative emotions (including stress) cause higher CL and lower learning outcomes, compared to the group where positive emotions were induced. Other researchers argue that stress (e.g., induced by pressure to perform) causes additional ECL, which leads to reduced learning within the CLT (Plass & Kalyuga, 2019; Quatieri et al., 2017).

Another interesting study conducted by Sato, Takenaka, and Kawahara (2012) tries to explain these mixed findings. They compared the selective attention on a visual search task (low and high CL) of a control group with a stress-induced group. Results suggest that stress had a positive effect on selective attention only in the low CL condition, not in the high CL. For the control group, the results were the other way around (Sato, Takenaka & Kawahara; 2012). These findings suggest that stress and perceptual load share the same attentional cognitive resources and further, lower stress may enhance cognitive processing until a certain stress level is reached.

Whereas many researchers disregard potential overlaps between CL and other psychophysical constructs, Fuentes-Garcia, Pereira, Castro, Santos, and Villafaina (2019) recommend their chosen measurements (HRV and EEG) as useful tools to either determine stress or cognitive load during cognitive tasks. In their study, a small sample of adolescents played chess with different complex scenarios and examined EEG activity (theta power spectrum) and HRV. Results show an increase in the sympathetic response with rising task difficulty (Fuentes-Garcia et al., 2019). The study's findings confirm that CL and stress are likely to evoke the same physical responses.

Summary. The psychophysical construct stress has been a broadly examined topic in research over the last decades. Especially the physical impact of stress is well investigated. The neuroendocrine response shows that the human body can deal well with acute stress, which is essential for survival. Since the stress response is profoundly examined, research offers several reliable, objective stress parameters such as (nor)adrenaline and cortisol. Using subjective methods is also a very popular and easy method to apply, as there is a common-sense among researchers that stress is triggered by individual mental evaluations.

Even though some empirical work includes stress parameters in their CL study design, very few research picks out the relation between both constructs as the central theme. This could be a sign of the methodical challenge to differentiate between both. As illustrated in Chapter *Stress and Cognitive Load.*, several studies demonstrate the impact of CL on the individual's stress and arousal level and also the effect of induced stress on CL and performance values. Even though there are mixed findings, it seems to be very likely that both constructs share common psychophysical outcomes. But different understandings and manipulations of both constructs make it difficult to comprehend the underlying mechanisms and distinction between both. Accordingly, Conway and colleagues (2013) emphasize that the presence of stress reflects a major challenge for CL detection, which can lead to biased findings.

Emotion as a Multidimensional Construct

Definition & Theories. Understanding emotions and their impact on human cognition and behavior has been the objective of several psychologists over the last decades. Rothermund and Eder (2011) summarize fundamental research findings regarding emotion as a multidimensional construct. If not stated otherwise, this book is used as a reference for the following section.

The word "emotion" is derived from the Latin word "emovere" which is translated as (to) drive, to set something in motion. The authors define emotions as "object-directed, involuntary triggered affective reactions that evoke temporal changes in inner experience and behavior of a person." (p.166). This definition combines important aspects of the psychophysical construct. Emotions have a subjective affective component, for instance, anger or happiness that can be assessed consciously through attention (Lambie & Marcel, 2002). Secondly, emotions are always related to something that evokes the emotional reaction, for instance feeling afraid of a spider at the wall. Because of this object-orientation, emotions are time-limited: They are linked to the appearance and disappearance of the object. This emotional reaction is an automatic response that cannot be suppressed. Emotion regulation strategies try to manipulate situations and cognitive evaluations so that emotions do or do not come up, but they cannot inhibit their release if they are triggered anyways. With this definition, it is possible to differentiate between object-related emotions and diffuse feelings and affective dispositions (stable personality traits).

Emotions represent a multidimensional construct that affects several psychophysical levels. Emotions evoke a change in the person's feelings and conscious experience. Barrett and Russell (1999) examined the basic dimensions of emotions and introduced the *Circumplex-Model*. In general, they differentiate between two bipolar but independent dimensions: the degree of pleasantness (valence) and degree of activation (arousal).

According to the model, the emotional experience evokes a combination of both dimensions. For instance, excitement reflects a high degree of activation and medium value for pleasantness, whereas calmness indicates low arousal and a medium degree of pleasantness.

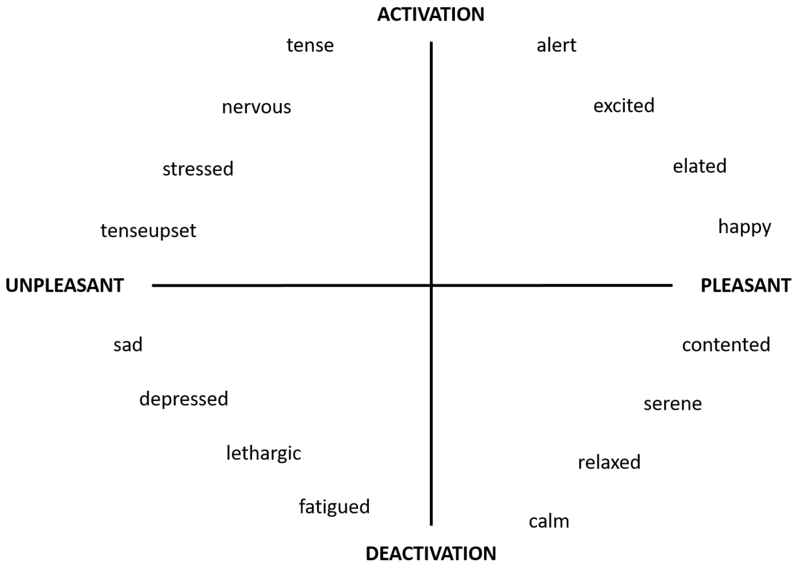


Figure 7. A schematic for the two-dimensional structure of affect. Adapted from Barrett & Russel (1999).

Emotions as a multidimensional construct also contain a cognitive element. Emotions always come along with an evaluation. Here, evaluation refers to an evaluative categorization from events (object-orientation) and their implications for the individual (Brosch, Pourtois & Sander, 2010). This means, in case of an emotional event, the person evaluates cognitively if it is positive or negative for the individual and releases a suitable emotion. It has to be mentioned that the same event can lead to different evaluations among people due to inter-individual differences and contextual aspects (Weiner, 1985).

Physical and Behavioral Reaction. There is also a physical reaction when experiencing emotions. Research suggests that emotions evoke changes in the activation of the autonomous nervous system (ANS). For instance, feeling afraid leads to physical arousal, such as an increase in pulse and GSR (*Galvanic Skin Response*). Hence, it is assumed that this reaction is needed to adapt adequately to important life events. Even though researchers agree that the ANS plays an important role regarding emotions, finding more detailed conclusions of certain emotion reaction profiles seems to be difficult. Cacioppo, Berntson, Larsen, Poehlmann, and Ito (2000) argue that their meta-analysis suggests that even basic emotions cannot be fully differentiated only by the visceral activity alone. Solely, the valence dimension (positive or negative emotion) can be derived from the ANS. This could be one of the reasons why the central nervous system (CNS) has been more and more the center of emotion research: Neuronal networks, and areas that are activated during an emotional state are from interest. For instance, the amygdala seems to be the center of the fear system and is connected to several other neuronal structures (LeDoux, 2000). Based on this assumption, there are therapeutic attempts to down-regulate the amygdala's activity by applying emotion regulation strategies and hence, reducing negative feelings such as fear, anger or sadness. Herwig and colleagues (2019) used fMRI sessions with neurofeedback (manipulating mental activity based on immediate feedback of the activity in the neuronal region) and the well-established strategy *Reality Check* to reduce the amygdala's activation when seeing emotional pictures. Four weekly neurofeedback sessions resulted in a significant decrease in amygdala activity, also compared to a control group (Herwig et al., 2019).

Even though there have been partly successful attempts and advanced technology to detect the neuronal and physical level of certain emotions, there have not been clear findings. Findings suggest that there are different systems and networks involved, indicating a complexity that is (still) difficult to capture (Dalgleish, Dunn & Mobs, 2009; Rothermund & Eder, 2011).

Emotions also manifest themselves in expressional behavior. This includes facial expressions, gestures, and voice characteristics. Especially facial cues have been the focus of research. Intercultural studies are indicating that basic emotions such as fear, anger, and happiness can be identified correctly across cultures by facial cues (Izard, 1994). Identifying emotions from facial cues is also of interest to detect deception. Some studies suggest that emotions lead to very brief involuntary, universal facial expressions (Frank &Ekman, 1997; Warren, Schertler & Bull, 2009). Researchers claim that with training, these so-called *Micro Expressions* can be detected (Warren, Schertler & Bull, 2009). Even though studies are supporting the idea of brief, uncontrolled emotional expression in the face, several results are mixed and inconclusive. For instance, a study conducted by Warren, Schertler & Bull (2009) compared emotional and unemotional lying and could see that their trained encoders performed better than chance only in the emotional lying condition. But the overall performance was not better than chance (Warren, Schertler & Bull, 2009). These mixed findings and the severe implications of deception detection techniques make it necessary to conduct further research before drawing explicit conclusions.

Intuitively, one would expect that emotions lead to certain expressions, not the other way around. Interestingly, the so-called *Facial-Feedback-Hypothesis* describes the phenomenon that facial expressions also have an impact on the emotional state. For instance, a (controlled) smile enhances the mood, whereas frowning worsens it (Soussignan, 2004). A more recent study by Davis, Senghas, Brandt, and Ochsner (2010) examined if botox injections (that paralyze muscles of facial expressions) have an impact on emotional experiencing. They could show that the “botox group” had a significant decrease in the strength of emotional experiencing compared to a control group (Davis et al., 2010).

Function of Emotions. For a long time, psychologists saw emotions more like a burden, which impairs rational thinking and acting. This has changed completely in emotion psychology; today, the majority sees emotions as fundamental to adapt adequately to the environment. Emotions are seen to come along with a willingness to act. Hence, they are linked to motivational aspects and necessary to cope with challenges. There have been attempts to specify the functionality of emotions. In 1980, Plutchik introduced *A Psychoevolutionary Theory of Emotion* that links certain emotions to a specific function. For instance, fear arises in the case of a threatening event and increases the willingness to flee or fight for protection. If seeing a potential mate, happiness would occur and leads to the willingness to court and approach the targeted person to fulfill the biological function of reproduction (Plutchik, 1980). Through these examples, the limitations of the theory get visible quickly. For instance, happiness is an emotion that can occur in several distinct contexts, not only in the case of sexual interest. Plutchik (1980) admits that certain emotion systems are (mis)used for other areas, but this argument shows that he still traces back a certain function to a specific emotion. Empirically, there seems to be a behavioral tendency that positive or negative emotional experiencing lead to a willingness to approach or avoid (Eder & Rothermund, 2008). Bradley, Codispoti, Cuthbert, and Lang (2001) confronted participants with different positive and negative emotional pictures and examined the eyelid closure reflex. The results show that in case of negative pictures participants significantly closed their eyelid more often (stronger defense reaction) than when seeing positive pictures (Bradley et al., 2001).

Emotions and Cognition. More recent studies investigated the impact of emotions on eye-related parameters. Knickerbocker and colleagues (2019) investigated the influence of emotion-laden words on eye movements (several fixation and gaze parameters) among three points of measurement. The results indicate that positive and negative words have a processing advantage (for instance, shorter fixation and gaze durations) over neutral words (Knickerbocker et al., 2019). More in detail, positive words had a benefit in all three measurement times (early, late, and post), whereas negative words showed effects only in late and post measurements (Knickerbocker at al., 2019). Scott, O'Donnell, and Sereno (2012) conducted a similar study capturing several fixation parameters while participants read sentences containing positive, negative and neutral words. Further, the authors differentiated between low and high frequency (rarely or often used words in daily speech). Results suggest an effect of word frequency and emotion, both positive and negative emotional words showed a lexical processing advantage (reading faster) in comparison to neutral words (except high-frequency negative words). This effect was modulated by word frequency (Scott, O'Donnell & Sereno, 2012). Hence, positive familiar words were the fastest read stimuli.

Further, the impact of emotions on memory and cognitive processing has been objective to current studies. Panasati, Ponsi, Monachesi, Lorenzini, Panasiti, and Aglioti (2018) compared performance and facial temperature between Psoriasis (skin disease) patients and a healthy control group during an emotional *n-back* task. The modified version of the *n-back* task contained positive, negative or neutral images between the test stimuli. Interestingly, when showing negative images, accuracy was significantly lower than for neutral and positive words in both groups (Panasati et al., 2018). Even though controls showed better overall performance, Psoriasis patients (associated with impaired emotion regulation skills) had a smaller difference in performance between the low and high CL condition (Panasati et al., 2018). Therefore, the authors suggest that emotions have an impact only on low CL and not on high CL modulations because, in the latter, there are no available resources to process emotional distractors. A similar study conducted by Li and Ouyang (2012) investigated the effect of emotions on verbal and spatial working memory (both *n-back* tasks) performance using EEG Technology. In the *0-back* (low CL) condition, they did not find an effect of the emotional state. But for the *2-back* (high CL) condition, results suggest that only a negative emotional state reduced performance accuracy (Li & Ouyang, 2012). This had an effect on certain *event-related potentials* (ERPs), which are related to working memory processes, but only for spatial working memory (Li & Ouyang, 2012). The authors imply that the interactive pattern of emotion and working memory is modulated by CL and justify the effect only found in the high CL condition by an attention resource competition between processing emotions and cognitive cost. DeFraine (2016) tries to give another explanation to the varying findings on how CL effects emotions by comparing emotion maintenance and a single emotion modulation. His findings suggest that CL reduced the intensity of negative emotions during a singular emotion modulation but not during emotion maintenance (DeFraine, 2016). Hence, he sees emotion maintenance as a key factor that influences the impact of CL on emotions, but it has to be mentioned that emotion maintenance was only operationalized by additional instructions to maintain the intensity of the feelings evoked by viewing emotional stimuli. Nevertheless, DeFraine (2016) emphasizes that mixed results can be due to different study designs and instructions.

As illustrated in Chapter *Physiological Measures*., empirical evidence suggests that CL evokes a pupillary response, among others. Since emotion and cognition seem to influence each other it is also from interest if the emotional state has an impact on widely used CL measures. There have been empirical findings suggesting that emotional arousal also evokes a change in pupil size. A study conducted by Bradley and colleagues (2008) investigated changes in pupil size when subjects looked at neutral, pleasant or unpleasant pictures under controlled luminance conditions. The authors could observe a main effect for emotional pictures, with a significant increase in pupil diameter for pleasant and unpleasant stimuli in comparison to neutral stimuli. Interestingly, there was no significant difference between positive and negative pictures (Bradley, Miccoli, Escrig & Lang, 2008). Bayer, Ruthmann, and Schacht (2017) went a step further and investigated the effect of personal-relevant emotional stimuli on cortex activity and pupil size. Results show a significant increase in emotion-related ERPs, pupil dilation, and higher Arousal ratings when presenting sentences that referred to the participants' spouse in comparison to unknown agents (Bayer, Ruthmann & Schacht, 2017). But it has to be noted that this study examined a small sample consisting of solely female participants.

Besides visual emotional stimuli, research findings suggest that this also applies to other senses. Already in 2003, Partala and Surakka investigated pupil size changes during auditory pupil stimulation. Findings indicate a significant increase in the pupil size when participants listened to positive or negative sounds in comparison to neutral sounds (Partala & Surakka, 2003). This empirical evidence of the emotional effect on different senses substantiates the assumption that emotions have a robust impact on cognitive processes.

Due to these robust findings, Plass and Kalyuga (2019) claim that the role of emotions has not been sufficiently considered within the CLT model in the past. They illustrate four possible integrations within the CLT. Corresponding to the findings of Li and Ouyang (2012), Plass and Kalyuga (2019) suggest that the emotional state can be seen as ECL since it requires additional mental effort to process current emotions, either in form of task-extra or task-irrelevant processing. Another perspective categorizes emotions as ICL when emotional outcomes are necessary to accomplish learning goals. Plass and Kalyuga (2019) see trainings to deliver bad news to a patient as an example. The third view on how emotions affect CL illustrated by Plass and Kalyuga (2019), emphasizes the effect of emotions on motivation. More in detail, the effect of emotions on learning is mediated by motivation. Whereas several research findings indicate that positive emotions enhance (intrinsic) motivation, which leads to better performance, the effect for negative emotions is not quite as explicit (Isen & Reeve, 2005; Plass & Kalyuga, 2019). Interestingly, some studies show that negative emotions can also lead to increased motivation (Plass & Kalyuga, 2019). According to the authors, this effect could be caused by emotion regulations to shift the attention away from a negative emotional state to focus on learning effectively. Another explanation could be that negative emotions cause increased motivation to avoid failure (Plass & Kalyuga, 2019). Hence, further research is needed to draw conclusions about the effect of negative emotions on motivation. Regarding the final view presented by Plass and Kalyuga (2019), research findings seem to be more conclusive: From this perspective, emotions influence the attention that directly reduces or increases (working) memory capacities. More in detail, studies have shown that especially negative stimuli can impair cognitive processing, while positive stimuli can enhance it (Li & Ouyang, 2012; Panasati et al., 2018; Plass & Kalyuga, 2019; Scott, O'Donnell & Sereno; 2012).

Summary. The perspective change of emotions and their functionality has revealed how broad and complicated this multidimensional construct seems to be. Even though we all have our own definition and idea of emotions, research shows that due to methodical challenges it is still difficult to reach a deeper understanding, elucidated by neuropsychological studies that come to oppositional or inconclusive results (Moreno, 2010; Plass & Kalyuga, 2019; Rothermund & Eder, 2011). The interconnectedness with other constructs also implies methodical challenges but emphasizes the important impact on several psychophysical aspects such as motivation or cognitive evaluations (Moreno, 2010). The latter is important to keep in mind when modulating mental processing in an experimental setting because studies have shown that processing emotions and mental effort seem to share similar cognitive resources (Panasati et al., 2018; Plass & Kalyuga, 2019).

Despite recent theoretical advances in our understanding, getting to the bottom of underlying mechanisms of emotions remains challenging. Despite mixed findings, it is clear that emotions influence our inner experience and behavior, either positively or negatively. Here, the reciprocal effect between emotions and performance was misleadingly considered unilateral in the past: performance influences emotions, not the other way around (Rothermund & Eder, 2011). But research suggests that emotional states also affect several areas, including performance, behavior and experiencing (Bradley et al., 2001; Scott, O'Donnell & Sereno, 2012; Soussignan, 2004).

All in all, recent research findings show that emotions deserve empirical attention to gain a deeper understanding since it seems to have an impact on several psychophysical areas and thus, broad implications for applied research.

Virtual Reality in Research

Since technology is rapidly advancing over the last decades, *Virtual Reality* (VR) is more and more used in empirical study designs. VR offers new possibilities for modulating and measuring psychophysical parameters. According to Vickers, Schultheis, and Manning (2018), VR can address previous methodical limitations and provides safe and sensitive measurements.

Virtual Reality in Psychology. Vickers, Schultheis, and Manning (2018) used VR technology to assess differences in driving performance in a VR driving simulation between subjects after brain injury and a healthy control group. Results show that additional mental demands (via *dual-task paradigm*) impact the driving performance of both groups negatively, and as expected, this effect was greater for people after brain injury (Vickers, Schultheis & Manning, 2018). The authors conclude that VR can provide sensitive metrics of driving and simulation scenarios that are too dangerous to apply “on-road”.

VR is increasingly used for *Exposure Therapy* (ET) in clinical practice and research fields (Cardos, David, David, 2017; Diemer & Zwanzger, 2019). ET is a widely-accepted method to treat anxiety disorders by exposing the individual with the anxiety source without causing any danger. According to Diemer and Zwanzger (2019), several VR studies show that a VR exposure can evoke a subjective, physiological and behavioral fear response. Landowska, Roberts, Eachus, Barrett (2018) conducted a study examining the effects of *Virtual Reality Exposure Therapy* (VRET) on acrophobia (fear of heights) using brain activity measures. Across three VRET sessions, the brain activity changed towards normal values during VR exposure, indicating that VRET is a method, which is easy to apply and effective to treat acrophobia. Cardos, David and David (2017) provide a meta-analysis regarding the effectiveness of VRET for flight anxiety. Findings show the advantages of VRET in comparison to control groups and similar effects when comparing to classical exposure interventions (Cardos, David & David, 2017). Interestingly, smaller and younger samples showed larger effect sizes of VRET (Cardos, David & David, 2017). Furthermore, the authors revealed outcome types, number of sessions and follow-up intervals as significant moderator variables. In line with these findings, Suso-Ribera and colleagues (2019) compared the efficiency of three forms of ET for small animal phobia: traditional ET (iVET), VRET, and *Augmented Reality* (ARET). Results revealed similar effects of all ET forms (Suso-Ribera et al., 2019). The authors point out that using VRET or ARET can have a higher likelihood of being accepted by patients since they frequently refuse traditional *en-vivo* exposure (Suso-Ribera et al., 2019).

VR has also been applied in learning and training research. Stark-Wroblewski and colleagues (2008) used VR Technology to expose Psychology students to psychological treatment approaches to reduce fear of flying. The authors compared prior and post knowledge about the presented (VR) content and conclude that VR can be useful to enhance students' understanding of academic topics (Stark-Wroblewski et al., 2008). Morrongiello, Corbett, Beer, and Koutsoulianos (2018) used a VR environment to test its efficiency to improve children's street-crossing behaviors. The VR program focused either on "where to cross" or "how to cross". Results show that children in both VR intervention groups made considerably fewer errors for all pedestrian safety variables compared to the prior test and the control group (Morrongiello, Corbett, Beer & Koutsoulianos, 2018).

Virtual Reality in Computer Science. Also, in Computer Science VR's popularity is rising. Teranishi and Yamagishi (2018) designed a VR environment that simulates assembling a computer allowing the user to move objects virtually with immediate feedback. A questionnaire revealed a significant improvement when comparing prior and post knowledge about the correct positions for separate computer parts, but not for the parts' names (Teranishi & Yamagishi, 2018).

Madathil and Greenstein (2017) conducted a study comparing different forms of *Usability* testing of a simulated online shopping website: They included a VR environment, a traditional lab environment, and a Cisco WebEx® Conference (screen sharing approach). *Usability* metrics, such as error rates and time to complete the tasks, did not differ among groups (Madathil & Greenstein, 2017). Additionally, the subjective CL of participants and test facilitators was rated lowest in the lab condition, but there was no significant difference between the VR and WebEx® group (Madathil & Greenstein, 2017). Interestingly, participants stated greater involvement and a more immersive experience in the VR condition compared to the WebEx group. Madathil and Greenstein (2017) see VR Technology as a promising new way of conducting *Usability* tests because their findings indicate only minor disadvantages over the traditional lab setting and emphasize the benefit of remote testing: This low-cost approach allows user and facilitator to participate from different locations (Madathil & Greenstein, 2017).

Another study conducted by Narasimha, Dixon, Bertrand, and Madathil (2019) examined the suitability of immersive VR systems for remote collaborative work. Participants completed a card sorting task either traditionally in-person via screen-sharing or VR environment (Narashimha et al., 2019). Performance parameters and workload (via NASA-TLX) showed no significant difference among the groups (Narashimha et al., 2019). The overall *Usability* rating (*IBM-Computer System Usability Questionnaire*) was significantly better for the VR card sorting compared to the other groups (Narasimha et al., 2019). All in all, these results indicate that VR can be used effectively to create remote collaborative work environments since it evoked the same or even better outcomes than the classic approaches that were included.

Dan and Reiner (2017) investigated whether visual learning via 2D or 3D has an impact on CL by determining EEG parameters (alpha and theta oscillations). Participants had to watch paper-folding (*origami)* instructions via a 2D Video and stereoscopic 3D Display. The CL index based on EEG parameters revealed a significant higher CL for the 2D condition (Dan & Reiner, 2018). Interestingly, subjects with lower spatial abilities profited more from the 3D presentation than the 2D instructions. This study shows that 3D representations can provide benefits compared with traditional 2D approaches since 3D reflects the human's natural environment more realistically.

Summary. These studies demonstrate that VR technology has found its way into research among several areas of applications. Researchers emphasize the benefits, especially in fields where traditional approaches imply safety or cost issues. This is why VR Technology reflects a promising alternative in areas such as anxiety treatment. Several studies underpin this by demonstrating that VR approaches obtain comparable results with traditional methods (e.g., Cardos, David & David, 2017; Madathil & Greenstein, 2017). Besides these encouraging findings, VR technology still needs to overcome some technical issues. In their review, Weech, Kenny, and Barnett-Cowan (2019) list the creation of a "sense of presence" in users as a main problem. This perceived "presence" seems to relate negatively to another popular barrier: *Cybersickness* (Weech, Kenny & Barnett-Cowan, 2019): The authors define *Cybersickness* as the bodily discomfort and malaise when being exposed to VR content. According to Weech, Kenny, and Barnett-Cowan (2019), these negative side effects underlie individual differences. Hence, VR Technology offers new opportunities in research and on the field, but in order to exploit its promising potential, some technical obstacles have to be overcome.

Aim of the Study & Hypotheses

This work contributes a new methodical approach in assessing CL. We used VR Technology to control better for light incidence when measuring the pupillary response (pupil diameter). CL is operationalized by three task difficulties using the *n-back* task that induces a heavy workload on working memory and therefore reflects a good method to assess CL (Guastello et al., 2015). Similar to the study design of Panasati and colleagues (2019), we chose a *2-back* condition to measure high CL. The *0-back* task serves as the low CL and *1-back* as the medium condition. Like Duchowski and colleagues (2018), we conducted the *Digit Span Memory Test* to assess participants' working memory capacity and hence, to use it as a covariate in the statistical analysis. To verify the CL modulation, we used the NASA-TLX questionnaire to assess the subjective workload. Further, we integrated the PASA survey to observe the effect of CL on the perceived stress level to contribute empirical work on how these two constructs relate to each other. Furthermore, we adapted a SAM questionnaire to collect data about the impact of CL on perceived valence, arousal and dominance, since research suggests considerable effects (e.g., Bradley et al., 2001; Scott, O'Donnell & Sereno, 2012; Soussignan, 2004). This is why we formulate the following research questions (RQs) and hypotheses (H):

RQ1: "Does the *n-back* task induce Cognitive Load?"

H1: The Pupil Diameter increases with increasing Task Difficulty

H2: The Number of Blinks increases with increasing Task Difficulty

H3: The Index of Pupillary Activity (IPA) increases with increasing Task Difficulty

H4: The subjective Cognitive Load increases with increasing Task difficulty

RQ2: "Do Performance Measures reflect increasing Cognitive Load?"

H5: The Error Rate increases with increasing Task Difficulty

H6: Reaction Time increases with increasing Task difficulty

RQ3: "Does Cognitive Load have an Impact on the perceived Stress Level?"

H7: Perceived Stress increases with increasing Task Difficulty

RQ4: "Does Cognitive Load have an Impact on Emotional States?"

H8: Perceived Valence decreases with increasing Task Difficulty

H9: Perceived Arousal increases with increasing Task Difficulty

H10: Perceived Dominance decreases with increasing Task Difficulty

Methods & Materials

Sample

Our study sample consisted of 31 subjects that were recruited via brochures and posters. Brochures (illustrated in Chapter *Appendix*) were distributed in the canteen and posters in designated places in several buildings of the University of Konstanz. The brochure and poster contained basic information about the study and a link to the meeting coordination platform *Calendly*® (https://calendly.com/de). Before potential participants could choose a time slot of an hour, they had to confirm that they fulfill the required criteria: Fluent in German (to avoid linguistic misunderstandings), owning an academic e-mail address (due to incidents with external participants in the past) and most importantly, no eye-related impairments to avoid physical biases when measuring the pupillary response. Hence, people wearing lenses were excluded from the experiment. Even though the conditions of participation were highlighted on brochures, flyers and the coordination platform, the fulfillment was checked again at the beginning of the experiment. Further, confirmed participants received more detailed information about the study procedure and were asked not to wear eye make-up during the experiment, since it can worsen the eye detection. In case this was forgotten, a make-up remover was available on-site. Participants received a sequential identification number to ensure data anonymization. Subjects received 10€ compensation for participating.

Materials

Main operations were conducted with two desktop monitors (HP© LP2475w, 24'), a *Logitech Ultra Flat Keyboard*©, and a *Logitech Click! Optical Mouse*© connected to a PC with *Windows 10 Enterprise*© installed. A *MacBook Air* (macOS Sierra©, version 10.12.3, 13') was only used to carry out the *Digit Span Memory Test*. Pupil Capture© and *Pupil Player*© eye-tracking software (version 1.13.29) were used to register data from the Pupil

Labs© lenses. These eye-tracking lenses were combined with a VR headset by *HTC VIVE© Pro Full Kit*. Participants used two matching controllers (*HTC VIVE© Pro Full Kit*) to navigate through instructions and to respond to test stimuli. The software SteamVR© (version 1.7.8) was used as an interface between HTC VIVE© devices and the computer.

Unity © (version 2.0.0) was used to build a VR application to execute the *n-back* task.

Study Design & Procedure

Our work is a follow-up study of von Bauer (2018), also dealing with assessing CL via physiological measures, concluding that the pupillary response detects the individual's cognitive state. He successfully used the *n-back* task to induce CL and assessed the pupil dilation among others (von Bauer, 2018). This is why some elements are adopted from this previous work.

We conducted a *single-blind within-subjects design* since participants completed all task difficulties. The group assessment was transparent for the study conductor, but not for participants. To avoid *order effects*, a counter-balanced design was chosen. Subjects were assigned sequentially to one of six groups. Hence, these six conditions cover all possible sequences of the three task difficulties of the *n-back* task. The study lasted about an hour per participant. A pilot study with three volunteers preceded the experiment. This pilot study was conducted to check whether instructions and the study procedure were comprehensive. Follow-up interviews with the volunteers provided helpful improvement suggestions. In the following, the final experimental procedure is illustrated.

Before the experiment started, all materials were prepared: copying *Pen&Paper* materials, starting *Pupil Capture®, SteamVR®, Unity®*, and the *Digit Span Memory Test*. In the case of *Cybersickness*, dextrose and water were provided on-site. First, the participants were welcomed and offered water to create a positive atmosphere. Right after, it was revised if the participant met the required criteria and the declaration of consent was handed out. After signing the consent, subjects were asked to turn off their cell phone, and make-up was removed if necessary. Then, they had to rate their current wakefulness on a 5-point Likert scale. After that, the *Digit Span Memory Test* was conducted, which was executed by the participants but monitored by the study conductor. Then, the VR phase began.

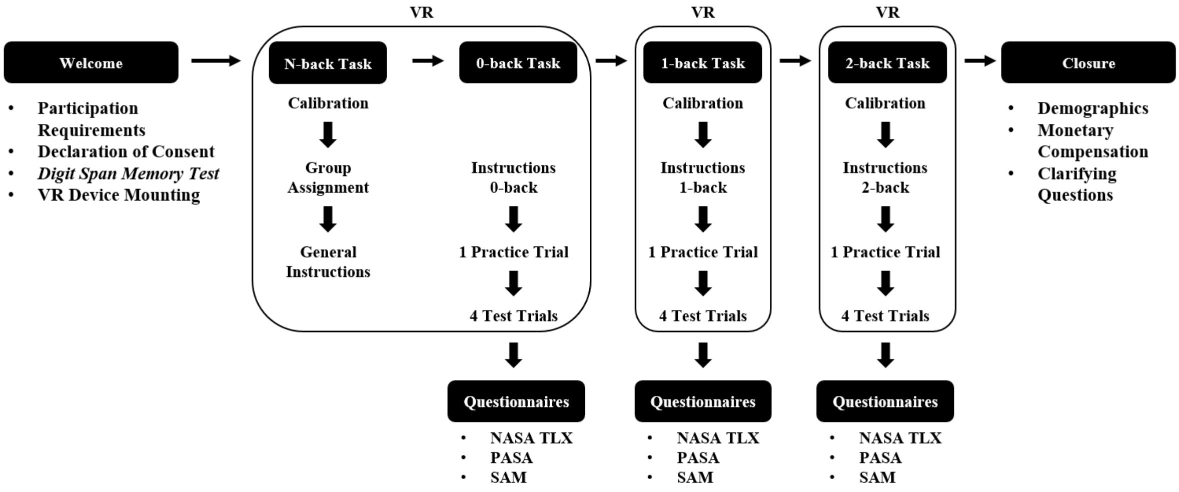


Figure 8. Study procedure illustrating one of six counter-balanced groups following the order: *0-back, 1-back* and *2-back*. Elements in the white box were conducted within the VR environment.

First, controller position (leaving hands with VR controllers on the table) and headset were adjusted. The right and left-hand click via VR controller were explained. The camera view of *Pupil Capture®* was used to check whether eyes were positioned well to detect the pupillary response. Visual eye markers were used to calibrate the eye position. After completing the calibration, *Unity®* was opened and recording (*Unity®* and *Pupil Capture®*) started. The sequential group assignment took place by the study conductor before the *n-back* instructions started. Within the VR environment, general instructions about the *n-back* task process (see Chapter *Independent Variable* for more details) were presented. The procedure was the same for all three task difficulties. The participant navigated autonomously through the experiment using the right-hand click. Depending on the counter-balanced condition, this was followed by specific instructions regarding the first *n-back* task difficulty. After that, participants had a practice trial with immediate feedback whether their response was right or wrong. After the practice trial, they had the chance to clarify uncertainties with the study conductor before the four test trials without performance feedback began. Right before practice and test trials started, participants were asked to always focus their gaze on the center and to give their best (see Chapter *Appendix* for more details).

After completing each *n-back* task condition, subjects were asked to put down the headset and fill out three questionnaires: NASA-TLX to assess the perceived CL, the PASA to assess the retrospective stress level, and the SAM to assess the current emotional state.

After completing the *0-back, 1-back* and *2-back* condition with three retrospective surveys each, programs and recording were stopped, and a demographics sheet (listed in Chapter *Appendix*) was handed out. Demographics were assessed at the end of the experiment to avoid unconscious *priming effects* on the *n-back* task performance. Finally, participants received 10€ compensation and, if necessary, questions about the experiment were clarified.

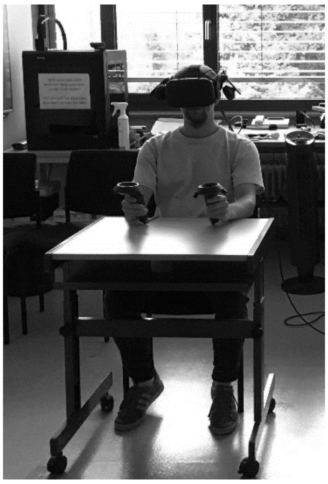


Figure 9. The participant's main working place to complete the VR *n-back* task and fill out questionnaires.

The follow-up work included organizing the completed materials, saving the recordings of the pupillary response, and preparing the laboratory for the next participant.

The room where the study was conducted provided a chair and a small table for the participant in the middle of the room (*Figure 9*). Table and chair positions were marked on the floor for standardizing purposes. The participant was seated there for the whole experiment: conducting the task within the VR environment and also to fill out *Pen&Paper* materials. The study conductor had a desktop chair and a table with two desktop monitors to supervise all involved software programs (*Figure 10*).

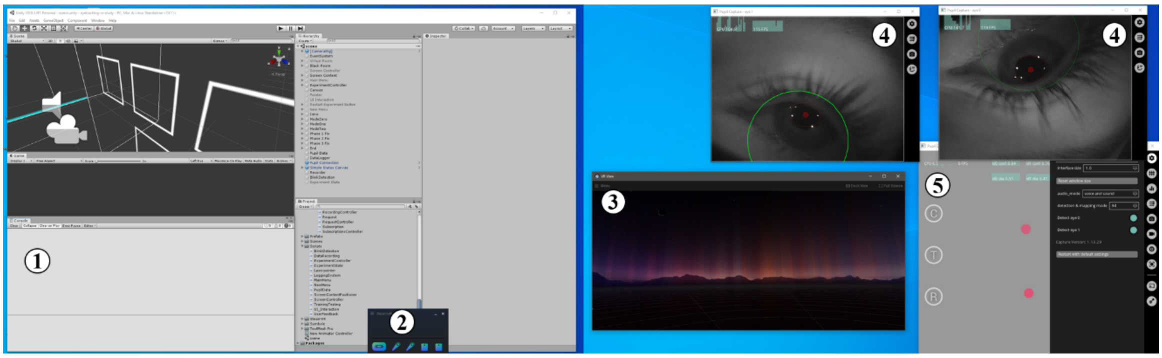


Figure 10. Screenshot of the supervised programs by the study conductor. 1= *Unity*® Console to operate the *n-back* task, 2= Supervising the status of the VR equipment, 3= Supervising the current VR view, 4= Controlling the eye detection, 5= Supervising the *real-time* values of eye diameter and its accuracy level.

Independent Variable

The three difficulties of the widely-used *n-back* task reflect the independent variable. The design was adopted from the preceding work from von Bauer (2018). As illustrated in Chapter *Performance Measures*., when conducting the *n-back* task, subjects have to compare the present letter with a letter *n* steps back. Von Bauer (2018) used a *3-back* condition, but his results suggest that it was too hard for participants and the author recommends the *2-back* task difficulty as the highest CL condition. This is why we included the following three task difficulties: *0-back, 1-back,* and *2-back*. For the *n-back* task, ten letters (C, D, F, H, K, N, P, R, V, Z) were chosen because they are highly readable and do not allow to form short words (von Bauer, 2018). Within the VR environment, a font size = 6 and "LiberationSans SDF" font type (without a font style) were chosen. Conducting the *0-back* task, participants had to decide whether the first letter matched each following letter. Conducting the *1-back* task, participants had to decide whether the present letter matched the last letter. Conducting the *2-*

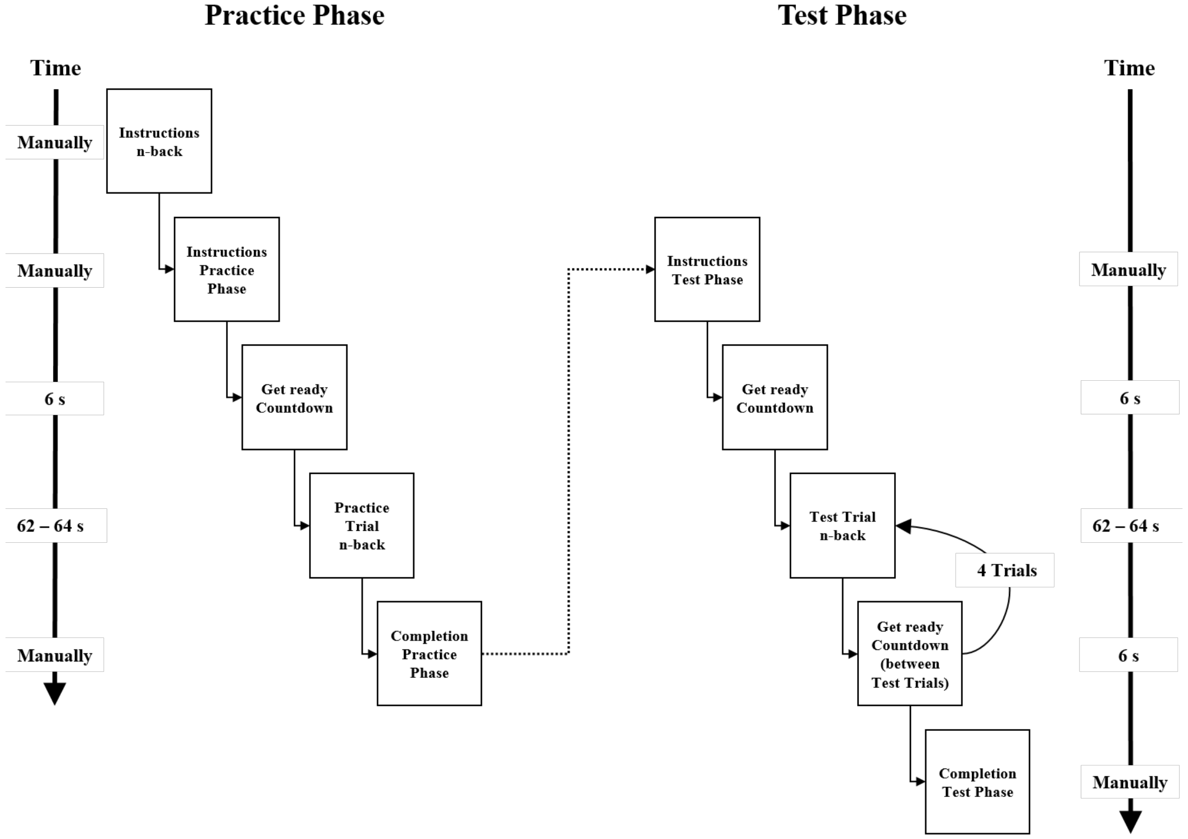


Figure 11. An *n-back* task run divided into practice (1 trial) and test phase (4 trials).

back task, participants had to decide whether the present letter matches the one before the last. One trial consisted of 30 test stimuli + n (reference) stimuli; hence, there were 30 responses for each trial registered. Participants had one practice trial before the four test trials began. One letter was presented for 1.5 s, which was followed by a break of 0.5 s. Within the 1.5 s of stimuli presentation, the participant had to decide whether the current letter matched the letter *n* steps back. The backside buttons of the VR controllers were used to either perform a right-hand click (match or forward instructions) or left-hand click (no match). If the person did not react within the time window, a false response was logged. One trial consisted of one-third matches and two-thirds no-matches with the default rule that there were no three sequential matches (von Bauer, 2018, Grimes, Tan, Hudson, Shenoy and Rao, 2008) There was a *Get ready Countdown* (6 s) at the beginning of the first trial and between them. Here, participants were additionally reminded to always center their gaze and to give their best. Additionally, in the *0-back* task difficulty, participants were reminded that the reference letter changed with each test trial.

A practice trial was included to avoid instructional misunderstandings. A rectangular feedback bar that was positioned beneath the letter gave immediate feedback after the participants' response: green if the response was correct, red if the response was wrong. Since immediate feedback can cause "distracting" emotions and stress (e.g., Raaijmakers, Baars, Schaap, Paas, & van Gog, 2017), we only included the feedback bar in the practice trial for learning purposes. During the test trials, the feedback bar was colored grey and only served as a confirmation that the participants' response was registered. The total test time was 272 s (4 trials + 4 *Get Ready countdowns*) for *0-back* and *1-back*. Because two instead of one reference letter are needed in the *2-back* condition, the total test time was 280 s (4 trials + 4 *Get Ready countdowns*). The *n-back* task was created with the program *Unity©*. Dark grey (#383838) was chosen for the VR environment background, White (#FFFFFF) for text and *n-*

back letters. For practice trials, the feedback bar was either green (#B8FA37) or red (#FA6037). For test trials, it was presented in grey (#E5E5E5). Importantly, the VR display was always "centralized", which means that instructions and *n-back* task remained fixed in the visual field, independently of the individuals' head movements.

Dependent Variables

To measure CL objectively, we used the *Pupil Lab©* Lenses integrated into the VR headset at a rate of 120 Hz for each eye. This means that the eye tracker provided 120 values of pupil diameter per second per eye. Similar to Knickerbocker and colleagues (2019), we only used one eye's data. The eye was chosen that provided more usable data (see Chapter *Data Acquisition and Pre-processing* for more details). The pupil diameter was assessed during the whole *n-back* task. For analysis purposes, we only considered data logged during test trials. These pupil diameter values were averaged within each task difficulty.

To detect the number of blinks, a blink detector provided by *Pupil Capture*® was activated. For analysis purposes we summed up the number of blinks for each task difficulty, resulting in three points of measurement.

To measure CL subjectively, the NASA-TLX questionnaire was used after completing each task difficulty. Similar to several other empirical works (Hart, 2006), we left out the weighting procedure and therefore used the RAW TLX. The NASA TLX shows a good re-test reliability, split-half reliability, Cronbach's Alpha, and internal consistency (Xiao, Wang, Wang & Lan, 2005). Further, the following subscales were used: *Performance*, *Mental Demand*, *Frustration*, *Effort,* and *Temporal Demand.* Similar to von Bauer (2018), we left out the subscale *Physical Demand* since the *n-back task* only requires mental demands.

For analysis purposes, we calculated the RAW TLX with five included dimensions resulting in an overall score from *MIN* = 0 to *MAX* = 100. The *Pen&Paper* version of the

NASA-TLX was handed out after completing each task difficulty. Hence, we have three times of measurement: *0-back*, *1-back,* and *2-back* condition.

Third Variables

To assess the perceived stress level of the participants, the PASA questionnaire introduced by Gaab (2005) was used. As illustrated in Chapter *Stress Modulation & Measurement*., this survey is based on the *Transactional Stress Theory* published by Lazarus and Folkman (1984) and includes two subscales: *Primary Appraisal* (threat and challenge) and *Secondary Appraisal* (self-concept and loss of control) with eight items each. Participants have to evaluate on a 6-point scale from “totally wrong” to “totally right”. Subtracting the *Primary Appraisal* from the *Secondary Appraisal* provides the overall Stressindex. The PASA always refers to a certain situation that is evaluated (Gaab, 2005). The PASA questionnaire was tested on a non-clinical male sample and showed a good internal consistency (for the subscales α: 0.61 – 0.83). A factor analysis confirms the expected factor structure (Gaab, 2005). In our study, each task difficulty reflects a situation that has to be evaluated. Hence, we have three times of measurement of the perceived stress level: after completing the *0-back*, *1-back* and *2-back* condition. The *pen&paper* version of the PASA was handed out after completing each task difficulty.

The widely-used the *Self-Assessment Manikin* (SAM) dimensions introduced by Bradley and Lang (1994) were used to assess the current state of emotions. SAM is a non-verbal pictorial assessment technique that directly measures valence, arousal, and dominance that are related to an individuals’ reaction to a wide variety of stimuli (Bradley & Lang, 1994). All dimensions consist of five figures each. The Valence dimension ranges from a smiling, happy figure to a frowning, unhappy character. The Arousal dimension ranges from an exciting figure to a relaxed, sleepy character. The Dominance dimension includes a range of different sizes of the figure to reflect changes of control. Hence, the smallest figure represents the minimum and the biggest figure, the maximum control of the situation. According to Bradley and Lang (1994), the SAM has been used to test the emotional reaction to a variety of stimuli. A factor analysis revealed that for the dimensions Valence, Arousal and Dominance accounted for 24%, 23% and 12% of the variance (Bradley & Lang, 1994). These results comply with the findings of the *Semantic Differential Scale* introduced by Mehrabian and Russel (1974) that contains 18 items to assess the three dimensions. Hence, the SAM provides a valid, easier, quicker and non-verbal method for assessing people's affective experience in comparison with widely-used surveys at that time (Bradley & Lang, 1994). Similar to Bradley and Lang (1994), we used a 9-point scale for each dimension. Hence, participants could choose a figure or a value between two figures. The *Pen&Paper* version of SAM was handed out after completing each task difficulty resulting in three points of measurement.

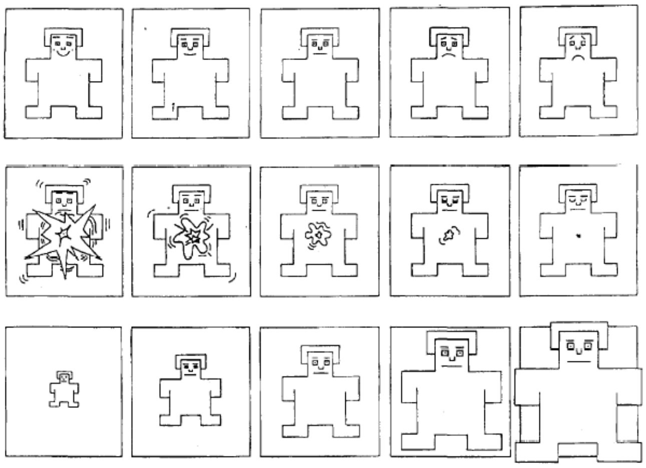


Figure 12. The Self-Assessment Manikin (SAM) used to rate the affective dimensions of Valence (top panel), Arousal (middle panel), and Dominance (bottom panel). Adapted from Bradley and Lang (1994).

Further, we included performance metrics in our study design. Error rates were calculated by adding up false responses during the *n-back* task conditions. For each task difficulty, participants had to respond to 30 stimuli per trial. Hence, per task difficulty, there were 120 responses registered. If the participant did not answer within the time window of 1500 ms, the answer was rated as false. Further, we calculated the reaction time during the *n-back* task by subtracting the timestamp of response from the timestamp where the stimuli appeared. If the participant did not answer within the time window, the maximum duration of stimuli presentation (1500 ms) was adopted. Reaction times were averaged per task difficulty resulting in three times of measurement.

The *Digit Span Memory Test* was conducted to measure the participant's working memory capacity. Digit Span Tests provide adequate psychometric properties (Waters & Caplan, 2003). In this study, we used an *open-source* online tool (https://timodenk.com/blog/digit-span-test-online-tool/) to conduct the *Digit Span Memory Test* (forward method). Within the online tool, the sound was deactivated, the sequence length started with four digits that appeared for 1000 ms each. Participants received written instructions before starting the test.

Adopted Statistical Analysis

All statistical analyses were conducted with the open-source *RStudio*® (version 1.2.1335) software, if not stated otherwise. Before conducting statistical analysis, variables were tested on normal distribution using the Shapiro-Wilk Test, since it provides good test power even for small samples (Shapiro & Wilk, 1965). Further, we chose the Mauchly's Test to test for sphericity for repeated measures. Testing the assumptions will be stated. To test the hypotheses, we used a one-way analysis of variance (ANOVA) with repeated measurements. Even though ANOVA is relatively robust to requirement violations (Duchowski et al., 2018; Schminder, Ziegler, Danay, Beyer & Bühner, 2010), we also provide a non-parametric analysis (Friedman Rank Sum Test) in case of requirement violations. In the case of significant findings, a post-hoc Pairwise Comparison (Tukey HSD correction) was conducted. For some analysis, additional variables were included as covariates. Effect sizes were calculated using the *partial eta squared* (η^2). Relations between variables were calculated using the *Pearson* Correlation (r^2). All analyses were conducted with a significance level of $p < 0.05$.

Results

Data Acquisition and Pre-processing

Statistical analyses are mainly based on two data sets per participant. *Pupil Capture©* created a "pupil_positions" (timestamps, pupil diameter, and its confidence among others) and "blinks" file automatically. Additionally, there was a script programmed with *Unity©* that combined the timestamp, pupil diameter, and its confidence from the "pupil_positions" responses for both eyes with difficulty level, letter, letter number per trial, and trial. This way, it was possible to extract the pupil diameter precisely per task difficulty. Recordings started when the general instructions were presented and ended (manually) when the *n-back* task with all task difficulties was finished.

The raw pupil diameter set consisted of *M*=414972 (*SD*=50243.46, *MIN*=209456, *MAX*=467272) values per participant. The great variance is mainly due to individual differences in reading speed (instructions), and questionnaire fill out time. These raw data sets were reduced as follows: First, general instructions (trial 0) and practice trials (trial 1, 6, 11) were excluded. Then, certain periods were cut out: values logged during specific instructions, after finishing the *n-back* task, short breaks between test trials, between presented letters during the *n-back* task, and between task difficulties (filling out questionnaires). After that, all pupil data with a confidence value (*Pupil Capture©*'s level of measurement accuracy) < 0.8 were removed. This way, blinks were excluded automatically. Finally, all pupil diameter values below two and above eight were removed, since they are out of the diameter's range (Kramer, 1990). After completing these steps, the cleared up data sets consisted of *M*=24679 (*SD*=16858.25, *MIN*=0, *MAX*=58383) observations per eye. These cleared up data sets formed the basis for excluding participants due to a poor amount of data. Included cleared up data sets had to consist of at least 1000 values for each task difficulty. Four participants did not meet this criterion and therefore, were excluded. Additionally, one person could not participate because the VR devices did not work, and one person was excluded because the eye cameras indicated that she had problems to keep her eyes open and showed other strong fatigue symptoms (very high number of blinks, yawning, and instructional misunderstandings). Hence, six participants were excluded. For the included data sets, the eye camera was chosen that provided more cleared up data (right-eye camera in 80% of the cases). Questionnaires were digitalized. For the SAM dimensions, Valence (1 = ”very unhappy”, 9 = “very happy”) and Arousal (1 = “very calm”, 9 = “very aroused”) scales were reversed.

Sample

31 subjects were invited, but the final sample consisted of 25 data sets. All participants were students of the University of Konstanz, with 36% studying at the Faculty of Science, 24% at the Faculty of Humanities, and 40% at the Faculty of Politics, Law, and Economics. The sample consisted of 48% male and 52% female participants. All counter-balanced groups consisted of four participants each, except for group F with five persons. Subjects were *M*=22.76 (*SD*=1.69, *MIN*=20, *MAX*=26) years old. Participants had to rate their wakefulness from 1 (“very tired”) to 5 (“very awake”). They felt relatively awake with *M*=3.68 (*SD*=0.748, *MIN*=2, *MAX*=5). Conducting the *Digit Span Memory Test*, participants reached a score of *M*=6.44 (*SD*=1.417, *MIN*=4, *MAX*=9) slightly below the “Magical Number 7” reported by Miller (1959). 96% of the participants were right-handed. One person (4%) reported a chronic disease (asthma). Four students (16%) reported current medication intake (all birth control pills). The sample had relatively little experience with VR devices beforehand with 48% no experience at all, 40% a one-time experience, and 12% rare use of VR technology. No one reported symptoms of *Cybersickness* during the study. Only two persons (8%) already knew the *n-back* task and three participants (12.5%) the *Digit Span Memory Test*. The participants rated their motivation to complete the *n-back* task best

possible with M=3.72 (*SD*=0.458, *MIN*=3, *MAX*=4) retrospectively. Hence, all participants stated to complete the *n-back* task either very or either motivated.

Hypotheses Testing

RQ1. Our first research question ("Does the *n-back* task induce Cognitive Load?".) deals with the CL modulation. Different parameters had been examined to confirm a successful CL manipulation by three task difficulties of the *n-back* task. Physiological metrics include pupil diameter and blink rate. Additionally, the IPA was calculated as a new method to measure CL using the pupillary response, and self-reported CL (via the NASA TLX questionnaire) was included.

The first hypothesis ("The Pupil Diameter increases with increasing Task Difficulty") deals with the impact of task difficulty on the pupil diameter. In *Table 1*, you can see the descriptive analysis of the pupil diameter per task difficulty. The pupil diameter distribution is also illustrated in *Figure 13*. The basis for pupil diameter analyses are mean values per task difficulty per participant. A Shapiro-Wilk Test confirmed a normal distribution for all task difficulties (*0-back*: W=.99, $p > .05$; *1-back*: W= .98, $p > .05$; *2-back*: W = .96, $p > .05$).

	n	***MIN***	***Q1***	**Median**	***Q3***	***MAX***	**Mean**	***SD***
0-back		2.81	4.48	5.02	5.57	7.21	5.00	0.97
1-back	25	2.65	4.11	4.80	5.53	6.63	4.84	1.04
2-back		2.97	4.92	5.58	6.18	7.39	5.51	1.12

Table 1. Descriptive Analysis of the pupil diameter (in mm) per task difficulty. *n* = Sample Size, *MIN* = Minimum Value, *Q1* = 1st Quartile, *Q3* = 3rd Quartile, *MAX* = Maximum Value, *SD* = Standard Deviation.

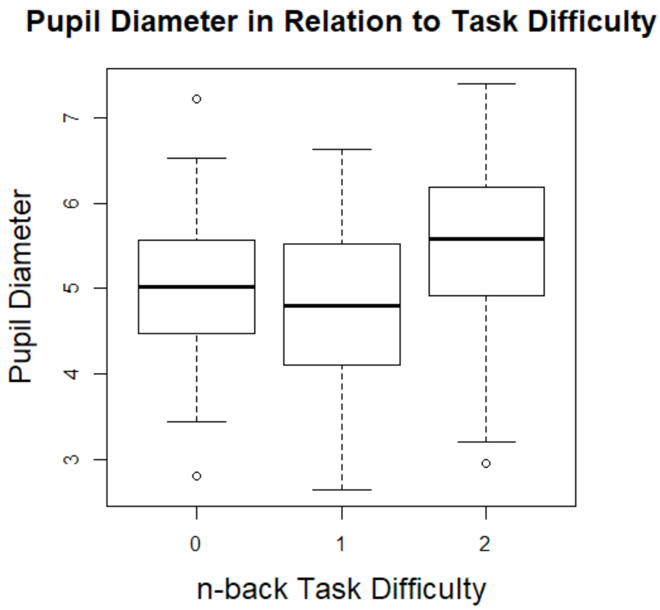


Figure 13. A Boxplot graph showing the distribution of the Pupil Diameter (in mm) per *n-back* task difficulty.

A one-way ANOVA with repeated measures was conducted to compare the effect of CL on the pupil diameter among *n-back* task difficulties. There was a significant effect of task difficulty ($F(2/48) = 7.766$, $p < .01$, $\eta^2 = .07$). A post-hoc pairwise comparison (HSD Tukey correction) revealed significant differences of the pupil diameter between the *0-back* and *2-back* ($p < .05$) and also between the *1-back* and *2-back* condition ($p < .01$).

There was no significant effect between the *0-back* and *1-back* task difficulty ($p > .05$). Conducting an Analysis of Covariance (ANCOVA) showed no significant relations ($p > .05$) of self-reported motivation, age, self-reported wakefulness, and Digit Span on the pupil diameter among task difficulties. Since Mauchly's Test indicated that the assumption of sphericity was violated ($p < .05$), a non-parametric Friedman Rank Sum Test was also conducted. Contrarily to the reported ANOVA, it revealed no significant difference among task difficulty conditions with χ^2 (2) = 6.08, $p > .05$). Further, Significant correlations ("Pearson") were found for the pupil diameter and reported Dominance for the *0-back* condition (r^2=-.49, $p < .05$) and in the *2-back* task difficulty with the Stressindex (r^2= .42, $p < .05$).

The second hypothesis ("The Number of Blinks increases with increasing Task Difficulty") examines the effect of CL (via task difficulty) on the accumulated number of blinks. *Table 2* includes descriptive analyses regarding the number of blinks per task difficulty. *Figure 14* illustrates the frequencies per task difficulty in the form of a boxplot.

A Shapiro-Wilk Test did not confirm a normal distribution for no task difficulty (*0-back*: W = .73, $p < .001$; *1-back*: W= .80, $p < .001$; *2-back*: W = .82, $p < .001$). A one-way ANOVA with repeated measures was conducted to compare the effect of CL (via task difficulty) on the number of blinks. It indicated no significant effect (F(2/48) = 0.7989, $p > .05$, η^2= .01). Since normal distribution and sphericity (Mauchly's test: $p < .05$) could not be confirmed, a Friedman Rank Sum Test among repeated measures was additionally conducted. Similar to the reported ANOVA, the chosen non-parametric test did not show a significant effect (χ^2 (2) = 4.56, $p > .05$).

	n	***MIN***	***Q1***	**Median**	***Q3***	***MAX***	**Mean**	***SD***
0-back	25	2	49	114	256	977	183.84	224.75
1-back		3	31	71	169	532	131.04	146.44
2-back		1	48	102	179	560	157.2	160.71

Table 2. Descriptive Analysis of the accumulated number of blinks per task difficulty. *n* = Sample Size, *MIN* = Minimum Value, *Q1* = 1st Quartile, *Q3* = 3rd Quartile, *MAX* = Maximum Value, *SD* = Standard Deviation.

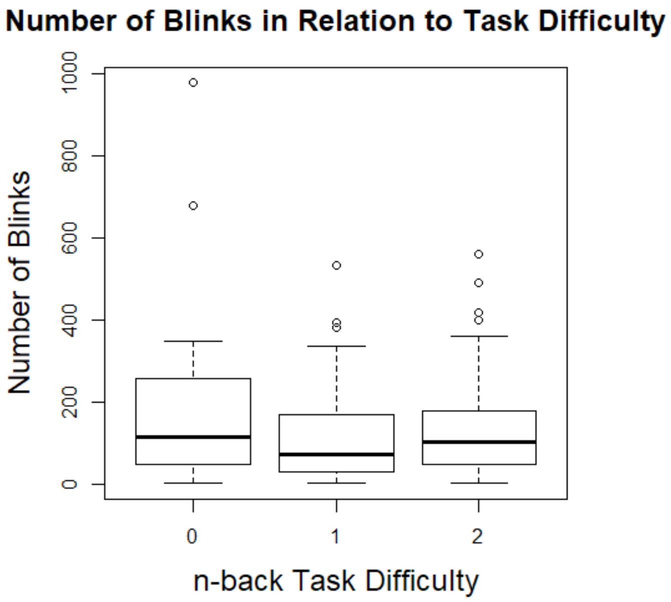


Figure 14. A Boxplot graph showing the distribution of blinks per *n-back* task difficulty.

The following hypothesis ("The Index of Pupillary Activity increases with increasing Task Difficulty") examines whether the IPA score (based on pupil diameter and registered blinks) introduced by Duchowski and colleagues (2018) discriminates significantly between task difficulties, hence CL. The python-script for the IPA calculation was partly provided by the authors (Duchowski et al., 2018). Since the IPA calculation is based on the automatically logged data files "pupil_positions" and "blinks" from *Pupil Capture*®, three participants were excluded additionally because these data files were not readable or incomplete. In total, we included *n* = 22. For further analysis purposes, an IPA value was calculated per task difficulty per participant.

	n	***MIN***	***Q1***	**Median**	***Q3***	***MAX***	**Mean**	***SD***
0-back	22	0.05	0.11	0.17	0.21	0.39	0.17	0.08
1-back		0.02	0.08	0.14	0.19	0.32	0.15	0.09
2-back		0.02	0.09	0.17	0.23	0.34	0.17	0.09

Table 3. Descriptive Analysis of IPA Score per task difficulty. *n* = Sample Size, *MIN* = Minimum Value, *Q1* = 1st Quartile, *Q3* = 3rd Quartile, *MAX* = Maximum Value, *SD* = Standard Deviation.

In *Table 3*, descriptive statistics are reported. *Figure 15* depicts the distribution of the IPA score (in Hz) among *n-back* task difficulties. A Shapiro Wilk test confirmed a normal distribution for all task difficulties (*0-back*: *W* = .94, *p* > .05; *1-back*: *W* = .96, *p* > .05; *2-back*: *W* = 0.96, *p* > .05). A one-way ANOVA for repeated measures did not render a significant difference of the IPA score among task difficulty (*F*(2/42) = 0.8457, *p* > .05, η^2 = .01). Since a Mauchly Test indicates a violation of the assumption of sphericity (*p* < .05), a non-parametric test was also conducted. In line with the reported ANOVA, a Friedman Rank Sum Test showed no significant effect (χ^2 (2) = 1.4545, *p* > .05). Hence, no further post-hoc analyses were conducted.

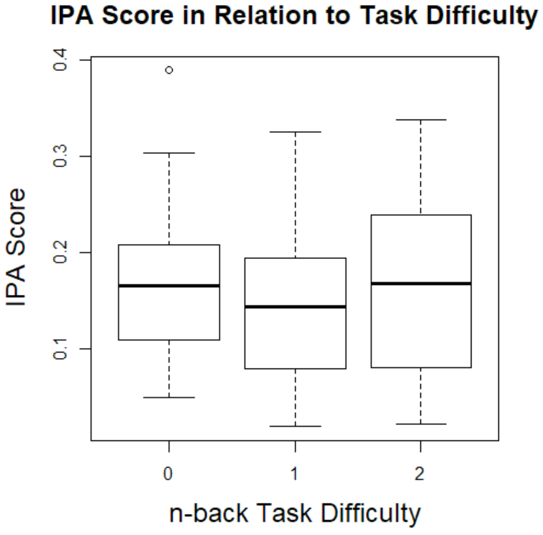


Figure 15. A Boxplot graph showing the distribution of the IPA score (in Hz) per *n-back* task difficulty.

The fourth hypothesis (“The subjective Cognitive Load increases with increasing Task Difficulty”) examines whether CL (induced by task difficulty) has an impact on the subjective CL assessed with the NASA TLX questionnaire. For analysis purposes, the overall (RAW) TLX score was calculated. Since one dimension was left out (*Physical Demand*), the accumulated overall TLX score has a value range from *MIN*=0 to *MAX*=100. In *Table 4,* related descriptive statistics are reported. *Figure 16* depicts the distribution of the overall TLX score among *n-back* task difficulties.

	n	***MIN***	***Q1***	**Median**	***Q3***	***MAX***	**Mean**	***SD***
0-back	25	5	12	19	31	72	25.12	17.06
1-back		11	27	32	52	69	35.72	15.83
2-back		38	50	62	75	89	63.92	14.91

Table 4. Descriptive Analysis of the overall TLX score per task difficulty. n = Sample Size, *MIN* = Minimum Value, *Q1* = 1st Quartile*, Q3* = 3rd Quartile, *MAX* = Maximum Value, *SD* = Standard Deviation.

The overall TLX score was tested for normal distribution, which was only partly confirmed by a Shapiro-Wilk Test (*0-back*: $W = .88$, $p < .01$; *1-back*: $W = .96$, $p > .05$; *2-back*: $W = 0.95$, $p > .05$). Sphericity was confirmed conducting a Mauchly Test ($p > .05$). A one-way ANOVA with repeated measures showed a strong significant effect ($F(2/48) = 59.347$, $p < .001$, $\eta^2 = .52$). Post-hoc Pairwise Comparisons (HSD Tukey correction) indicated a significant effect between all task difficulties: *0-back* and *1-back* ($p < .05$), *0-back* and *2-back* ($p < .001$) and also *1-back* and *2-back* ($p < .001$).

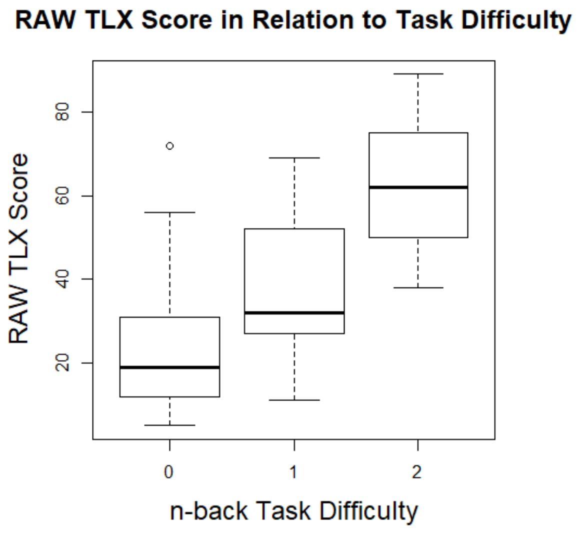


Figure 16. A Boxplot graph showing the distribution of the overall TLX score per *n-back* task difficulty.

Since normal distribution is only partly confirmed, a non-parametric test for repeated measures was also conducted: The Friedman Rank Sum Test (χ^2 (2) = 33.63, $p < .001$) also indicated a significant effect of induced CL on the subjective CL, but contrarily to the reported ANOVA, this was only the case between *0-back* and *2-back* ($p < .001$) and *1-back* and *2-back* ($p < .001$) conducting Pairwise Comparisons (HSD Tukey correction).

RQ2. Our second research question ("Do Performance Measures reflect increasing Cognitive Load?") examines whether the chosen performance metrics error rate and reaction time support the success of the CL manipulation. Here, we have a total sample of n = 23 per task difficulty, since data revealed retrospectively that two participants only responded in case of a match and not when there was no match. Hence, their error rates and reaction times are highly biased due to this instructional misunderstanding.

The first performance-related hypothesis ("The Error Rate increases with increasing Task Difficulty") examines whether CL has an impact on the error rate among task difficulty. Error rates were accumulated per participant per task difficulty. *Table 5* shows descriptive statistics of the error rate among task difficulty. Since we included four test trials à 30 test stimuli, the error rate could range from *MIN*=0 to *MAX*=120 per task difficulty. In case a participant did not respond within the time window, a "false" respond was logged. In case multiple responses for one stimulus were registered, only the first input was included. *Figure 17* depicts the distribution of the error rate among *n-back* task difficulties.

	n	***MIN***	***Q1***	**Median**	***Q3***	***MAX***	**Mean**	***SD***
0-back	23	0	0	0	1.5	4	0.91	1.24
1-back		0	1	3	4	16	3.43	3.93
2-back		3	10	16	24.5	44	17.65	10.80

Table 5. Descriptive Analysis of the error rate per task difficulty. *n* = Sample Size, *MIN* = Minimum Value, *Q1* = 1st Quartile, *Q3* = 3rd Quartile, *MAX* = Maximum Value, *SD* = Standard Deviation.

A Shapiro Wilk Test only confirmed a normal distribution for the *2-back* task (*0-back*: W = .74, p < .001; *1-back*: W = .78, p < .001; *2-back*: W = .93, p > .05). A one-way ANOVA for repeated measures indicated a significant effect of CL on the error rate (F(2/44) = 40.617, p < .001, η^2= .56). Post-hoc Pairwise Comparisons (HSD Tukey correction) revealed a significant difference between *0-back* and *2-back* (p < .001) and between *1-back* and *2-back* (p < .001) condition. There was no significance reached between the *0-back* and *1-back* (p > .05) condition. An ANCOVA procedure showed no significant relations (p > .05) of self-reported motivation, age, self-reported wakefulness, and Digit Span on the error rate among task difficulties.

Since normal distribution was only partly confirmed and a Mauchly Test did not confirm sphericity (p < .001), the non-parametric Friedman Rank Sum Test was also conducted. In line with the reported ANOVA, the Friedman Test also indicated a significant effect (χ^2 (2) = 33.724, p < .001). Here, post-hoc Pairwise Comparisons (HSD Tukey correction) also showed only a significant difference between *0-back* and *2-back* (p < .001) and between *1-back* and *2-back* (p < .001). Significant correlations ("Pearson") were found for error rate and the overall TLX score among all task difficulties (*0-back*: r^2 (23) = .47, p < .05; *1-back*: r^2(23) = .54, p < .01; *2-back*: r^2(23) = .42, p < .05). Further, error rate and the Stressindex correlated significantly for two task difficulties (*0-back*: r^2(23) = .46, p < .05; *2-back*: r^2(23) = .65, p < .001).

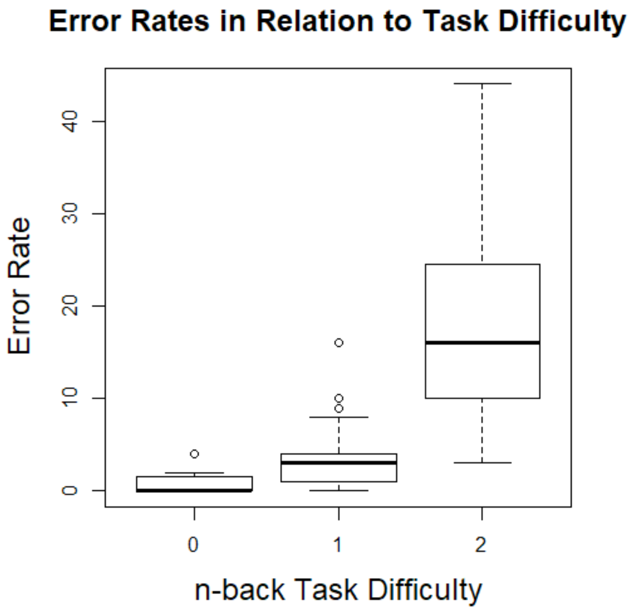


Figure 17. A Boxplot graph showing the distribution of the error rate per *n-back* task difficulty.

The sixth hypothesis ("Reaction Time increases with increasing Task Difficulty") deals with the effect on CL on the reaction time (in ms) per task difficulty. For analysis purposes, the reaction time was averaged per task difficulty per participant. *Table 6* illustrates the descriptive analysis. Since participants had to respond to the test stimuli within 1500 ms, reaction times could range from *MIN*=0 and *MAX*=1500 ms. In case a participant did not respond in time, the maximum reaction time was used. In the case of more than one inputs, the reaction time for the first response was included only. *Figure 18* depicts the averaged distribution of reaction times (in ms) among *n-back* task difficulties.

	n	***MIN***	***Q1***	***Median***	***Q3***	***MAX***	***Mean***	***SD***
0-back		355.31	411.40	439.62	497.88	565.73	457.18	61.08
1-back	23	430.70	525.90	580.78	626.88	821.05	591.77	96.78
2-back		568.38	705.20	867.23	975.30	1078.76	844.03	149.61

Table 6. Descriptive Analysis of the averaged reaction times (in ms) per task difficulty. *n* = Sample Size, *MIN* = Minimum Value, *Q1* = 1st Quartile, *Q3* = 3rd Quartile, *MAX* = Maximum Value, *SD* = Standard Deviation.

The Shapiro Wilk Test confirmed a normal distribution for all task difficulties (*0-back*: $W = .95, p > .05$; *1-back*: $W = .97, p > .05$; *2-back*: $W = .94, p > .05$). A one-way ANOVA with repeated measures indicated a significant effect of task difficulty on reaction times ($F(2/44) = 17.101, p < .001, \eta^2 = .69$). Pairwise Comparisons (HSD Tukey correction) showed significant differences between all task difficulties (*0-back* and *1-back*: $p < .001$; *0-back* and *2-back*: $p < .001$; *1-back* and *2-back*: $p < .001$). An ANCOVA procedure showed no significant relations ($p > .05$) of self-reported motivation, age, and self-reported wakefulness on the reaction time among task difficulties. Only the covariate *Digit Span* was significantly related to reaction time ($F(1/44) = 4.660, p < .05, \eta^2 = .21$).

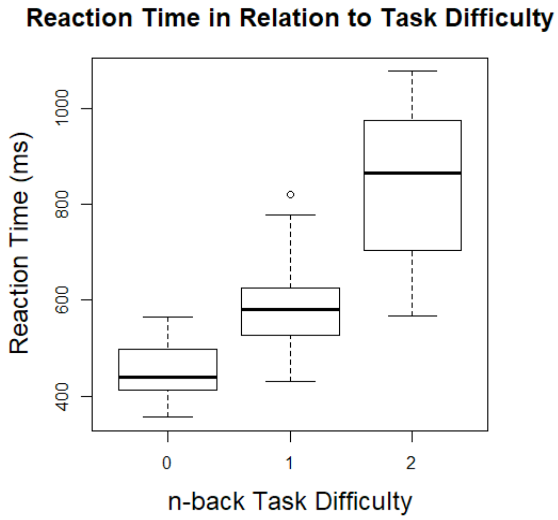


Figure 18. A Boxplot graph showing the distribution of the averaged reaction times (in ms) per *n-back* task difficulty.

Since Mauchly's Test indicated that the assumption of sphericity was violated (p < .05), a Friedman Rank Sum Test was also conducted. This non-parametric test of differences among repeated measures rendered a value of χ^2 (2) = 44.087, which was significant (p < .001). In line with the reported ANOVA, post-hoc Pairwise Comparisons (HSD Tukey correction) also found significant differences between all task difficulties: *0-back* and *1-back* (p > .01), *0-back* and *2-back* (p < .001), and between *1-back* and *2-back* (p < .01). Significant correlations ("Pearson") regarding reaction times were only found in the *2-back* condition. Here, reaction times correlate significantly with the Stressindex (r^2(23) = .43, p < .05), the RAW TLX (r^2(23)= .54, p < .01) and Error Rate (r^2(23)= .59, p < .01).

RQ3. The third research question ("Does Cognitive Load have an impact on the perceived Stress Level?") examines whether our CL manipulation has an impact on the subjective stress level that was reported retrospectively after completing each task difficulty by filling out the PASA questionnaire. The overall Stressindex was calculated using an EXCEL® template that was provided by the PASA developer (Gaab, 2005). Hence, we have an overall self-reported stress value per task difficulty per participant.

Table 7 illustrates the descriptive analysis of the overall Stressindex per task difficulty. The overall Stressindex is calculated by subtracting the *Primary Appraisal* from the *Secondary Appraisal*. Hence, the overall Stressindex can range from *MIN*=-5 to *MAX*=5.

	n	***MIN***	***Q1***	**Median**	***Q3***	***MAX***	**Mean**	***SD***
0-back	25	-5	-2.88	-2.38	-2	-0.38	-2.50	1.08
1-back		-4.13	-2.5	-2	-1.38	-0.13	-2.07	0.92
2-back		-2.88	-1.88	-0.88	-0.13	0.75	-0.99	0.99

Table 7. Descriptive Analysis of the overall Stressindex per task difficulty. *n* = Sample Size, *MIN* = Minimum Value, *Q1* = 1st Quartile, *Q3* = 3rd Quartile, *MAX* = Maximum Value, *SD* = Standard Deviation.

Figure 19 depicts the self-reported stress level per task difficulty. We proposed the hypothesis that the perceived stress level increases with increasing task difficulty. To test the hypothesis, a Shapiro Wilk Test confirmed a normal distribution of the overall Stressindex values for all *n-back* conditions (*0-back*: $W = .95, p > .05$; *1-back*: $W = .99, p > .05$; *2-back*: $W = .97, p > .05$). A one-way ANOVA for repeated measures revealed a significant difference for the self-reported stress level between task difficulties ($F(2/48) = 19.445, p < .001, \eta^2 = .30$). Post-hoc Pairwise Comparisons (HSD Tukey correction) showed significant differences between *0-back* and *2-back* ($p < .001$) and also between *1-back* and *2-back* ($p < .001$), which was not the case for *0-back* and *1-back* ($p > .05$).

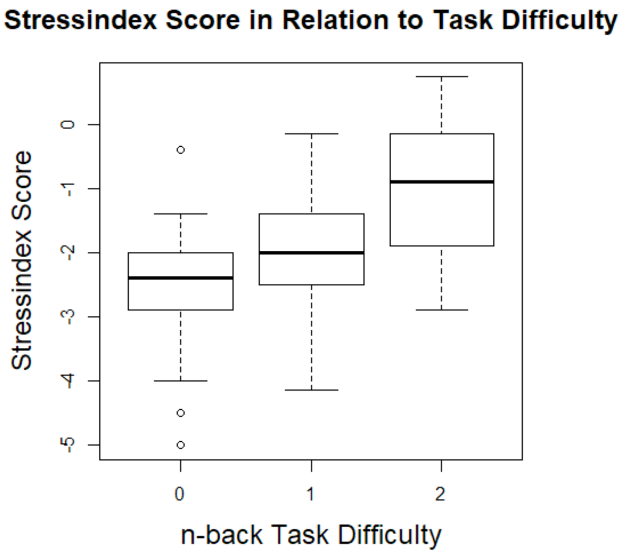


Figure 19. A Boxplot graph showing the distribution of the overall Stressindex per *n-back* task difficulty.

Since the Mauchly Test indicated that the assumption of sphericity has been violated ($p < .05$), a non-parametric test was also conducted. A Friedman Rank Sum Test of differences among repeated measures was conducted and rendered a value of $\chi^2(2) = 24.869$, which was significant ($p < .001$). Similar to the reported ANOVA, Pairwise Comparisons (HSD Tukey correction) indicated a significant effect of task difficulty on the overall Stressindex between *0-back* and *2-back* (p < .001) and also between *1-back* and *2-back* (p < .001). Significant correlations ("Pearson") showed significant relations between the overall Stressindex and the RAW TLX score within all task difficulties (*0-back*: r^2(25) = .59, p < .01; *1-back*: r^2(25) = .45, p < .05; *2-back*: r^2(25) = .48, p < .05).

RQ4. Our fourth research question ("Does Cognitive Load have an impact on Emotional States?") examines the effect of the CL manipulation on the self-reported emotional state using the SAM dimensions. Hence, every participant rated three dimensions (valence, arousal, and dominance) after completing each task difficulty. The provided figures included a scale from *MIN*=1 to *MAX*=9 per dimension.

		n	*MIN*	*Q1*	Median	*Q3*	*MAX*	Mean	*SD*
Valence	***0-back***	25	4	6	7	8	9	6.96	1.14
	1-back		2	6	7	7	9	6.60	1.47
	2-back		3	5	6	7	8	5.92	1.47
Arousal	***0-back***	25	1	2	3	5	7	3.32	1.91
	1-back		1	2	4	4	7	3.56	1.47
	2-back		1	2	5	6	7	4.28	2.07
Dominance	***0-back***	25	3	5	6	8	9	5.88	2.11
	1-back		3	5	6	7	9	6.04	1.79
	2-back		2	5	5	7	9	5.40	1.98

Table 8. Descriptive Analysis of the SAM Dimensions (Valence, Arousal and Dominance) per task difficulty. *n* = Sample Size, *MIN* = Minimum Value, *Q1* = 1st Quartile, *Q3* = 3rd Quartile, *MAX* = Maximum Value, *SD* = Standard Deviation.

Table 8 illustrates the descriptive analysis regarding self-reported emotional states using the SAM dimensions per task difficulty.

Figure 20 depicts each SAM dimension for each task difficulty. To test our eighth hypothesis ("Perceived Valence decreases with increasing Task Difficulty"), we conducted a Shapiro Wilk Test that did not confirm the assumption of a normal distribution for all SAM dimensions (*0-back*: $W = .91$, $p < .05$; *1-back*: $W = .91$, $p < .05$; *2-back*: $W = .93$, $p < .05$). The Mauchly Test confirmed the assumption of sphericity ($p > .05$). A one-way ANOVA for repeated measures revealed a significant effect for task difficulty on self-reported Valence ($F(2/48) = 4.52$, $p < .05$, $\eta^2 = .09$). Pairwise Comparisons (HSD Tukey correction) rendered a significant difference of Valence ratings only between the *0-back* and *2-back* condition ($p <$ .05). Since none of the task difficulties were normally distributed, the non-parametric Friedman Rank Sum Test of differences among repeated measures was conducted and revealed a value of χ^2 (2) = 13.286, which was also significant ($p < .01$). Similar to the reported ANOVA, post-hoc Pairwise Comparisons (HSD Tukey correction) only showed a significant difference between the task difficulties *0-back* and *1-back* ($p < .05$).

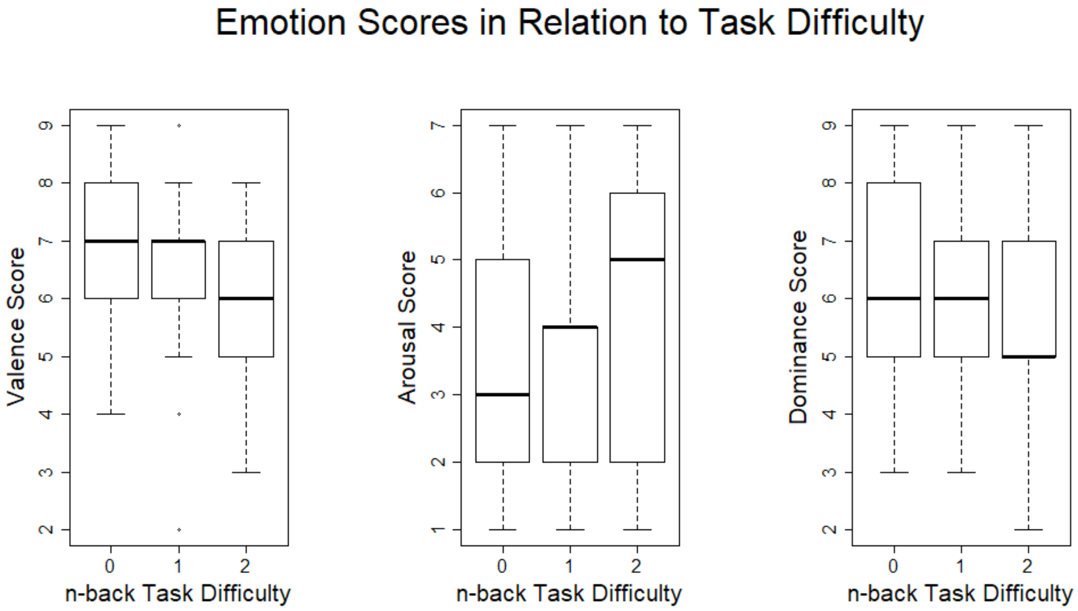


Figure 20. A Boxplot graph for each SAM Dimension (Valence, Arousal and Dominance) showing their distribution per *n-back* task difficulty.

There was one significant correlation (“Pearson”) found in the *2-back* condition for self-reported Valence and Stressindex ($r^2(25) = -.42$, $p < .05$). Further, Valence ratings and number of blinks correlated significantly in the *0-back* condition ($r^2 = -.46$, $p < .05$).

To test the following hypothesis (“Perceived Arousal increases with increasing Task Difficulty”), a Shapiro Wilk Test was conducted. It confirmed a normal distribution only partly (*0-back*: $W = .88$, $p < .01$; *1-back*: $W = .93$, $p > .05$; *2-back*: $W = .87$, $p < .01$). A Mauchly Test confirmed the assumption of sphericity ($p > .05$). A one-way ANOVA for repeated measures showed a significant effect for task difficulty ($F(2/48) = 4.001$, $p < .05$, $\eta^2 = .05$). Post-hoc Pairwise Comparisons (HSD Tukey correction) indicated that Arousal ratings differed significantly only between the *0-back* and *2-back* condition ($p < .05$). Since the normal distribution was only partly confirmed, a non-parametric test was also conducted. Contrarily to the reported ANOVA, a non-parametric Friedman Rank Sum Test of differences among repeated measures was conducted and rendered a value of $\chi^2\,(2) = 4.5455$, which was not significant ($p > .05$). Further, Arousal ratings correlated negatively with the Valence ratings only in the *0-back* condition ($r^2(25) = -.55$, $p < .01$). In the *2-back* condition, Arousal ratings and the RAW TLX score correlated positively ($r^2(25) = .58$, p < .01).

The last hypothesis (“Perceived Dominance decreases with increasing Task Difficulty”) tested whether self-reported Dominance values increase with increasing CL via task difficulty. A Shapiro Wilk Test confirmed a normal distribution only for the *1-back* and *2-back* condition (*0-back*: $W = .89$, $p < .05$; *1-back*: $W = .95$, $p > .05$; *2-back*: $W = .92$, $p > .05$). A Mauchly Test confirmed sphericity for task difficulty ($p > .05$). A one-way ANOVA indicated no significant difference of Dominance ratings among task difficulty ($F(2/48) = 2.202$, $p > .05$, $\eta^2 = .02$) Since normal distribution was only partly confirmed, a non-parametric test was also conducted. Similar to the reported ANOVA, a Friedman Rank Sum Test rendered a value of $\chi^2\,(2) = 3.377$, which was not significant ($p > 0.05$). Hence, no post-hoc analyses were conducted. Further, a correlation ("Pearson") matrix revealed that in the *0-back* and *1-back* condition Dominance ratings and the self-reported Stressindex correlated significantly (*0-back*: $r^2(25) = -.42$, $p < .05$; *1-back*: $r^2(25) = -.50$, $p < .05$).

Discussion

In the following, the presented results of this interdisciplinary work are discussed. The postulated RQs include related findings that support the interpretation of our empirical data. Further, limitations and future work ideas for each RQ are included. Since this pioneer work examined the suitability of VR Technology to measure CL, it is discussed in detail.

RQ1

In order to examine our first RQ ("Does the *n-back* task induce Cognitive Load?") whether the *n-back* task is an appropriate method to induce CL, we postulated four hypotheses dealing with different CL parameters, combining objective and subjective methods.

Assessing the pupil diameter during cognitive processing revealed a mean of *MIN*=4.84 and *MAX*=5.51 mm. This seems to be in line with other pupil diameter measurements. Chen and Epps (2014) reported an average pupil size change of 0.1604 mm and 0.5352 mm for different tasks, which is in line with our findings ($\Delta_{0\text{-}1}$=0.16 and $\Delta_{1\text{-}2}$=0.67 mm). But it has to be mentioned that the authors did not find a significant increase in the pupil's diameter change among task difficulty (Chen & Epps, 2014). Scharinger, Soutschek, Schubert, and Gerjets (2015) also used a *0-back*, *1-back,* and *2-back* task to induce CL. They measured similar, but slightly bigger averaged pupil sizes (*0-back*: 5.59 mm; *1-back*; 5.72 mm; *2-back*: 6.14 mm) with considerable small standard deviations. This empirical consistency suggests that our pupil diameter measurements are ranged realistically. Even though mean or median values are close together among task difficulties, the ANOVA revealed a significant diameter's increase between *0-* and *2-back* and *1-* and *2 back*. These findings indicate that we successfully manipulated low (*0-back*) and high (*2-back*) CL. Medium CL (*1-back*) was not confirmed empirically. Results suggest that *0-back* and *1-back* demand similar mental effort, hence including only one of them can be used as a sufficient low CL manipulation. Another study using the same *n-back* levels (*0-back*, *1-back,* and *2-back*) could find a main effect of the *n-back* level with a considerable effect size (η^2= .73) for pupil size (Scharinger et al., 2015). Contrarily to our findings, they found a significant increase in the pupil size from *0-back* to *1-back* (even though the change between *1-back* and *2-back* was also bigger than between *0-back* and *1-back*). Even though our effect size (η^2= .07) indicates a medium to large effect, according to Cohen (1988), a non-parametric test (no violation of requirements) showed no significant increase of the pupil's diameter among task difficulty. Apparently, it depends on the chosen statistical analysis whether the effect reaches significance or not. Further empirical work or different statistical analyses are needed to deeper understand the relation between CL and pupil dilation since previous research has provided mixed findings. Several studies suggest that an increase of pupil size indicates increasing cognitive processing (e.g., Chen & Epps, 2013; Krejtz, Duchowski, Niedzielska, Biele & Krejtz; 2018; Pomplum & Sunkara, 2003). Contrarily, other empirical work could not confirm this relation (e.g., Chen & Epps, 2014). Our partly inconclusive findings can be due to inaccurate measurements or an unsuccessful CL manipulation. On the other side, it can also be more accurate than other results due to the novel VR approach (better-controlled light conditions). Thus, our results may reflect study outcomes that see the light reflex as being mainly responsible for pupil size changes (e.g., Kramer, 1990; Pomplum & Sunkara, 2003). Similar to Duchowski and colleagues (2018), we also included working memory capacity (using the *Digit Span Memory Test*) as a covariate. Similar to their IPA calculations, our results also suggest that the individual working memory capacity did not have an impact on the pupil's diameter among task difficulty. Even though working memory capacity is related to several higher-order cognitive tasks (Duchowski et al., 2018), this could mean that the pupil diameter successfully discriminates between task difficulties independently of the individual working memory capacity. Further, included third variables, such as self-reported motivation and wakefulness, age, and *Digit Span* did not have an impact on the pupil's diameter among task difficulty. This could also arise from small ranges within our study sample consisting of students only. For instance, self- reported motivation indicates that all participants were "very" or "either motivated" and relatively young (*M*=22.76*; SD*=1.69).

There was a significant positive correlation found for the pupil diameter and the overall Stressindex, but only within the *2-back* condition. This could mean that the pupil diameter increases with increasing perceived stress but only when high CL is induced. This result supports the assumption that CL and stress are two constructs that influence each other (without assuming causality). Moreover, pupil diameter and self-reported Dominance correlated negatively ($p < .05$) only within the *0-back* condition. This means that the bigger the pupil diameter, the lower the Dominance rating, and vice versa. Hence, the pupil diameter could be an indirect indicator of perceived Dominance under the condition that either no or low CL is induced. Since this relation is only found when inducing low CL, it is unlikely that there are systematical relations among task difficulty. An explanation could be that the pupil dilation is more sensitive to CL than to emotions when higher CL is induced (competition for working memory resources), and this is why this effect is only found in the lowest CL condition.

Interestingly, Tullis and Albert (2013) see the pupil diameter as a *Usability* metric for measuring the user's arousal level. This could not be confirmed by our data since no significant relation was found. All in all, our first hypothesis can only be partly confirmed, since we only found a significant increase of the pupil diameter for low (*0-back*) and high (*2-back*) CL and medium (*1-back*) and high (*2-back*) CL.

The number of blinks within each task difficulty was also calculated since it is seen as an indicator of cognitive processing (Irwin & Thomas, 2010; Ledger, 2013). Unfortunately, the descriptive statistics reveal high standard deviations, which can indicate a biased and incorrect blink detection. Hence, it is highly unlikely that within the lowest task difficulty (which reflects a normal state the most), the accumulated number of blinks ranges from 2 to 977 times. On the other side, research also shows a ranging mean of blinks per minute. For instance, Irwin and Thomas (2010) reported a blink rate of 2 - 4 times per minute, whereas Portello, Rosenfield, and Chu (2013) published a blink rate average of 11.6 blinks (but with an *SD* = 7.84) per minute. Further, the latter is probably underestimated since participants had a reading task on a desktop computer, which can cause the *Computer Vision Syndrome*, hence a reduced number of blinks (Portello, Rosenfield & Chu, 2013). Another methodical challenge is to control for voluntarily made blinks that are executed consciously. So, it is also possible that our data is not biased but underlies several confounding factors and great inter-individual differences. Not surprisingly, parametric (ANOVA) and non-parametric (Friedman Rank Sum Test) analysis showed no significant differences regarding the blink rate among task difficulty. There are also controversial views on how the blink rate reflects cognitive processing in general. Some researchers suggest a reduced blink rate (e.g., Irwin & Thomas, 2010; Ledger, 2013), others found an increase (e.g., Chen & Epps, 2013; Tanaka & Yamaoka; 1993) with increasing mental demand. Our dispersed data can only provide a small tendency that the blink rate decreases with increasing task difficulty since the number of blinks is slightly higher for low CL (*0-back*) compared to medium CL (*1-back*) and high CL (*2-back*).

Future work should have a closer look at blinks' timestamps within a trial run since Stern, Walrath, and Goldstein (1984) found a blink boost at the beginning and end of cognitive processing, which could explain mixed findings. Another potential explanation is the type of chosen task. Hence, it would be interesting to compare blink rates of different tasks, such as visual search, arithmetic, and semantic tasks. All in all, our results do not confirm our second hypothesis.

We also integrated a new analysis method introduced by Duchowski and colleagues (2018) called *Index of Pupillary Activity* (IPA). This approach is a measure of the rate of change of pupil diameter and not based on averaged differences. The authors claim that this method is less influenced by light and reflex dilation (Duchowski et al., 2018). Descriptive statistics reveal a similar structure than the averaged pupil diameter: All IPA scores lie close to each other with a slight *U-curve* with unexpected low values for medium CL (*1-back*). Fortunately, the adaption of this novel method revealed realistic values since Duchowski and colleagues (2018) reported a similar range. Unfortunately, the authors did not mention the potential range of the IPA score; this would have helped to classify our results whether the IPA score reflects relatively low or high CL. The parametric and non-parametric analysis did not reveal a significant increase of the IPA score among task difficulty. Even though Duchowski and colleagues (2018) report a significant ANCOVA analysis for the IPA score on task difficulty, their effect size is rather small (η^2= .05). Further, one important step to calculate the IPA score includes cutting out 200 ms before and after a registered blink. Hence, a reliable blink detector is essential. Since our blink rate varies noticeably even within one task difficulty, vague blink detection could be the reason why the IPA score did not differentiate between task difficulties. Hence, we cannot confirm our third hypothesis, but this may be due to biased blink data.

The IPA is a novel method that needs to be examined in more studies to gain further insight regarding its suitability to detect CL. More empirical data and experience with the IPA method in the research field would make it easier to interpret results and evaluate if it reflects a promising approach for future work.

With our fourth hypothesis, we examined whether self-reported CL differentiates between task difficulties. Since we left out the *Physical Demand* dimension, the maximum RAW TLX score was *MAX*=100. Descriptive analysis indicates that participants reported relatively low CL for all task difficulties (*0-back*: *M*=25.12; *1-back*: *M*=35.72; *2-back*: *M*=63.92), but high standard deviations suggest high inter-individual differences. This is in line with a related work conducted by Luque-Casado, Perales, Cárdenas, and Sanabria (2016). They only included the *2-back* task that induced very similar subjective CL ratings using the TLX score (*M*= 68.1, *SD*=14.4). Interestingly, the *2-back* task-induced higher TLX scores in comparison to a vigilance and duration discrimination task (Luque-Casado et al., 2016). This confirms our choice of working memory task to induce CL. The descriptive distribution shows an increase of the averaged RAW TLX score with increasing task difficulty. This was confirmed by parametric and non-parametric analyses. Contrarily to the pupil diameter, post-hoc analyses revealed considerable differences between all task difficulties. According to Cohen (1988), the effect size (η^2= .52) indicates a very large effect of the subjective CL among task difficulty. This is even more than Duchowski and colleagues (2018) reported (η^2= .38) within a similar study design. Hence, our data confirm that the RAW (without the weighting procedure) TLX score was sensitive to task difficulty. Research suggests that the TLX score correlates with the error rate (e.g., Grigg, Garrett & Benson, 2012). Our data confirm that, since significant, positive relations were found within all task difficulties. All in all, our data strongly confirms our fourth hypothesis that the subjective CL increases with increasing task difficulty.

Our first RQ ("Does the *n-back* task induce Cognitive Load?") tested with four hypotheses if the *n-back* task-induced CL, which indicates a successful CL manipulation. Our pupil diameter data seem to be realistic since we came to a similar value range than related empirical work. This is important because to answer the first RQ, our pupil data form the basis for further interpretations. Even though we cannot confirm all hypotheses, our data strongly suggest that the *n-back* task caused CL. This is in line with studies demonstrating that working memory tasks can be successfully used to produce CL (e.g., Zuo et al., 2019). But our data also reflects the ambiguous opinions in the research field about how many steps back should be used (see more details in Chapter *Performance Measures*.). In our case, high CL (*2-back*) was successfully induced. The averaged pupil diameter and subjective CL indicate that *0-back* and *1-back* seem to cause rather similar CL. This is also in line with three participants who asked about the study purpose after finishing the experiment. They reported that the *2-back* condition was definitely the most difficult in comparison to *0-back* and *1-back*. The difficulty rise between *0-back* and *1-back* was not so transparent for them. For future work, it seems enough to include either the *0-back* or *1-back* task to induce low CL. Another explanation could be that *0-back* and *1-back* address different neuronal networks, which could lead to biased findings. In the *0-back* condition, participants had to compare the very first letter with all that followed, whereas in the *1-back* condition, the last letter had to be compared to the current. So, it may be possible that *0-back* addresses more long-term than short-term neuronal networks (as it is for *1-back* and *2-back*). Another disadvantage of the *0-back* condition is that if the person forgets or misses the first letter, it is impossible to accomplish the trial. This could lead to frustration and highly biased physiological and performance measures. Hence, based on our results, it is recommendable to include *1-back* (low CL) and *2-back* (high CL) in future work.

Further, a control condition would be recommendable, to see how the pupil diameter reacts, when no further CL is induced. This way, a baseline value would help to interpret pupil diameter changes when modulating CL. Hence, it would be possible to assess the "pure" CL that is induced by the chosen study setting (e.g., using VR Technology). Including a control condition in a "within-subject" design is recommendable, since it controls for inter-individual differences.

If the same materials are used again, the *Pupil Capture*® Blink Detector should be tested in advance to review whether our spread blink data is due to measurement errors or due to inter-individual differences. As stated above, a control group would provide helpful baseline values to interpret our dispersed data additionally.

RQ2

To examine the second RQ ("Do Performance Measures reflect increasing Cognitive Load?") whether performance measures confirm the *n-back* task as our CL modulation, we proposed two hypotheses including error rate and reaction time.
The descriptive data of error rates among task difficulty suggests that participants were much more inaccurate in the high CL condition (*2-back*) than in the low (*0-back*) and medium (*1-back*) task difficulty. This is in line with our results regarding the pupil diameter, which underpins both findings and suggests a systematic relation between pupil diameter and error rate: Both parameters could significantly differentiate between condition *0-back* and *2-back* and *1-back* and *2-back*. Contrarily to our pupil diameter results, the non-parametric test showed the same significant differences. Interestingly, Scharinger and colleagues (2015) found similar patterns. Accuracy in the *2-back* condition differed significantly from the *0-back* and *1-back* task with no significant difference between the two latter. The great variance of error rates within the *2-back* task indicates high inter-individual differences when high CL

is induced. Similar to the pupil diameter, self-reported motivation and wakefulness, age, and working memory capacity (*Digit Span*) seem to have no impact on error rate.

In general, our experiment shows relatively small error rates among all task difficulties (*2-back* as highest: *M*=17.65, *SD*=10.80), when we consider *MAX*=120. Since our *n-back* task had a ratio of one-third matches and two-thirds no-matches, it can be ruled out that participants always answered the same (error rate for responding "yes": 80; error rate for responding "no": 40). Responding by chance would also result in higher error rates. The impression that participants made relatively few errors can be supported by Scharinger and colleagues (2015). When comparing their reported accuracy (*0-back*: 88%, *1-back*: 86%, *2-back*: 79%; *n* = 22) among *n-back* task difficulties with our data (*0-back*: 99%, *1-back*: 98%, *2-back*: 85%, *n* = 23), our participants seemed to be more accurate. This disparity could be due to longer trials (causing exhaustion) chosen by the authors or study design differences. For instance, the enclosed VR environment could reduce distraction from the surroundings. Another explanation of our lower error rates could be different practice approaches. Scharinger and colleagues (2015) let participants practice all *n-back* task difficulties at the beginning of the study until they reached at least 60% accuracy. Contrarily, participants of our study practiced separately in a fixed time window at the beginning of each task difficulty. The separated and more recent practice phase may be a reason why we report greater accuracy.

Examining relations between our measured variables revealed a significant relationship between error rate and overall RAW TLX score within all task difficulties. This means that the higher the error rate, the higher the subjective CL was reported. Further, there were significant correlations found for error rate and self-reported stress level within the *0-back* and *2-back* condition. Hence, the higher the error rate, the higher stress levels were reported. Interestingly, this relation was not significant for the *1-back* condition. This could be another indicator that this condition did not induce medium CL as expected. Moreover, our data revealed a great correlation between error rate and perceived stress level ($r^2(23) = .65$, $p < .001$) when inducing high CL (*2-back*). This supports our assumption that the *2-back* condition successfully induced high CL and hence, high stress levels.

The descriptive data of averaged reaction times shows an increase with increasing task difficulty. Similar to error rates, reaction times underlie high variance when inducing high CL (*2-back*). Mean values suggest a greater difficulty increase towards *2-back* than between the *0-back* and *1-back* task. Parametric, as well as non-parametric statistics, rendered a significant difference between all task difficulties with a considerable effect size (η^2= .69), which is slightly higher than the effect size reported by Scharinger and colleagues (2015) who also found reaction times significantly increasing with increasing *n-back* difficulty. Comparing with this related study design (*0-back*: 462 ms, *1-back*: 506 ms, *2-back*: 632 ms), our averaged reaction times seem to be similar to their *0-back* and *1-back* condition, but our participants reacted about 200 ms slower in the *2-back* condition.

Similar to the previously reported ANCOVA results, self-reported motivation, wakefulness, and age did not have an impact on how fast participants respond to test stimuli. Interestingly, *Digit Span* was significantly related to reaction times. Hence, the individual working memory capacity (*Digit Span*) has an impact on how fast the individual responded in our study. Intuitively, this must also be the case for error rates, which is not the case for our data. Since it seems that research provides mixed results (e.g., Duchowksi et al., 2018), further research is needed to understand to which extent individual working memory capacity has an impact on the performance of several higher-order cognitive tasks. It could be possible that different manipulations of working memory capacity (here *Digit Span*) could lead to different conclusions: Cognitive tasks may activate distinct neuronal networks, which is why working memory capacity seems to have an impact only on some mental tasks.

Reaction times correlated significantly with some variables but only within the *2-back* condition. When inducing high CL (*2-back*), the longer the reaction time, the higher stress levels were reported. This result is in line with the empirical work of Chen and colleagues (2016) that could also observe that a performance decline can cause stress and negative affect. An even higher relation was found for reaction time and the overall RAW TLX score: the longer the reaction time, the higher subjective CL was reported.

Our RQ2 (“Do Performance Measures reflect increasing Cognitive Load?”) examined whether performance measures support our CL manipulation. Comparing our findings with related empirical work suggests a successful measurement of error rates and reaction times since our data is within the same range. This is also the case for reported performance measures of the preceding work conducted by von Bauer (2018). Significantly increasing reaction times among *n-back* task difficulty confirms that the *n-back* task successfully induced CL. Error rates also support our CL manipulation, except for the *1-back* condition. Hence, performance measures reflect our RQ1 findings: low CL (*0-back*) and high CL (*2-back*) were successfully induced, whereas the *1-back* condition does not seem to be the right choice to induce medium CL. Furthermore, performance metrics support our pupil diameter metric as a CL assessment, since our performance measures and pupil diameter increased both (partly) significantly with increasing task difficulty. This is additionally supported by considerable relations with self-reported CL.

RQ3

To examine the third RQ (“Does Cognitive Load have an impact on the perceived Stress Level?”), we included a self-reported stress level of each task difficulty using the PASA questionnaire.

The descriptive data revealed a relatively small self-reported stress level with negative means. Hence, among all task difficulties, the perceived coping skills (*Secondary*

Appraisal) were greater than the perceived threat (*Primary Appraisal*). Wirtz and colleagues (2007) induced psychosocial stress (using the TSST) and reported a higher overall Stressindex (*M*=2.2, *SD*=2.4). This could be an indicator that the multidimensional construct stress can be divided into different types of stress. Hence, our study could have triggered stress that is performance-related by exceeding working memory capacity. Since the TSST combines performance (interview and arithmetic task) and social pressure (in front of a jury), it makes sense that in this study, higher PASA scores were reported (Wirtz et al., 2007). Hence, we would have observed higher stress levels if participants had to perform the *n-back* task visibly in front of (judging) people. Moreover, the PASA questionnaire seems to be sensitive to both types of perceived stress.

Similar to previous findings, parametric and non-parametric analyses showed that the self-reported stress level significantly differentiated between task difficulties except for medium CL (*1-back*) with a large effect size (η^2= .30). This means that the perceived stress level increased with increasing task difficulty when excluding the *1-back* condition. Furthermore, there were significant correlations between the perceived stress level and perceived CL within all task difficulties found. Hence, the higher the perceived CL, the higher the stress level was rated and vice versa. Hence, our assumed close relation between CL and stress is supported. This is in line with several empirical works regarding CL or stress that use the same metrics, for instance, GSR or HR(V). But it remains a methodical obstacle to differentiate between both psychophysical constructs objectively.

Our third RQ ("Does Cognitive Load have an impact on the perceived Stress Level?") examines the influence of induced CL on the perceived stress level. Similar to RQ1 and RQ2, self-reported stress levels increased with increasing task difficulty significantly, but not between low (*0-back*) and medium (*1-back*) CL. Hence, our seventh hypothesis can be (partly) confirmed. Data show that participants were not very stressed in general (all mean

Stress level scores below 0). This can be explained by the *Transactional Stress Model* by Lazarus and Folkman (1984): Conducting the *n-back* task was not evaluated as a threat or challenge: A good performance in our study was not important to them, and no severe consequences were expected in case of bad outcomes (*Primary Appraisal*). Another explanation could be that even though the *n-back* task was evaluated as stressful, the belief in their skills to pass successfully (*Secondary Appraisal*) was greater. Low values could also be the result of the retrospective assessment. Participants may have reported higher stress levels before (successfully) completing each task difficulty. This small change in future studies could provide a more realistic self-reported stress level.

Several empirical studies demonstrate that CL influences the perceived stress level of an individual. Hart (2006), who developed the NASA TLX questionnaire to measure subjective CL, sees stress as one aspect of the human cost that is caused by CL. But different understandings and manipulations of both constructs make it difficult to comprehend the underlying mechanisms and distinction between both constructs. Also, it might be that the physiological stress reaction plays a mediating role when measuring CL: CL triggers the physiological stress reaction that activates the ANS that is (also) responsible for higher pupil dilation (Pedrotti et al., 2017). This would also explain why the same parameters are used to measure both CL and stress in research (e.g. GSR signal).

One possible explanation for the relation between CL and stress could be that CL causes bad performance outcomes that, in turn, trigger the stress reaction. Future work could include two groups whereas one group gets immediate negative feedback (manipulated) on performance and the other group not. Differences in the perceived stress level between both groups could indicate a moderator role of performance (evaluation) between CL and stress. Within this study design, physiological metrics used in CL and stress research (e.g., HR or GSR) could provide further information about their sensitivity to both constructs: If the "stress" group shows higher GSR values than the "no stress" group, it could be an indicator that this metric is more sensitive to stress than CL.

With our study design, we could observe the impact of CL on self-reported stress levels. But we cannot provide any deeper understanding of how stress affects cognitive processing. Since stress causes neuroendocrine changes in the brain and body, it seems very plausible that stress also impacts CL. But mixed findings make it hard to come to clear assumptions. Some studies support the view that stress enhances cognitive abilities; others observed a performance decrease when stress was induced. These opposite results could be explained by the numerously replicated *Yerkes-Dadson Law*: They could observe a U-shaped relation between the arousal level and performance where low and high arousal lead to reduced performance, and a medium arousal state generates the optimal performance (Yerkes & Dodson, 1908). Future work could add a "stress" group and compare performance metrics with a "no-stress" group. Stress could be manipulated by using the TSST procedure (see Chapter *Stress Modulation & Measurement*. for more information) that demonstrably induces (psycho-social) stress. Instead of an arithmetic task, the *n-back* task could be performed in front of a judging committee. The "no-stress" group would perform the *n-back* task within the same setting but without any audience. Here, results could further demonstrate how stress influences performance. One of the few published studies focusing on the differentiation between CL and stress developed a similar study design. Conway and colleagues (2013) let participants complete math tasks (three difficulty levels). First, with the information that their performance was not of interest. After that, they were told that they would be monitored, had limited time to solve the math tasks and immediate feedback was provided (Conway et al., 2013). Statistical analyses of GSR values showed no sensitivity regarding task difficulty, but the "stress" condition showed considerable higher GSR values than the "no stress" condition. Hence, GSR seems to be a more adequate metric to measure stress than CL. It would be from interest to replicate their study design and use other physical metrics such as eye-related parameters. This replication could provide deeper knowledge about the sensitivity of eye-related parameters regarding CL and stress.

All in all, most people have a common understanding of stress and CL, but research has its challenges to differentiate and hence better understand both constructs. Empirical work shows that both share same psychophysical outcomes, which demonstrates the “nearness” of both. Further research and a common “scientific” understanding of both constructs are needed to differentiate between them effectively. Moreover, researchers should be sensitized about potential overlaps with other psychophysical constructs when measuring CL and interpreting outcomes. As stated by Conway and colleagues (2013): “A major challenge for CL detection is the presence of stress, which may affect physiological measurements in ways that confound reliable detection of CL” (p. 659).

RQ4

Our last RQ (“Does Cognitive Load have an impact on Emotional States?”) investigates whether CL influences the emotional states of participants. Emotions were assessed by using the widely-established SAM dimensions.

The Valence dimension measured the current self-reported positive (high value) or negative (low value) affect. Descriptive statistics show relatively high values with a slight decrease in Valence ratings with increasing difficulty. Statistical analyses indicated only a significant difference between ratings of the *0-back* (low CL) and *2-back* (high CL) condition. This finding is again in line with previous findings suggesting that the *1-back* condition did not induce medium CL.

Duchowski and colleagues (2018) also assessed self-reported Valence in a control, easy, and difficult task (arithmetic tasks) condition. The authors report slightly lower values and similar effects: After completing the difficult task condition, Valence ratings were significantly lower than after the easy task and control condition, but the reported effect size (η^2= .03), suggests a rather small effect (Duchowski et al., 2018). Duchowski and colleagues (2018) could not confirm a significant decrease of self-reported Valence between the control and easy task condition, which may be consistent with our findings. These results suggest that a task difficulty only has an impact on valence on higher levels, not if task difficulty is low. Unfortunately, negative correlations between Valence and performance measures rise with increasing task difficulty but do not reach significance in our study. Adding even more difficult tasks, such as a *3-back* condition, could provide more conclusions.

Moreover, performance could play a mediating role: The more difficult a task is, the more performance is reduced, which in turn leads to a negative emotional state. An interesting study conducted by Raaijmakers and colleagues (2017) examined the effect of manipulated performance feedback on self-reported invested mental effort after a problem-solving task. Their findings suggest that when given negative feedback, self-reported mental investment was rated significantly higher than when receiving positive feedback (Raaijmakers et al., 2017). This underpins our decision to only include feedback for practice purposes and could support the idea that performance (evaluation) plays a key role between CL and subjective reports. Furthermore, it should be considered that feedback can cause biases when measuring self-reported CL. Hence, our data suggest an effect of CL on perceived valence when a certain level of cognitive processing is reached, but further research is needed to gain a deeper understanding of this relation.

Arousal describes a state of being physically alert, awake, and attentive that was rated after completing each task difficulty in our study. Even though descriptive analyses show a slight increase in Arousal ratings with increasing task difficulty, high fluctuations can be noticed. Similar to Valence ratings, Arousal ratings only differed significantly between the low (*0-back*) and high (*2-back*) task difficulty. As stated above, it would be of interest to include greater task difficulties and a control group in future work to see whether CL has an impact on Arousal ratings only if a certain level of mental demand is addressed. Supported is this assumption by a significant correlation between Arousal ratings and subjective CL when high CL (*2-back*) is induced. Hence, the higher the subjective CL is rated, the higher is the self-reported Arousal level and vice versa only when high CL is induced. As it could be the case for valence, performance (evaluation) could play a mediating role between CL and Arousal ratings. But our arousal-related results have to be interpreted cautiously since a non-parametric test rendered no significance even though its requirements were not violated. Duchowiski and colleagues (2018) also included Arousal ratings and report very similar mean values for the easy (*M*=3.21) and difficult (*M*=4.18) task condition, which indicates a reliable data acquisition within our study. A study conducted by Li, Markkula, Li, and Merat (2018) investigated the effect of CL (driving) on physical arousal since CL seems to improve lane-keeping performance (keeping the car on the street lane). Their results suggest that CL leads to increased arousal, which improved driving performance (Li et al., 2018). This supports the assumed relation between CL and arousal level. But it has to be stated that they quantified Arousal with the SCL (*Skin Conductance Level*), which is also a popular metric to measure cognitive processing or stress (e.g., Nourbakhsh et al., 2012). With this background, it seems plausible that some researchers see stress as a form of negative emotions, for instance, arousal (Plass & Kalyuga, 2019), which underpins the closeness between these multidimensional constructs. Once again, further research and common construct understanding (e.g., differentiation between arousal and stress) are needed to come to clearer conclusions regarding the effect of CL on (physical) arousal.

Dominance describes an emotional reaction of superiority that was rated after each task difficulty in our study. Similar to the other SAM dimensions (valence, arousal) we can observe high fluctuations of Dominance ratings. Similar to other reported findings, ratings of the low (*0-back*) and medium (*1-back*) CL showed almost equal Dominance ratings that decreased slightly when high (*2-back*) CL was induced. Thus, it is not surprising that statistical analyses did not render significant differences. Interestingly, there were significant correlations found for Dominance ratings and stress ratings when low (*0-back*) and medium (*1-back*) CL were induced. This means that the higher stress was rated, the lower Dominance was rated, and vice versa. This finding is contrary to Valence and Arousal ratings that seem to be influenced only when high (*2-back*) CL was induced. Similar study designs leave out the Dominance scale, which could be a sign that they did not find comprehensive data neither (e.g., Duchowski et al., 2018). In our study, three participants also had questions about this dimension, which could be an indicator that there were comprehension biases that also led to these inconclusive results.

Our data on self-reported emotions suggest comprehensive findings only partly. High variances of answers between participants indicate high inter-individual differences. In general, rather low emotional changes were registered. Since we did not put additional performance pressure in the form of immediate feedback, social evaluation or a reward/penalty on participants, there was not much at stake for people. Hence, this might be one reason for the relatively low effect on emotions in our study. Increasing performance pressure could provide more evidence in case stronger emotional responses are registered. This is in line with Brosch, Pourtois, and Sander (2010) who emphasize that the mental evaluation of a situation defines and precedes the release of an emotion.

Nevertheless, there was an interesting pattern found. Findings indicate the consequences of CL on the emotional state, particularly when high mental demand is induced. As stated above, this could be due to a reduced appraisal of importance to perform well, and therefore, a higher level of mental demand is needed to have an impact on the emotional state. Besides manipulating the importance to perform well (as illustrated above), future work should include a more heterogeneous sample since ours only contained relatively young students (*M*=22.76 years; *SD*=1.69), which could also lead to biases regarding the effect of CL on the emotional state (for instance, outstanding emotion-regulation skills).

Other studies also show how cognitive processing effects the emotional state (e.g., Duchowski et al., 2018). Especially in HCI, this is of interest. When developing a *User-Centered Design*, it is important to consider the emotional impact since it can influence the overall success of a product (Butz & Krüger, 2017). Norman (2004), a leading *Usability* professional, even argues that the emotional part of a design is more crucial to a product's success than the practical side. Hence, it is essential to understand the underlying mechanisms of triggered emotions not only for HCI research but also for the practice field to establish a holistic approach when measuring *Usability* and *User Experience*.

For Cognition research, the impact of emotions on cognitive processing is also from interest. Li and Ouyang (2012) could observe an effect of emotion on working memory, only when high CL (also *2-back*) was induced. This is in line with our results, suggesting that the interaction between emotions and cognitive processing is less visible until high mental effort is addressed. Several studies indicate that positive emotions enhance learning and cognitive processing (e.g., Panasati et al., 2018), which is also of interest for HCI research and practice since this can affect the product's handling and in turn, *Usability* evaluations of users.

The role of emotions when measuring CL should not be underestimated. Our study and several other empirical work suggest that CL has an impact on emotions (e.g., Panasati et al., 2012), which in turn affects cognitive processing (e.g., Li & Ouyang, 2012). Furthermore, emotional responses also have been examined on a physical level using metrics, such as eye-related or EEG parameters (e.g., Li & Ouyang, 2012; Bradley et al., 2008). Bradley and colleagues (2008) could observe an effect of emotion processing on the pupil size, which should be definitely considered in CL research (Bradley at al., 2008; Partala & Surakka, 2003). To understand better the role of emotions on CL, future work could include an *Emotion n-back task* condition, as used by Panasati and colleagues (2012) by integrating positive, negative, and neutral images between *n-back* stimuli. As stated above, a control group would provide valuable baseline observations, additionally. This design could provide further insights into what proportion CL and the emotional state explain effects on physical (e.g., pupil size) and performance measures.

Limitations

Limitations of the presented study design are partly mentioned above already. In general, a lack of a control group makes it difficult to interpret our results but would have required more money and time resources. A control group would provide baseline values that could help to understand better some of our findings. For instance, a baseline pupil size could have provided information about the CL (or stress level) only caused by the VR environment or unsatisfied expectations since VR Technology was completely new to most participants.

Due to upcoming *Cybersickness* symptoms, we let participants fill out the questionnaires outside the VR environment. This way we provided short breaks. But mounting the VR headset and starting a new calibration before each task difficulty could have led to measurement inaccuracies.

Further, the room conditions could have led to measurement biases. When participants filled out questionnaires, the study conductor was present. Even though the study conductor turned her back to participants demonstratively to provide a sense of privacy, this may have provoked *social desirability* biases. Also, we did not control the questionnaire fill out order. Most of the time, participants started with the NASA TLX, followed by the PASA and finally SAM dimensions. It is possible that this order influenced given answers of the following PASA and SAM questionnaire.

Just as most academic studies, a very homogeneous sample forms the basis for the presented findings. This is very usual since it is very difficult for academic researchers to address and recruit potential participants outside the academic context. Even though this challenge is very common, it is important to keep in mind that (our) empirical findings are based on young students, and hence, population conclusions should be made cautiously.

Within our study design, we wanted to gain deeper insights on how to differentiate between CL, stress, and emotions. We could observe that task difficulty has an impact on CL, perceived stress, and emotions, and significant relations between them were found partially. These findings support the intuitive closeness between all constructs. We would have gained far more insight, when stress and/or the emotional state would have been manipulated, too. As mentioned above, this could have been realized by including a "stress" group or an additional *Emotional n-back task* to deeper investigate the effect of both constructs on CL. But these implementations would have required considerable higher monetary and time resources and, therefore, should be seen as potential future work. All in all, it remains an empirical obstacle to differentiate between these multidimensional constructs in a deeper way since research shows a multi-directional relation between them. For instance, (high) cognitive processing seems to cause stress, and in turn, perceived stress impacts cognitive processing. This complicated relation becomes apparent when taking a look at the use of physiological metrics: There are CL (e.g., Nourbakhsh et al., 2012), stress (Perala & Sterling, 2007) and emotion (Shivakumar & Vijaya, 2015) studies that all use GSR signals to measure different constructs. Nevertheless, research should seek to gain deeper insights even though there are methodical challenges to overcome.

VR Technology – A New Promising Approach?

Inducing CL within a VR environment was a new approach to address measurement obstacles (luminance and head movement biases) when using *Task-Evoked*

Pupillary Responses (TEPRs) to detect CL. Formulating an explicit statement whether VR Technology provides more accurate measurements than classic approaches is difficult since we do not know the “true” effect of CL on the pupillary response (pupil diameter), which could serve as a classification for precise measurement. Besides the well-known physiological challenges, other variables (such as the emotional state) also have an impact on pupil size changes, which are also difficult to control. Nevertheless, this pioneer work provides a first assessment of the usefulness of VR Technology in comparison to classic approaches, which inspires future work adjustments. And it is worth to explore its utility further.

When discussing the use of VR Technology for research purposes, concerns about side effects for participants often rise. We tried to address the so-called *Cybersickness* (bodily discomfort and malaise) by giving participants small breaks to fill out questionnaires between task difficulties. Data shows that none of the included participants had any sickness feelings during the experiment. Hence, future work could include longer testing phases within the VR environment. This way, questionnaires, etc. could be filled out without taking off the VR headset, which in turn could prevent biases due to multiple calibration procedures.

An advantage of VR technology is that head movements can be better controlled. Several studies use a chin rest (e.g., Krejtz et al., 2018) to control for head movements, which can be uncomfortable for participants. We “centered” the *n-back* task within the VR environment; hence, the participant could move the head, and the content remained at the same position in the visual field. Thus, we could not totally control for light (letters had to be illuminated for reading purposes), but we could at least control for the angle of light incidence with higher comfort for participants.

As stated above, we do not know benchmark values for the pupil dilation that are solely caused by CL. But comparisons with other empirical findings can increase the probability of precise measurement. As mentioned above, the pupil diameter values and most of our measured variables seem to be in the same range as similar studies report. Hence, our measurements (except blinks) seem to be realistic.

Our data suggest no such strong sensitivity of the pupil diameter for task difficulty than other related studies. As stated above, this could be due to the included *n* steps or other factors (e.g., VR Technology) that either provoked more or less precise eye-related measurements. Since our descriptive data seem to be realistic, it is possible that the use of VR Technology revealed an overestimation of TEPRs, when inducing CL. This would be in line with few studies arguing that light and near reflexes provoke bigger changes in pupil size than cognitive processing (Kramer, 1990; Pomplum & Sunkara, 2003). As we intended to control better for light reflexes, reducing the brightness of the *n-back task* to a minimum and preventing external light incidence due to the VR headset, one can argue that this might be an indicator that light is much more responsible for TEPRs than generally presumed in the research field. Another argument is that calculated IPA scores were even less sensitive to task difficulty than the pupil diameter. Since the authors (Duchowski et al., 2018) of the IPA approach claim that it is less light-sensitive, we expected a higher sensitivity of the IPA score to detect CL. Before questioning the widely-accepted metric pupil diameter to detect CL, more experience should be gained, since VR hard- and software and IPA calculations are relatively novel at this faculty. Future work will provide more indications of whether VR Technology reflects added values for CL measurement or not.

In the following, some ideas are presented to gain a deeper understanding of whether VR Technology could be pursued further as a new approach. A control condition is not only helpful to interpret eye-related parameters when inducing CL but also helps to evaluate if the use of a VR environment provides better data. A control condition within the VR environment could be adapted by showing an empty screen (same background of the *n-*

back task) with the instruction to only focus on the center. Hence, the setting would be the same but without task-induced CL. These baseline values could indicate the CL that is only caused by using VR hardware. This is possible since most of the participants did not have experienced VR Technology before. Thus, a part of our measured pupil dilation could be due to process the VR environment. If these baseline values differ significantly from pupil dilation when CL is induced, it is more likely that pupil dilation is caused by task-induced CL (which provides a better understanding of whether this widely-accepted metric is valid or not).

	X		**Y**		**cd/m^2**	
VR Lens	**Left**	**Right**	**Left**	**Right**	**Left**	**Right**
VR turned off	0	0	0	0	0	0
Only Background (#383838)	0	0.295	0	0.371	2.84	4.18
C	0.231	0.181	0.325	0.359	13.7	11.3
D	0.274	0.317	0.292	0.528	12.6	31.2
F	0.137	0.139	0.483	0.578	17.1	25.2
H	0.269	0.143	0.586	0.628	21.8	36.8
K	0.424	0.143	0.242	0.481	13.9	16.1
N	0.428	0.184	0.414	0.198	17.2	7.14
P	0.439	0.125	0.350	0.586	10.4	23.1
R	0.414	0.148	0.432	0.576	19.0	37.7
V	0.233	0.127	0.434	0.642	17.1	29.8
Z	0.276	0.188	0.46	0.541	18.4	25.9
M(SD)	0.313(0.10)	0.170(0.05)	0.402(0.10)	0.512(0.13)	16.12(3.24)	24.42(9.74)
	0.241(0.08)		0.45675(0.11)		20.272(6.49)	

Table 9. Measured light density (cd/m^2) and chromaticity (X and Y coordinates) through both VR headset lenses:

A remaining obstacle to overcome is controlling for the pupillary light reflex. Several researchers have tried to eliminate the influence of light, but until today there is no *real-time* and easy approach known. The simple but yet complex idea is to develop an automatic application that subtracts the pupil dilation that is caused by light from the measured pupil size to display the “pure” pupil dilation that is due to cognitive processing.

Even though we prevent external light incidence, the VR environment needs a certain brightness, so test stimuli get visible. To have an idea of how much light density (cd/m²) is released during our test phase, we used the *Tekronix® J18 LumaColor™ II* that records light density (cd/m²) and chromaticity (CIE 1931 color spaces). *Table 9* shows measured values within the VR environment (via left and right VR lens): When the VR headset was turned off, only the background and test letter were displayed. With these values, we provide first steps to reduce light biases in the VR environment. When the VR headset was turned off, there was no light (density). Interestingly, when only the background was shown (#383838), no Y and X values were recorded for the left lens, but for the right one. This might be a sign of systematically higher brightness measurements of the right VR lens. Additionally, the cd/m² calculations also show higher values for the right VR lens on average than the left one. This observation should be deeper revised before collecting further eye-related data with this VR hardware.

According to the photometer's manual, the measured X and Y coordinates indicate a white color (Y = X = .333) of the test letters. Since we want to gain deeper insight into how to differentiate between pupil size changes due to luminance and cognitive processing, collected data about light density (cd/m²) is more of interest than chromaticity. Noticeable of our measured light density (cd/m²) is that we collected relatively low values among all test stimuli (*M*=20.424, *SD*=6.49). According to Becker and Herrmann (2003), an LCD monitor has a light density of 150 - 250 cd/m². Since TEPRs are induced using PC monitors classically (e.g., Chen & Epps, 2014), we can assume that using VR Technology can offer new ways to reduce (and control) brightness better.

Here, the following study design would be of interest: Participants' pupil diameter is measured when focusing on a display that gives off different light densities (cd/m²). This way, it would be possible to gain further data to which extent the pupil reacts only on several light densities. Another idea to control better for light is to replicate our study design and add a new condition (could be combined with a general control condition): Presenting the *n-back* letters but with the instruction to solely focus on the test stimuli. This way we would have pupil size changes only due to the brightness that, in turn, could be subtracted from the pupil size collected when the same letter stimulus has to be cognitively processed. This would provide crucial insights on how much light influences pupil dilation fluctuations. But it has to be mentioned that this calculation would imply time-consuming data pre-processing.

Another suggestion to control better for light is to inhibit any light incidence during the experiment using VR technology. This could be made when CL is induced without visual stimuli. Monk, Jackson, Nielsen, Jefferies, and Oliver (2011) used auditory synthesized digits as *n-back* stimuli in combination with speech recognition software to record the participants' responses. This way, the authors wanted to establish a more practical *Secondary Task* when conducting the *dual-task paradigm* (see more details in Chapter *Performance Measures*.) to induce working memory load (Monk et al., 2011). Another study used German consonants spoken by a female voice as *n-back* stimuli (Jaeggi, Buschkuehl, Perrig & Meier, 2010). The authors compared different variants of the *n-back* task and their results suggest that the auditory version induces more CL than visuospatial *n-back* stimuli (Jaeggi et al., 2010). This is of great interest for us, since our visual stimuli only differentiated partly between task difficulties. Using an auditory *n-back* task within a VR environment could prevent luminance biases and be more suitable to induce CL than visual letters. An easier approach to induce CL without visual stimuli (and hence brightness) would be to apply a study design conducted by Pecchinenda and Petrucci (2016): They let participants either count backward by seven (high CL) or counting forwards by two (low CL).

Thus, there have been few but successful alternative methods being implemented without the use of visual stimuli for different reasons. Here, a combined approach of an auditory *n-back* or arithmetic task with VR Technology could reduce brightness to a minimum and thus, provides data on how the pupil reacts to cognitive processing without luminance biases. Nevertheless, there are new potential obstacles implied when planning this new approach. Since participants would face an obscure VR environment for a longer time, a pilot study should examine in advance whether symptoms of agoraphobia (perceiving an unsafe environment with no easy way to escape), dizziness or other troubles occur. Furthermore, closing eyes or eye movements could occur, since participants do not have to focus on visual stimuli. Here, this potential issue has to be addressed in instructions and monitored by the study conductor who corrects the participant if necessary.

All in all, the use of VR Technology to gain more precise CL data is a promising approach, since it may explain mixed findings whether the TEPR is a valid method to measure CL or not. As mentioned above, we did not find such a strong relation between CL and pupil dilation as reported by other empirical work, which can be due to less or more valid measurements. Our work provides first insights on the suitability of VR Technology. This study aims to motivate to further examine the VR approach and provides recommendations regarding future experiments that could reveal comprehensive evidence for or against its use.

Conclusion

This interdisciplinary work makes a contribution to improve CL measurement by using VR and Eye-Tracking Technology. Additionally, psychophysiological constructs (stress and emotions) were assessed as potential confounding variables. Hence, this work addresses methodical obstacles in CL research.

The main focus of this work lied on CL measurement using the *Task-Evoked Pupillary Response* (TEPR) since it reflects a widely-accepted CL parameter. We used a VR environment to better control for the well-known luminance biases when measuring the pupil dilation. Our data suggest a rather small effect of pupil size change among task difficulty compared to related empirical work. Correspondingly, the novel IPA calculation revealed no sensitivity of the pupil diameter to task difficulty. Contrarily, self-reported CL and performance measures (error rate and reaction time) indicate a successful CL manipulation. These findings may challenge the widely-used pupil dilation as a parameter to detect cognitive processing in research. Putting together all results, it seems more plausible that the empirical lack of sensitivity to CL is due to a more precise CL measurement, since we reduced brightness to a minimum (via VR Technology). Additionally, the light-insensitive IPA score did not detect CL at all. Hence, TEPRs might be overestimated and biased in CL research. Since this finding would entail a better understanding and measurement of CL on a big scale it is worth to integrate VR Technology in future work. Chapter *VR Technology – A New Promising Approach*? illustrates several approaches to gain further insights in detail. More precisely, a control group would support to interpret our empirical findings. Further, our data indicate that the low CL (*0-back*) and medium CL (*1-back*) condition induced a similar CL level. Future work can include either of them to induce low CL.

Besides addressing the luminance bias via VR Technology, we addressed the overlap issue with several other psychophysiological constructs that also represent confounding variables when measuring CL. As mentioned before, same dependent variables (e.g., GSR signal) are commonly used to measure either cognitive processing, stress or emotional states. Very few studies try to differentiate empirically between these constructs to examine to what extent all constructs impact the pupillary response. Within our experiment, CL caused small changes in the self-reported stress level and emotional state: With increasing task difficulty perceived stress and arousal increased, whereas valence and dominance decreased. It is possible that these rather small changes are due to the laboratory setting (e.g., no severe consequences in the case of bad performance). This is also supported by the observed pattern that CL has an impact on the perceived stress level and emotional state, particularly when high CL was induced. Since stress and emotional reactions influence typical CL parameters (e.g., pupil size or SCL), future work should try to understand better the relationship between these constructs. This could lead to new approaches that control better for these confounding variables.

In conclusion, this work contributed first insights into VR Technology as a new approach to measure CL more precisely. Results suggest that this promising approach is worth to be broader investigated. Additionally, findings demonstrated the often neglected "closeness" between CL and other psychophysiological constructs, which should be considered more when conducting studies and interpreting empirical findings in CL research. Since a precise CL measurement is beneficial to several high-risk areas it should be further pursued, even though methodical obstacles have to be overcome. Moreover, CL measurement is a good example of interdisciplinary research that pools different perspectives on the topic and therefore, provides holistic findings that should be pursued more in the academic context.

References

Albert, W., & Tullis, T. (2013). *Measuring the User Experience: Collecting, analyzing, and presenting usability metrics.* Morgan Kaufmann.

Andersen, S. A. W., Mikkelsen, P. T., Konge, L., Cayé-Thomasen, P., & Sørensen, M. S. (2016). Cognitive load in mastoidectomy skills training: virtual reality simulation and traditional dissection compared. *Journal of surgical education*, *73*(1), 45-50.

Antonenko, P., Paas, F., Grabner, R., & Van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review*, *22*(4), 425-438.

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). Academic Press.

Ayres, P. (2015). State- of- the- art research into multimedia learning: A commentary on Mayer's Handbook of Multimedia Learning. *Applied Cognitive Psychology*, *29*(4), 631-636.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In *Psychology of learning and motivation* (Vol. 8, pp. 47-89). Academic press.

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). Memory. East Sussex.

Baddeley, A., Lewis, V., & Vallar, G. (1984). Exploring the articulatory loop. *The Quarterly Journal of Experimental Psychology Section A*, *36*(2), 233-252.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of experimental psychology: General*, *104*(1), 54.

Barrett, L. F., & Russell, J. A. (1999). The structure of current affect: Controversies and emerging consensus. *Current directions in psychological science*, *8*(1), 10-14.

Bayer, M., Ruthmann, K., & Schacht, A. (2017). The impact of personal relevance on emotion processing: evidence from event-related potentials and pupillary responses. *Social cognitive and affective neuroscience*, *12*(9), 1470-1479.

Beatty, J. (1988). Pupillometric signs of selective attention in man. *Neurophysiology and psychophysiology: Experimental and clinical applications*, 138-143.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. *Handbook of psychophysiology*, *2*(142-162).

Becker, M.E., & Herrmann, H.J. (2003). LCD-Bildschirme – ergonomisch. *Computer-Fachwissen*. 10/2003. Retrieved November 18, 2019, from https://www.display-messtechnik.de/fileadmin/template/main/docs/lcd-bildschirm-ergonomie-compfach.pdf

Bohringer, A., Schwabe, L., Richter, S., & Schachinger, H. (2008). Intranasal insulin attenuates the hypothalamic–pituitary–adrenal axis response to psychosocial stress. *Psychoneuroendocrinology*, *33*(10), 1394-1400.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49-59.

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: defensive and appetitive reactions in picture processing. *Emotion*, *1*(3), 276.

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional Arousal and autonomic activation. *Psychophysiology*, *45*(4), 602-607.

Brosch, T., Pourtois, G., & Sander, D. (2010). The perception and categorisation of emotional stimuli: A review. *Cognition and emotion*, *24*(3), 377-400.

Brown, T. M., & Fee, E. (2002). Walter Bradford Cannon: Pioneer physiologist of human emotions. *American Journal of Public Health*, *92*(10), 1594-1595.

Butz, A., & Krüger, A. (2017). *Mensch-Maschine-Interaktion*. Walter de Gruyter GmbH & Co KG.

Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., & Ito, T. A. (2000). The psychophysiology of emotion. *Handbook of emotions*, *2*, 173-191.

Cannon, W. B. (1914). The emergency function of the adrenal medulla in pain and the major emotions. *American Journal of Physiology-Legacy Content*, *33*(2), 356-372.

Cardoş, R. A., David, O. A., & David, D. O. (2017). Virtual reality exposure therapy in flight anxiety: a quantitative meta-analysis. *Computers in Human Behavior*, *72*, 371-380.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, *4*(1), 55-81.

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., & Conway, D. (2016). *Robust multimodal cognitive load measurement*. Cham: Springer International Publishing.

Chen, S., & Epps, J. (2013). Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer methods and programs in biomedicine*, *110*(2), 111-124.

Chen, S., & Epps, J. (2014). Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction*, *29*(4), 390-413.

Chen, S., Epps, J., & Chen, F. (2013). Automatic and continuous user task analysis via eye activity. In *Proceedings of the 2013 international conference on Intelligent user interfaces* (pp. 57-66). ACM.

Chong, R. K., Mills, B., Dailey, L., Lane, E., Smith, S., & Lee, K. H. (2010). Specific interference between a cognitive task and sensory organization for stance balance control in healthy young adults: visuospatial effects. *Neuropsychologia*, *48*(9), 2709-2718.

Chrousos, G. P. (2009). Stress and disorders of the stress system. *Nature reviews endocrinology*, *5*(7), 374.

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load?. *Computers in Human Behavior*, *25*(2), 315-324.

Cochran, B. (2017). The Impact of Working Memory on Response Order Effects and Question Order Effects in Telephone and Web Surveys.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Abingdon.

Conrad, R. (1964). Acoustic confusions in immediate memory. *British journal of Psychology*, *55*(1), 75-84.

Conway, D., Dick, I., Li, Z., Wang, Y., & Chen, F. (2013). The effect of stress on cognitive load measurement. In *IFIP Conference on Human-Computer Interaction* (pp. 659-666). Springer, Berlin, Heidelberg.

Cranford, K. N., Tiettmeyer, J. M., Chuprinko, B. C., Jordan, S., & Grove, N. P. (2014). Measuring load on working memory: the use of heart rate as a means of measuring chemistry students' cognitive load. *Journal of Chemical Education*, *91*(5), 641-647.

Dalgleish, T., Dunn, B. D., & Mobbs, D. (2009). Affective neuroscience: Past, present, and future. *Emotion Review*, *1*(4), 355-368.

Dan, A., & Reiner, M. (2017). EEG-based cognitive load of processing events in 3D virtual worlds is lower than processing events in 2D displays. *International Journal of Psychophysiology*, *122*, 75-84.

Daniel, T. A., & Katz, J. S. (2018). Primacy and recency effects for taste. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(3), 399.

Davis, J. I., Senghas, A., Brandt, F., & Ochsner, K. N. (2010). The effects of BOTOX injections on emotional experience. *Emotion*, *10*(3), 433.

De Groot, A. D. (1965). Thought and Choice in Chess. 1978. *The Hague, Netherlands: Mouton Publishers*.

De la Torre, G., Ramallo, M. A., & Cervantes, E. (2016). Workload perception in drone flight training simulators. *Computers in Human Behavior*, *64*, 449-454.

De Renzi, E., Liotti, M., & Nichelli, P. (1987). Semantic amnesia with preservation of autobiographic memory. A case report. *Cortex*, *23*(4), 575-597.

Debue, N., & Van De Leemput, C. (2014). What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in psychology*, *5*, 1099.

DeFraine, W. C. (2016). Differential effects of cognitive load on emotion: Emotion maintenance versus passive experience. *Emotion, 16*(4), 459-467.

Delaney, P. F., & Sahakyan, L. (2007). Unexpected costs of high working memory capacity following directed forgetting and contextual change manipulations. *Memory & cognition*, *35*(5), 1074-1082.

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of educational psychology*, *100*(1), 223.

Dickerson, S. S., & Kemeny, M. E. (2004). Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, *130*(3), 355.

Diemer, J., & Zwanzger, P. (2019). Die Entwicklung virtueller Realität als Expositionsverfahren. *Der Nervenarzt*, *90*(7), 715-723.

Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., ... & Giannopoulos, I. (2018). The index of pupillary activity: measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 282). ACM.

Edwards, S., Clow, A., Evans, P., & Hucklebridge, F. (2001). Exploration of the awakening cortisol response in relation to diurnal cortisol secretory activity. *Life sciences*, *68*(18), 2093-2103.

Esler, M., Jackman, G., Bobik, A., Kelleher, D., Jennings, G., Leonard, P., ... & Korner, P. (1979). Determination of norepinephrine apparent release rate and clearance in humans. *Life Sciences*, *25*(17), 1461-1470.

Felton, E. A., Williams, J. C., Vanderheiden, G. C., & Radwin, R. G. (2012). Mental workload during brain–computer interface training. *Ergonomics*, *55*(5), 526-537.

Ferreira, C. T., Ceccaldi, M., Giusiano, B., & Poncet, M. (1998). Separate visual pathways for perception of actions and objects: evidence from a case of apperceptive agnosia. *Journal of Neurology, Neurosurgery & Psychiatry*, *65*(3), 382-385.

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.

Fisk, A. D., Derrick, W. L., & Schneider, W. (1986). A methodological assessment and evaluation of *dual-task paradigm*s. *Current Psychological Research & Reviews*, *5*(4), 315-327.

Frank, M. G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high-stake lies. *Journal of personality and social psychology*, *72*(6), 1429.

Fraser, K., Huffman, J., Ma, I., Sobczak, M., McIlwrick, J., Wright, B., & McLaughlin, K. (2014). The emotional and cognitive impact of unexpected simulated patient death: a randomized controlled trial. *Chest*, *145*(5), 958-963.

Frosina, P., Logue, M., Book, A., Huizinga, T., Amos, S., & Stark, S. (2018). The effect of cognitive load on nonverbal behavior in the cognitive interview for suspects. *Personality and Individual Differences*, *130*, 51-58.

Fuentes-García, J. P., Pereira, T., Castro, M. A., Santos, A. C., & Villafaina, S. (2019). Psychophysiological stress response of adolescent chess players during problem-solving tasks. *Physiology & behavior*, *209*, 112609.

Gaab, J. (2009). PASA-Primary Appraisal Secondary Appraisal-Ein Fragebogen zur Erfassung von situationsbezogenen kognitiven Bewertungen. *Verhaltenstherapie*, *19*(2), 114-115.

Gaab, J., Blättler, N., Menzi, T., Pabst, B., Stoyer, S., & Ehlert, U. (2003). Randomized controlled evaluation of the effects of cognitive–behavioral stress management on cortisol responses to acute stress in healthy subjects.*Psychoneuroendocrinology*, *28*(6), 767-779.

Gaab, J., Rohleder, N., Nater, U. M., & Ehlert, U. (2005). Psychological determinants of the cortisol stress response: the role of anticipatory cognitive appraisal. *Psychoneuroendocrinology*, *30*(6), 599-610.

Gareau, A., & Gaudreau, P. (2017). Working memory moderates the effect of the integrative process of implicit and explicit autonomous motivation on academic achievement. *British Journal of Psychology*, *108*(4), 701-720.

Gerjets, P., Scheiter, K., Opfermann, M., Hesse, F. W., & Eysink, T. H. (2009). Learning with hypermedia: The influence of representational formats and different levels of learner control on performance and learning behavior.*Computers in Human Behavior*, *25*(2), 360-370.

Gevins, A., & Smith, M. E. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science*, *4*(1-2), 113-131.

Gilbert, A. M., & Fiez, J. A. (2004). Integrating rewards and cognition in the frontal cortex. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(4), 540-552.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of verbal learning and verbal behavior*, *5*(4), 351-360.

Goldstein, E. B. (2018). *Cognitive psychology: Connecting mind, research, and everyday experience*. Australia: Wadsworth Cengage Learning.

Golenhofen, K. (1997). *Physiologie: Lehrbuch, Kompendium, Fragen und Antworten; mit 7 Tabellen*. Urban & Schwarzenberg.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, *15*(1), 20-25.

Grigg, S. J., Garrett, S. K., & Benson, L. C. (2012). Using the NASA-TLX to assess first year engineering problem difficulty. In *IIE Annual Conference. Proceedings* (p. 1). Institute of Industrial and Systems Engineers (IISE).

Grimes, D., Tan, D. S., Hudson, S. E., Shenoy, P., & Rao, R. P. (2008). Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 835-844). ACM.

Guastello, S. J., Reiter, K., Malon, M., Timm, P., Shircel, A., & Shaline, J. (2015). Catastrophe models for cognitive workload and fatigue in *N-back* tasks. *Nonlinear dynamics, psychology, and life sciences*.

Hammerfald, K., Eberle, C., Grau, M., Kinsperger, A., Zimmermann, A., Ehlert, U., & Gaab, J. (2006). Persistent effects of cognitive-behavioral stress management on cortisol responses to acute stress in healthy subjects—a randomized controlled trial. *Psychoneuroendocrinology*, *31*(3), 333-339.

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908.

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). North-Holland.

Huang, L. Y., She, H. C., Chou, W. C., Chuang, M. H., Duann, J. R., & Jung, T. P. (2013). Brain oscillation and connectivity during a chemistry visual working memory task. *International Journal of Psychophysiology*, *90*(2), 172-179.

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). IEEE.

Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2004). Changes in mental workload during task execution. In *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology*.

Irwin, D. E., & Thomas, L. E. (2010). Eyeblinks and cognition. In V. Coltheart (Ed.), *Macquarie monographs in cognitive science. Tutorials in visual cognition* (pp. 121-141). New York, NY, US: Psychology Press.

Isen, A. M., & Reeve, J. (2005). The influence of positive affect on intrinsic and extrinsic motivation: Facilitating enjoyment of play, responsible work behavior, and self-control. *Motivation and emotion*, *29*(4), 295-323.

Izard, C. E. (1994). Innate and universal facial expressions: evidence from developmental and cross-cultural research. *Psychological Bulletin*, 115, 288–299.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the *N-back* task as a working memory measure. *Memory*, *18*(4), 394-412.

Jergović, M., Tomičević, M., Vidović, A., Bendelja, K., Savić, A., Vojvoda, V., ... & Sabioncello, A. (2014). Telomere shortening and immune activity in war veterans with posttraumatic stress disorder. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *54*, 275-283.

Just, M. A., Carpenter, P. A., & Miyake, A. (2003). Neuroindices of cognitive workload: Neuroimaging, pupillometric and event-related potential studies of brain work. *Theoretical Issues in Ergonomics Science*, *4*(1-2), 56-88.

Juster, R. P., Perna, A., Marin, M. F., Sindi, S., & Lupien, S. J. (2012). Timing is everything: Anticipatory stress dynamics among cortisol and blood pressure reactivity and recovery in healthy adults. *Stress*, *15*(6), 569-577.

Kadziolka, M. J., Di Pierdomenico, E. A., & Miller, C. J. (2016). Trait-like mindfulness promotes healthy self-regulation of stress. *Mindfulness*, *7*(1), 236-245.

Kaluza, G. (2012). Gelassen und sicher im Stress (4. Aufl.). *Berlin Heidelberg: Springer-Verlag.*

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need?. *Educational Psychology Review*, *23*(1), 1-19.

Karatekin, C., Couperus, J. W., & Marcus, D. J. (2004). Attention allocation in the dual- task paradigm as measured through behavioral and psychophysiological responses. *Psychophysiology*, *41*(2), 175-185.

Kennedy, D. O., & Scholey, A. B. (2000). Glucose administration, heart rate and cognitive performance: effects of increasing mental effort. *Psychopharmacology*, *149*(1), 63-71.

Khawaja, M. A. (2010). Cognitive load measurement using speech and linguistic features.

King, A. C., Schluger, J., Gunduz, M., Borg, L., Perret, G., Ho, A., & Kreek, M. J. (2002). Hypothalamic-pituitary-adrenocortical (HPA) axis response and biotransformation of oral naltrexone: preliminary examination of relationship to family history of alcoholism. *Neuropsychopharmacology*, *26*(6), 778-788.

Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, *28*(1-2), 76-81.

Klauer, K. C., & Zhao, Z. (2004). Double dissociations in visual and spatial short-term memory. *Journal of Experimental Psychology: General*, *133*(3), 355.

Klimesch, W., Schack, B., & Sauseng, P. (2005). The functional significance of theta and upper alpha oscillations. *Experimental psychology*, *52*(2), 99-108.

Knickerbocker, F., Johnson, R. L., Starr, E. L., Hall, A. M., Preti, D. M., Slate, S. R., & Altarriba, J. (2019). The time course of processing emotion-laden words during sentence reading: Evidence from eye movements. *Acta psychologica*, *192*, 1-10.

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. *Multiple-task performance*, 279-328.

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.

Kudielka, B. M., Hellhammer, D. H., Kirschbaum, C., Harmon-Jones, E., & Winkielman, P. (2007). Ten years of research with the Trier Social Stress Test—revisited. *Social neuroscience: Integrating biological and psychological explanations of social behavior*, *56*, 83.

Kuebler, U., Wirtz, P. H., Sakai, M., Stemmer, A., & Ehlert, U. (2013). Acute stress reduces wound-induced activation of microbicidal potential of ex vivo isolated human monocyte-derived macrophages. *PLoS One*, *8*(2), e55875.

Kuebler, U., Zuccarella-Hackl, C., Arpagaus, A., Wolf, J. M., Farahmand, F., von Känel, R., ... & Wirtz, P. H. (2015). Stress-induced modulation of NF-κB activation, inflammation-associated gene expression, and cytokine levels in blood of healthy men. *Brain, behavior, and immunity*, *46*, 87-95.

Lambie, J. A., & Marcel, A. J. (2002). Consciousness and the varieties of emotion experience: A theoretical framework. *Psychological review*, *109*(2), 219.

Landowska, A., Roberts, D., Eachus, P., & Barrett, A. (2018). Within-and between-session prefrontal cortex response to Virtual Reality Exposure Therapy for acrophobia. *Frontiers in human neuroscience*, *12*.

Lazarus, R. S., & Folkman, S. (1984). Stress, appraisal, and coping. *Springer publishing company*.

Ledger, H. (2013). The effect cognitive load has on eye blinking. *The Plymouth Student Scientist*, *6*(1), 206-223.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual review of neuroscience*, *23*(1), 155-184.

Levy, M. Z., Allsopp, R. C., Futcher, A. B., Greider, C. W., & Harley, C. B. (1992). Telomere end-replication problem and cell aging. *Journal of molecular biology*, *225*(4), 951-960.

Li, J., Zhang, M., Loerbroks, A., Angerer, P., & Siegrist, J. (2015). Work stress and the risk of recurrent coronary heart disease events: A systematic review and meta-analysis. *International journal of occupational medicine and environmental health*, 1-12.

Li, P., Markkula, G., Li, Y., & Merat, N. (2018). Is improved lane keeping during cognitive load caused by increased physical Arousal or gaze concentration toward the road center?. *Accident Analysis & Prevention*, *117*, 65-74.

Li, X., Ouyang, Z., & Luo, Y. J. (2012). The cognitive load affects the interaction pattern of emotion and working memory. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, *6*(2), 68-81.

Lin, T., & Imamiya, A. (2006). Evaluating usability based on multimodal information: an empirical study. In *Proceedings of the 8th international conference on Multimodal interfaces* (pp. 364-371). ACM.

Lin, T., Li, X., Wu, Z., & Tang, N. (2013). Automatic cognitive load classification using high-frequency interaction events: An exploratory study. *International Journal of Technology and Human Interaction (IJTHI)*, *9*(3), 73-88.

Lin, T., Omata, M., Hu, W., & Imamiya, A. (2005). Do physiological data relate to traditional usability indexes?. In *Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future* (pp. 1-10). Computer-Human Interaction Special Interest Group (CHISIG) of Australia.

Lini, S., Favier, P. A., Hourlier, S., Vallespir, B., Bey, C., & Baracat, B. (2012, September). Influence of a temporally-customizable HMI on pilots' cognitive load in civil aviation: a comparative study. In *Proceedings of the HCI Aero conference*.

Lobato-Rincón, L. L., Cabanillas-Campos, M. D. C., Bonnin-Arias, C., Chamorro-Gutiérrez, E., Murciano-Cespedosa, A., & Sánchez-Ramos Roda, C. (2014). Pupillary behavior in relation to wavelength and age. *Frontiers in human neuroscience*, *8*, 221.

Luque-Casado, A., Perales, J. C., Cárdenas, D., & Sanabria, D. (2016). Heart rate variability and cognitive processing: The autonomic response to task demands. *Biological psychology*, *113*, 83-90.

Madathil, K. C., & Greenstein, J. S. (2017). An investigation of the efficacy of collaborative virtual reality systems for moderated remote usability testing. *Applied ergonomics*, *65*, 501-514.

Marshall, S. P. (2002). The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the IEEE 7th conference on Human Factors and Power Plants* (pp. 7-7). IEEE.

Martin, S. (2014). Measuring cognitive load and cognition: metrics for technology-enhanced learning. *Educational Research and Evaluation*, *20*(7-8), 592-621.

Mathur, A., Gehrmann, J., & Atchison, D. A. (2013). Pupil shape as viewed along the horizontal visual field. *Journal of vision*, *13*(6), 3-3.

Mayer, R. E. (2005). Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning*, *3148*.

Mayer, R. E., & Pilegard, C. (2005). Principles for managing essential processing in multimedia learning: Segmenting, pretraining, and modality principles. *The Cambridge handbook of multimedia learning*, 169-182.

Mayer, R., & Pilegard, C. (2014). Principles for managing essential processing in multimedia learning: Segmenting, pre-training, and modality principles. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd edn, pp. 316–344). New York, N.Y.: Cambridge University Press.

McEwen, B. S. (1998). Stress, adaptation, and disease: Allostasis and allostatic load. *Annals of the New York academy of sciences*, *840*(1), 33-44.

McEwen, B. S., & Sapolsky, R. M. (1995). Stress and cognitive function. *Current opinion in neurobiology*, *5*(2), 205-216.

Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. the MIT Press.

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.

Miller, L. H., Smith, A. D., & Rothstein, L. (1994). *The stress solution: An action plan to manage the stress in your life*. Pocket.

Monk, A. F., Jackson, D., Nielsen, D., Jefferies, E., & Olivier, P. (2011). *N-back*er: An auditory *n-back* task with automatic scoring of spoken responses. *Behavior research methods*, *43*(3), 888.

Morath, J., Moreno-Villanueva, M., Hamuni, G., Kolassa, S., Ruf-Leuschner, M., Schauer, M., ... & Kolassa, I. T. (2014). Effects of psychotherapy on DNA strand break accumulation originating from traumatic stress. *Psychotherapy and Psychosomatics*, *83*(5), 289-297.

Moreno, R. (2010). Cognitive load theory: More food for thought. *Instructional Science*, *38*(2), 135-141.

Moreno, R. E., & Park, B. (2010). Cognitive load theory: Historical development and relation to other theories.

Morrongiello, B. A., Corbett, M., Beer, J., & Koutsoulianos, S. (2018). A pilot randomized controlled trial testing the effectiveness of a pedestrian training program that teaches children where and how to cross the street safely. *Journal of Pediatric Psychology*, *43*(10), 1147-1159.

Mulder, L. J. M. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological psychology*, *34*(2-3), 205-236.

Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*(5), 482-488.

Murison, R. (2016). The neurobiology of stress. *Neuroscience of Pain, Stress, and Emotion* (pp. 29-49). Academic Press.

Murray, D. J. (1968). Articulation and acoustic confusability in short-term memory. *Journal of Experimental Psychology*, *78*(4p1), 679.

Nakajima, Y., & Sato, K. (1989). Distractor difficulty and the long-term recency effect. *The American journal of psychology*, 511-521.

Nater, U. M., Whistler, T., Lonergan, W., Mletzko, T., Vernon, S. D., & Heim, C. (2009). Impact of acute psychosocial stress on peripheral blood gene expression pathways in healthy men. *Biological psychology*, *82*(2), 125-132.

Nickel, P., & Nachreiner, F. (2000). Psychometric properties of the 0.1 Hz component of HRV as an indicator of mental strain. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 44, No. 12, pp. 2-747). Sage CA: Los Angeles, CA: SAGE Publications.

Nickel, P., & Nachreiner, F. (2003). Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload. Human factors, 45(4), 575-590.
Norman, D. A. (2004). *Emotional design: Why we love (or hate) everyday things*. Basic Civitas Books.

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. (2012, November). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference* (pp. 420-423). ACM.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). Multimedia learning with intelligent tutoring systems. *Cambridge handbook of multimedia learning*, 705-728.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. *Handbook of Perception and Human Performance*. Volume 2. Cognitive Processes and Performance. ed. KR Boff, L. Kaufman and JP Thomas, pp. 42–1–42–49.

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. (1994a). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational psychology review*, *6*(4), 351-371.

Paas, F. G., & Van Merriënboer, J. J. (1994b). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of educational psychology*, *86*(1), 122.

Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. *The Cambridge handbook of multimedia learning*, *27*, 27-42.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, *38*(1), 63-71.

Paivio, A. (1990). *Mental representations: A dual coding approach* (Vol. 9). Oxford University Press.

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 141-144). ACM.

Panasiti, M. S., Ponsi, G., Monachesi, B., Lorenzini, L., Panasiti, V., & Aglioti, S. M. (2019). Cognitive load and emotional processing in psoriasis: a thermal imaging study. *Experimental brain research*, *237*(1), 211-222.

Park, B., & Brünken, R. (2015). The Rhythm Method: A New Method for Measuring Cognitive Load—An Experimental Dual- Task Study. *Applied Cognitive Psychology*, *29*(2), 232-243.

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International journal of human-computer studies*, *59*(1-2), 185-198.

Pecchinenda, A., & Petrucci, M. (2016). Emotion unchained: Facial expression modulates gaze cueing under cognitive load. *PloS one*, *11*(12), e0168111.

Pedrotti, M., Mirzaei, M. A., Tedesco, A., Chardonnet, J. R., Mérienne, F., Benedetto, S., & Baccino, T. (2014). Automatic stress classification with pupil diameter analysis. *International Journal of Human-Computer Interaction*, *30*(3), 220-236.

Perala, C. H., & Sterling, B. S. (2007). *Galvanic skin response as a measure of soldier stress* (No. ARL-TR-4114). Army Research Lab Aberdeen Proving Ground Md Human Research and Engineering Directorate.

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of experimental psychology*, *58*(3), 193.

Pintrich, P. R. (2003). Motivation and classroom learning. *Handbook of psychology*, 103-122.
Plutchik, R. (1980). Emotion. *A psychoevolutionary synthesis*.

Pomplun, M., & Sunkara, S. (2003). Pupil dilation as an indicator of cognitive workload in human-computer interaction. In *Proceedings of the International Conference on HCI* (Vol. 2003).

Portello, J. K., Rosenfield, M., & Chu, C. A. (2013). Blink rate, incomplete blinks and computer vision syndrome. *Optometry and Vision Science*, *90*(5), 482-487.

Postman, L., & Phillips, L. W. (1965). Short-term temporal changes in free recall. *The Quarterly Journal of Experimental Psychology, 17*(2), 132-138.

Quatieri, T. F., Williamson, J. R., Smalt, C. J., Perricone, J., Patel, T., Brattain, L., ... & Eddy, M. (2017). Multimodal biomarkers to discriminate cognitive state. *The Role of Technology in Clinical Neuropsychology*, *409*.

Raaijmakers, S. F., Baars, M., Schaap, L., Paas, F., & Van Gog, T. (2017). Effects of performance feedback Valence on perceptions of invested mental effort. *Learning and Instruction*, *51*, 36-46.

Rasmussen, S. R., Konge, L., Mikkelsen, P. T., Sørensen, M. S., & Andersen, S. A. (2016). Notes from the field: *Secondary Task* precision for cognitive load estimation during virtual reality surgical simulation training. *Evaluation & the health professions*, *39*(1), 114-120.

Richer, F., & Beatty, J. (1985). Pupillary dilations in movement preparation and execution. *Psychophysiology*, *22*(2), 204-207.

Roberts, W. (2017). The Use of Cues in Multimedia Instructions in Technology as a way to Reduce Cognitive Load. *Journal of Educational Multimedia and Hypermedia*, *26*(4), 373-412.

Rohleder, N., Nater, U. M., Wolf, J. M., Ehlert, U., & Kirschbaum, C. (2004). Psychosocial stress-induced activation of salivary alpha-amylase. *Annals of the New York Academy of Sciences*, *1032*, 258-263.

Rosenbaum, R. S., Köhler, S., Schacter, D. L., Moscovitch, M., Westmacott, R., Black, S. E., ... & Tulving, E. (2005). The case of KC: contributions of a memory-impaired person to memory theory. *Neuropsychologia*, *43*(7), 989-1021.

Rothermund, K., & Eder, A. B. (2011). *Allgemeine psychologie: Motivation und emotion* (1. Aufl. ed.). Wiesbaden: VS-Verl. doi:10.1007/978-3-531-93420-4

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of experimental psychology*, 89(1), 63.

Sanada, M., Ikeda, K., Kimura, K., & Hasegawa, T. (2013). Motivation enhances visual working memory capacity through the modulation of central cognitive processes. *Psychophysiology*, *50*(9), 864-871.

Sato, H., Takenaka, I., & Kawahara, J. I. (2012). The effects of acute stress and perceptual load on distractor interference. *Quarterly journal of experimental psychology*, *65*(4), 617-623.

Scharinger, C., Soutschek, A., Schubert, T., & Gerjets, P. (2015). When flanker meets the n-back: What EEG and pupil dilation data reveal about the interplay between the two central-executive working memory functions inhibition and updating. *Psychophysiology*, *52*(10), 1293-1304.

Schminder, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution. *European Research Journal of Methods for the Behavioral and Social Sciences*, *6*, 147-151.

Schuster, R. M., Hammitt, W. E., & Moore, D. (2003). A theoretical model to measure the appraisal and coping response to hassles in outdoor recreation settings. *Leisure Sciences*, *25*(2-3), 277-299.

Schwonke, R., Renkl, A., Salden, R., & Aleven, V. (2011). Effects of different ratios of worked solution steps and problem solving opportunities on cognitive load and learning outcomes. *Computers in Human Behavior*, *27*(1), 58-62.

Scott, G. G., O'Donnell, P. J., & Sereno, S. C. (2012). Emotion words affect eye fixations during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 783.

Selye, H. (1950). The physiology and pathology of exposure to stress.

Selye, H. (1970). The evolution of the stress concept: Stress and cardiovascular disease. *The American journal of cardiology*, *26*(3), 289-299.

Selye, H., & Fortier, C. (1950). Adaptive reactions to stress. Life stress and bodily disease. Edit. MG Wolff, SG Wolf, and CC Hare.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591-611.

Shivakumar, G., & Vijaya, P. A. (2015). Investigation of Individual Emotions with GSR and FTT by Employing LabVIEW. In *Handbook of Research on Synthesizing Human Emotion in Intelligent Systems and Robotics* (pp. 214-228). IGI Global.

Soussignan, R. (2004). Regulatory function of facial actions in emotion processes. *Advances in psychology research*, *31*, 173-198.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological monographs: General and applied*, *74*(11), 1-29.

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.

Stark, L., Campbell, F. W., & Atwood, J. (1958). Pupil unrest: an example of noise in a biological servomechanism. *Nature*, *182*(4639), 857.

Stark-Wroblewski, K., Kreiner, D. S., Boeding, C. M., Lopata, A. N., Ryan, J. J., & Church, T. M. (2008). Use of virtual reality technology to enhance undergraduate learning in abnormal psychology. *Teaching of Psychology*, *35*(4), 343-348.

Stern, J. A., Walrath, L. C., & Goldstein, R. (1984). The endogenous eyeblink. *Psychophysiology*, *21*(1), 22-33.

Storch, M., Gaab, J., Küttel, Y., Stüssi, A. C., & Fend, H. (2007). Psychoneuroendocrine effects of resource-activating stress management training. *Health Psychology*, *26*(4), 456.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, *4*(4), 295-312.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring cognitive load. In *Cognitive load theory* (pp. 71-85). Springer, New York, NY.

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

Szabo, S., Tache, Y., & Somogyi, A. (2012). The legacy of Hans Selye and the origins of stress research: a retrospective 75 years after his landmark brief "letter" to the editor of nature. *Stress*, *15*(5), 472-478.

Szatkowska, I., Bogorodzki, P., Wolak, T., Marchewka, A., & Szeszkowski, W. (2008). The effect of motivation on working memory: An fMRI and SEM study. *Neurobiology of learning and memory*, *90*(2), 475-478.

Tanaka, Y., & Yamaoka, K. (1993). Blink activity and task difficulty. *Perceptual and Motor Skills*, *77*(1), 55-66.

Teranishi, S., & Yamagishi, Y. (2018). Educational Effects of a Virtual Reality Simulation System for constructing Self-Built PCs. *Journal of Educational Multimedia and Hypermedia*, *27*(3), 411-423.

Tulving, E. (1985). How many memory systems are there?. *American psychologist*, *40*(4), 385.

Turner, J. R., & Carroll, D. (1985). Heart rate and oxygen consumption during mental arithmetic, a video game, and graded exercise: Further evidence of metabolically-exaggerated cardiac adjustments?. *Psychophysiology*, *22*(3), 261-267.

Ungerleider, L. G., Courtney, S. M., & Haxby, J. V. (1998). A neural system for human visual working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(3), 883–890. doi:10.1073/pnas.95.3.883

Unsworth, N. (2009). Variation in working memory capacity, fluid intelligence, and episodic recall: A latent variable examination of differences in the dynamics of free recall. *Memory & Cognition*, *37*(6), 837-849.

Unsworth, N., & Brewer, G. A. (2010). Variation in working memory capacity and intrusions: Differences in generation or editing?. *European Journal of Cognitive Psychology*, *22*(6), 990-1000.

Van Gerven, P. W., Paas, F., Van Merriënboer, J. J., & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology*, *41*(2), 167-174.

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.

Vickers, K. L., Schultheis, M. T., & Manning, K. J. (2018). Driving after brain injury: Does dual-task modality matter?. *NeuroRehabilitation*, *42*(2), 213-222.

Vitaliano, P. P., Scanlan, J. M., Zhang, J., Savage, M. V., Hirsch, I. B., & Siegler, I. C. (2002). A path model of chronic stress, the metabolic syndrome, and coronary heart disease. *Psychosomatic medicine*, *64*(3), 418-435.

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, *438*(7067), 500.

Vogels, J., Demberg, V., & Kray, J. (2018). The Index of Cognitive Activity as a measure of cognitive processing load in dual task settings. *Frontiers in Psychology*, *9*.

Von Bauer, P., (2018). Cognitive State Classification Using Psychophysiological Measures. University of Konstanz. (not published)

Wang, J., Zhou, T., Qiu, M., Du, A., Cai, K., Wang, Z., ... & Chen, L. (1999). Relationship between ventral stream for object vision and dorsal stream for spatial vision: An fMRI+ ERP study. *Human Brain Mapping*, *8*(4), 170-181.

Warren, G., Schertler, E., & Bull, P. (2009). Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior*, *33*(1), 59-69.

Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 550-564.

Weech, S., Kenny, S., & Barnett-Cowan, M. (2019). Presence and cybersickness in virtual reality are negatively related: a review. *Frontiers in psychology*, *10*, 158.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological review*, *92*(4), 548.

Wheeler, S. A., Gregg, D., & Singh, M. (2019). Understanding the role of social desirability bias and environmental attitudes and behaviour on South Australians' stated purchase of organic foods. *Food quality and preference*, *74*, 125-134.

Wilson, G. F., & Russell, C. A. (2003). Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human factors*, *45*(4), 635-644.

Wirtz, P. H., Ehlert, U., Emini, L., Rüdisüli, K., Groessbauer, S., Gaab, J., ... & von Känel, R. (2006). Anticipatory cognitive stress appraisal and the acute procoagulant stress response in men. *Psychosomatic medicine*, *68*(6), 851-858.

Wirtz, P. H., von Känel, R., Emini, L., Suter, T., Fontana, A., & Ehlert, U. (2007). Variations in anticipatory cognitive stress appraisal and differential proinflammatory cytokine expression in response to acute stress. *Brain, Behavior, and Immunity*, *21*(6), 851-859.

Wu, Y., Miwa, T., & Uchida, M. (2017). Using physiological signals to measure operator's mental workload in shipping–an engine room simulator study. *Journal of Marine Engineering & Technology*, *16*(2), 61-69.

Xiao, Y. M., Wang, Z. M., Wang, M. Z., & Lan, Y. J. (2005). The appraisal of reliability and validity of subjective workload assessment technique and NASA-task load index. *Chinese journal of industrial hygiene and occupational diseases*, *23*(3), 178-181.

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.

Yu, J. H., Albaum, G., & Swenson, M. (2003). Is a central tendency error inherent in the use of semantic differential scales in different cultures?. *International Journal of Market Research*, *45*(2), 1-16.

Zagermann, J., Pfeil, U., & Reiterer, H. (2016). Measuring cognitive load using eye tracking technology in visual computing. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization* (pp. 78-85). ACM.

Zhao, X., Li, X., & Yao, L. (2017). Localized fluctuant oscillatory activity by working memory load: a simultaneous EEG-fMRI study. *Frontiers in behavioral neuroscience*, *11*, 215.

Appendix

Correlation Matrix ("Pearson")

n0

r^2	diameter	Stress	Valence	Arousal	Dominance	RAW_TLX	blinks	ER	RT
diameter	1.00	0.20	-0.22	-0.14	-0.49	0.00	-0.20	-0.14	-0.14
Stress	0.20	1.00	-0.36	-0.09	-0.42	0.59	-0.32	0.46	-0.23
Valence	-0.22	-0.36	1.00	-0.55	0.18	-0.40	-0.46	-0.14	0.10
Arousal	-0.14	-0.09	-0.55	1.00	-0.08	0.37	0.39	0.19	-0.10
Dominance	-0.49	-0.42	0.18	-0.08	1.00	-0.10	0.11	-0.10	0.01
RAW_TLX	0.00	0.59	-0.40	0.37	-0.10	1.00	-0.06	0.47	-0.25
blinks	-0.20	-0.32	-0.46	0.39	0.11	-0.06	1.00	-0.15	-0.12
ER	-0.14	0.46	-0.14	0.19	-0.10	0.47	-0.15	1.00	-0.05
RT	-0.14	-0.23	0.10	-0.10	0.01	-0.25	-0.12	-0.05	1.00

p	diameter	Stress	Valence	Arousal	Dominance	RAW_TLX	blinks	ER	RT
diameter	NA	0.36	0.32	0.51	0.02	1.00	0.35	0.53	0.52
Stress	0.36	NA	0.09	0.69	0.04	0.00	0.14	0.03	0.30
Valence	0.32	0.09	NA	0.01	0.42	0.06	0.03	0.52	0.63
Arousal	0.51	0.69	0.01	NA	0.72	0.09	0.06	0.40	0.64
Dominance	0.02	0.04	0.42	0.72	NA	0.66	0.63	0.64	0.96
RAW_TLX	1.00	0.00	0.06	0.09	0.66	NA	0.79	0.02	0.24
blinks	0.35	0.14	0.03	0.06	0.63	0.79	NA	0.50	0.59
ER	0.53	0.03	0.52	0.40	0.64	0.02	0.50	NA	0.83
RT	0.52	0.30	0.63	0.64	0.96	0.24	0.59	0.83	NA

n1

r^2	diameter	Stress	Valence	Arousal	Dominance	RAW_TLX	blinks	ER	RT
diameter	1.00	0.23	0.13	-0.24	-0.29	0.07	0.37	0.30	0.04
Stress	0.23	1.00	-0.31	0.09	-0.50	0.45	0.35	0.31	-0.05
Valence	0.13	-0.31	1.00	-0.12	0.04	-0.25	-0.16	-0.20	-0.04
Arousal	-0.24	0.09	-0.12	1.00	-0.11	0.34	-0.32	0.05	-0.09
Dominance	-0.29	-0.50	0.04	-0.11	1.00	-0.19	0.07	-0.14	-0.17
RAW_TLX	0.07	0.45	-0.25	0.34	-0.19	1.00	0.17	0.54	0.21
blinks	0.37	0.35	-0.16	-0.32	0.07	0.17	1.00	0.13	0.12
ER	0.30	0.31	-0.20	0.05	-0.14	0.54	0.13	1.00	-0.22
RT	0.04	-0.05	-0.04	-0.09	-0.17	0.21	0.12	-0.22	1.00

p	diameter	Stressindex	Valence	Arousal	Dominance	RAW_TLX	blinks	ER	RT
diameter	NA	0.30	0.57	0.28	0.17	0.76	0.08	0.17	0.86
Stressindex	0.30	NA	0.15	0.70	0.02	0.03	0.10	0.15	0.83
Valence	0.57	0.15	NA	0.58	0.87	0.25	0.47	0.37	0.84
Arousal	0.28	0.70	0.58	NA	0.63	0.12	0.13	0.81	0.67
Dominance	0.17	0.02	0.87	0.63	NA	0.38	0.74	0.52	0.44

RAW_TLX	0.76	0.03	0.25	0.12	0.38	NA	0.44	0.01	0.34
blinks	0.08	0.10	0.47	0.13	0.74	0.44	NA	0.55	0.58
ER	0.17	0.15	0.37	0.81	0.52	0.01	0.55	NA	0.32
RT	0.86	0.83	0.84	0.67	0.44	0.34	0.58	0.32	NA

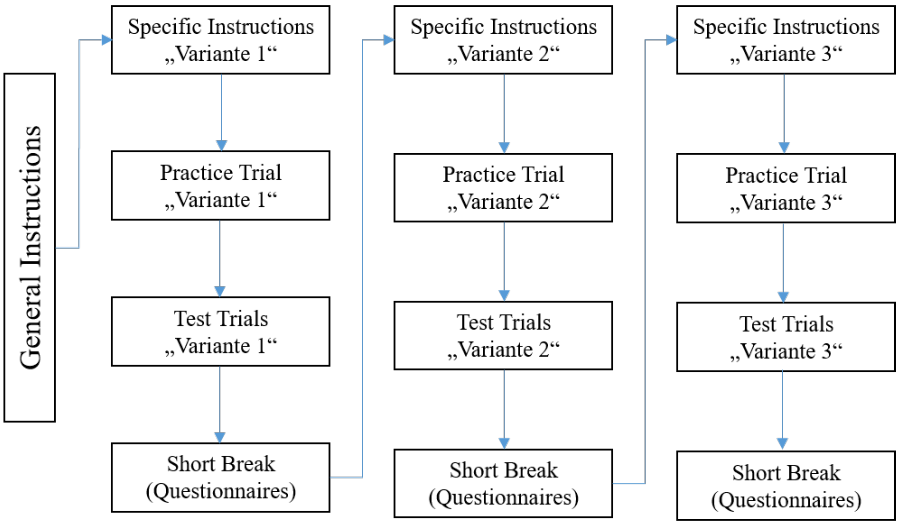
n2

r²	diameter	Stress	Valence	Arousal	Dominance	RAW_TLX	blinks	ER	RT
diameter	1.00	0.42	-0.33	0.14	-0.23	-0.04	-0.09	0.06	-0.11
Stress	0.42	1.00	-0.42	0.20	-0.12	0.48	-0.05	0.65	0.43
Valence	-0.33	-0.42	1.00	-0.41	0.22	-0.39	-0.04	-0.23	-0.23
Arousal	0.14	0.20	-0.41	1.00	-0.27	0.58	-0.36	0.00	0.01
Dominance	-0.23	-0.12	0.22	-0.27	1.00	-0.08	0.35	0.32	-0.01
RAW_TLX	-0.04	0.48	-0.39	0.58	-0.08	1.00	-0.17	0.42	0.54
blinks	-0.09	-0.05	-0.04	-0.36	0.35	-0.17	1.00	0.17	-0.03
ER	0.06	0.65	-0.23	0.00	0.32	0.42	0.17	1.00	0.59
RT	-0.11	0.43	-0.23	0.01	-0.01	0.54	-0.03	0.59	1.00

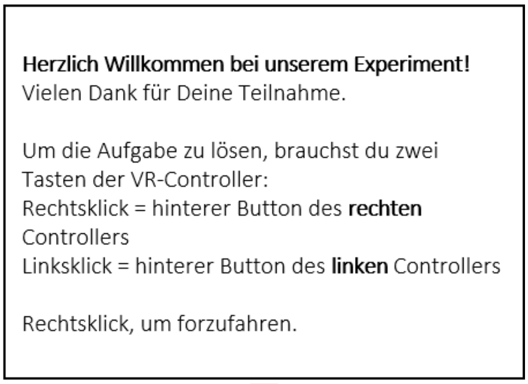
p	diameter	Stressindex	Valence	Arousal	Dominance	RAW_TLX	blinks	ER	RT
diameter	NA	0.05	0.13	0.52	0.28	0.84	0.68	0.78	0.61
Stressindex	0.05	NA	0.04	0.37	0.59	0.02	0.81	0.00	0.04
Valence	0.13	0.04	NA	0.05	0.31	0.06	0.87	0.30	0.28
Arousal	0.52	0.37	0.05	NA	0.22	0.00	0.10	1.00	0.97
Dominance	0.28	0.59	0.31	0.22	NA	0.73	0.10	0.14	0.96
RAW_TLX	0.84	0.02	0.06	0.00	0.73	NA	0.44	0.04	0.01
blinks	0.68	0.81	0.87	0.10	0.10	0.44	NA	0.45	0.89
ER	0.78	0.00	0.30	1.00	0.14	0.04	0.45	NA	0.00
RT	0.61	0.04	0.28	0.97	0.96	0.01	0.89	0.00	NA

N-back task procedure

(modified version due to printing purposes)


General Instructions
Specific Instructions „Variante 1“
Practice Trial „Variante 1“
Test Trials „Variante 1“
Short Break (Questionnaires)
Specific Instructions „Variante 2“
Practice Trial „Variante 2“
Test Trials „Variante 2“
Short Break (Questionnaires)
Specific Instructions „Variante 3“
Practice Trial „Variante 3“
Test Trials „Variante 3“
Short Break (Questionnaires)

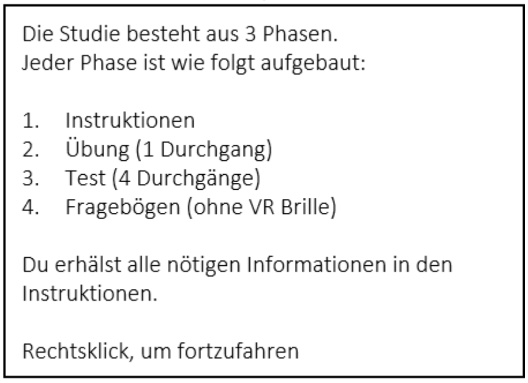
General Instructions



Herzlich Willkommen bei unserem Experiment!
Vielen Dank für Deine Teilnahme.

Um die Aufgabe zu lösen, brauchst du zwei Tasten der VR-Controller:
Rechtsklick = hinterer Button des **rechten** Controllers
Linksklick = hinterer Button des **linken** Controllers

Rechtsklick, um forzufahren.

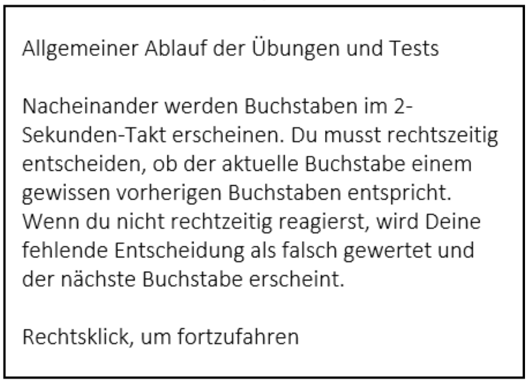


Die Studie besteht aus 3 Phasen.
Jeder Phase ist wie folgt aufgebaut:

1. Instruktionen
2. Übung (1 Durchgang)
3. Test (4 Durchgänge)
4. Fragebögen (ohne VR Brille)

Du erhälst alle nötigen Informationen in den Instruktionen.

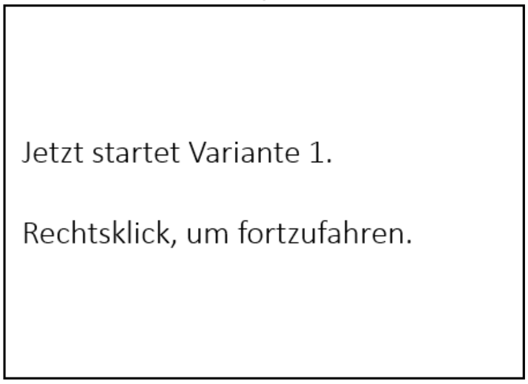
Rechtsklick, um fortzufahren



Allgemeiner Ablauf der Übungen und Tests

Nacheinander werden Buchstaben im 2-Sekunden-Takt erscheinen. Du musst rechtszeitig entscheiden, ob der aktuelle Buchstabe einem gewissen vorherigen Buchstaben entspricht. Wenn du nicht rechtzeitig reagierst, wird Deine fehlende Entscheidung als falsch gewertet und der nächste Buchstabe erscheint.

Rechtsklick, um fortzufahren

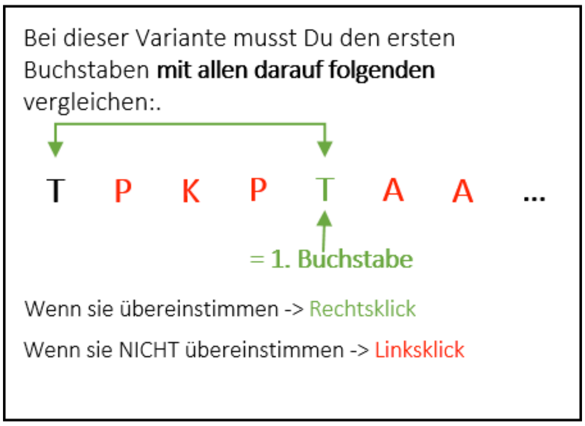


Jetzt startet Variante 1.

Rechtsklick, um fortzufahren.

Specific Instructions

0-back



Bei dieser Variante musst Du den ersten Buchstaben **mit allen darauf folgenden** vergleichen:.

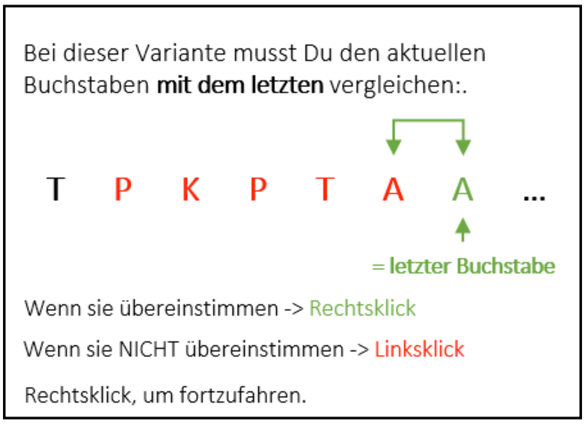
T P K P T A A ...

= 1. Buchstabe

Wenn sie übereinstimmen -> Rechtsklick

Wenn sie NICHT übereinstimmen -> Linksklick

1-back



Bei dieser Variante musst Du den aktuellen Buchstaben **mit dem letzten** vergleichen:.

T P K P T A A ...

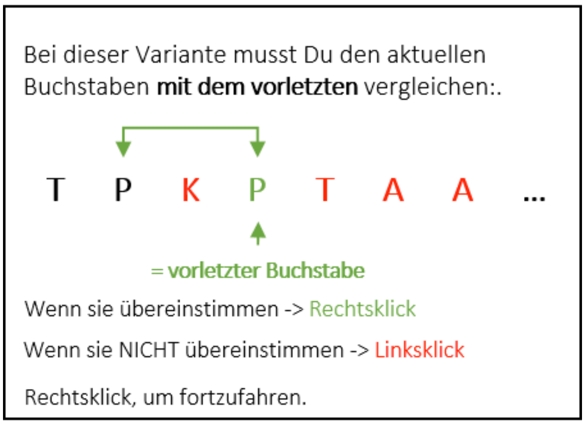
= letzter Buchstabe

Wenn sie übereinstimmen -> Rechtsklick

Wenn sie NICHT übereinstimmen -> Linksklick

Rechtsklick, um fortzufahren.

2-back



Bei dieser Variante musst Du den aktuellen Buchstaben **mit dem vorletzten** vergleichen:.

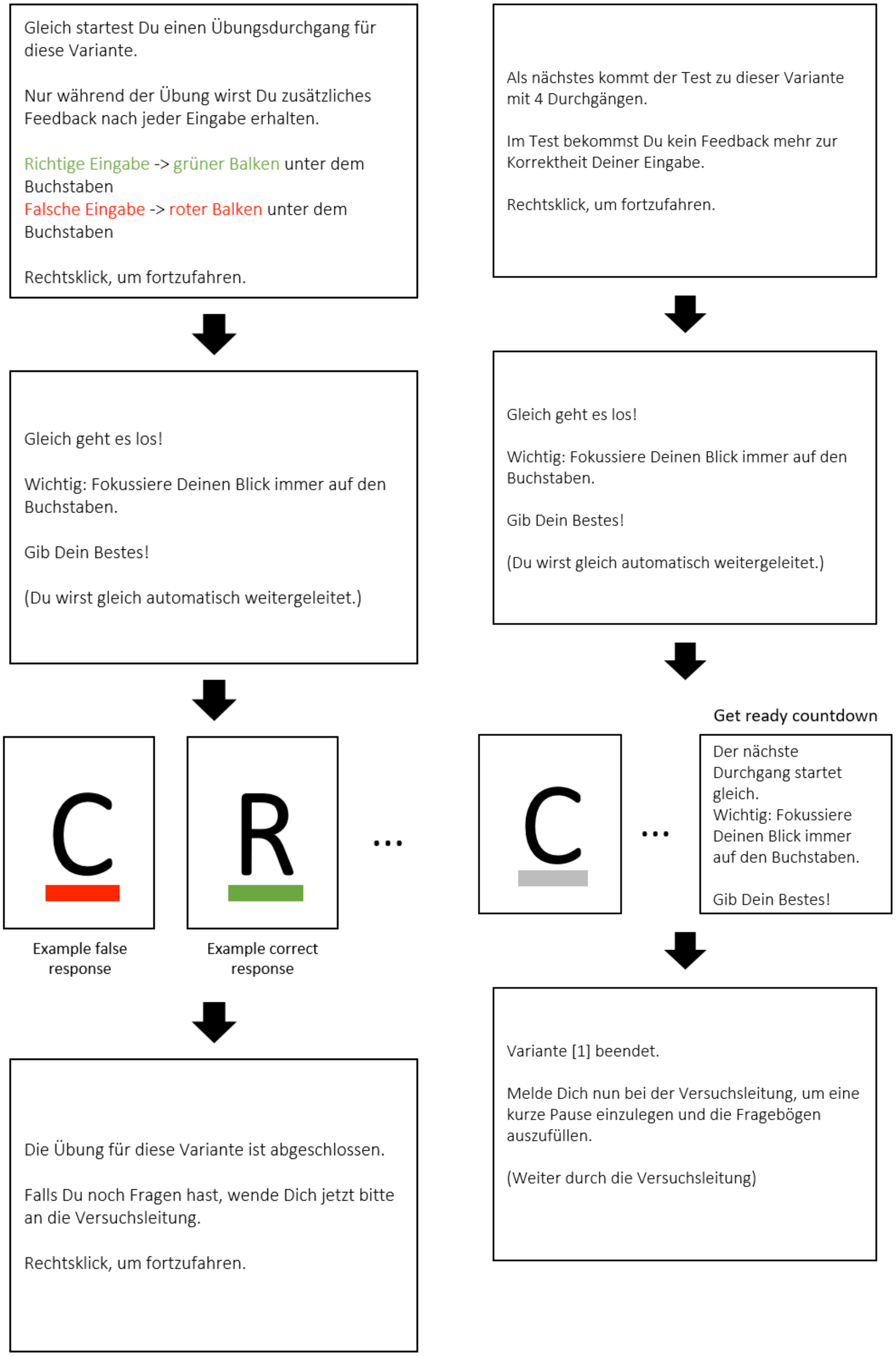
T P K P T A A ...

= vorletzter Buchstabe

Wenn sie übereinstimmen -> Rechtsklick

Wenn sie NICHT übereinstimmen -> Linksklick

Rechtsklick, um fortzufahren.



Practice Trial

Gleich startest Du einen Übungsdurchgang für diese Variante.

Nur während der Übung wirst Du zusätzliches Feedback nach jeder Eingabe erhalten.

Richtige Eingabe -> grüner Balken unter dem Buchstaben
Falsche Eingabe -> roter Balken unter dem Buchstaben

Rechtsklick, um fortzufahren.

Gleich geht es los!

Wichtig: Fokussiere Deinen Blick immer auf den Buchstaben.

Gib Dein Bestes!

(Du wirst gleich automatisch weitergeleitet.)

Example false response

Example correct response

Die Übung für diese Variante ist abgeschlossen.

Falls Du noch Fragen hast, wende Dich jetzt bitte an die Versuchsleitung.

Rechtsklick, um fortzufahren.

Test Trials

Als nächstes kommt der Test zu dieser Variante mit 4 Durchgängen.

Im Test bekommst Du kein Feedback mehr zur Korrektheit Deiner Eingabe.

Rechtsklick, um fortzufahren.

Gleich geht es los!

Wichtig: Fokussiere Deinen Blick immer auf den Buchstaben.

Gib Dein Bestes!

(Du wirst gleich automatisch weitergeleitet.)

Get ready countdown

Der nächste Durchgang startet gleich.
Wichtig: Fokussiere Deinen Blick immer auf den Buchstaben.

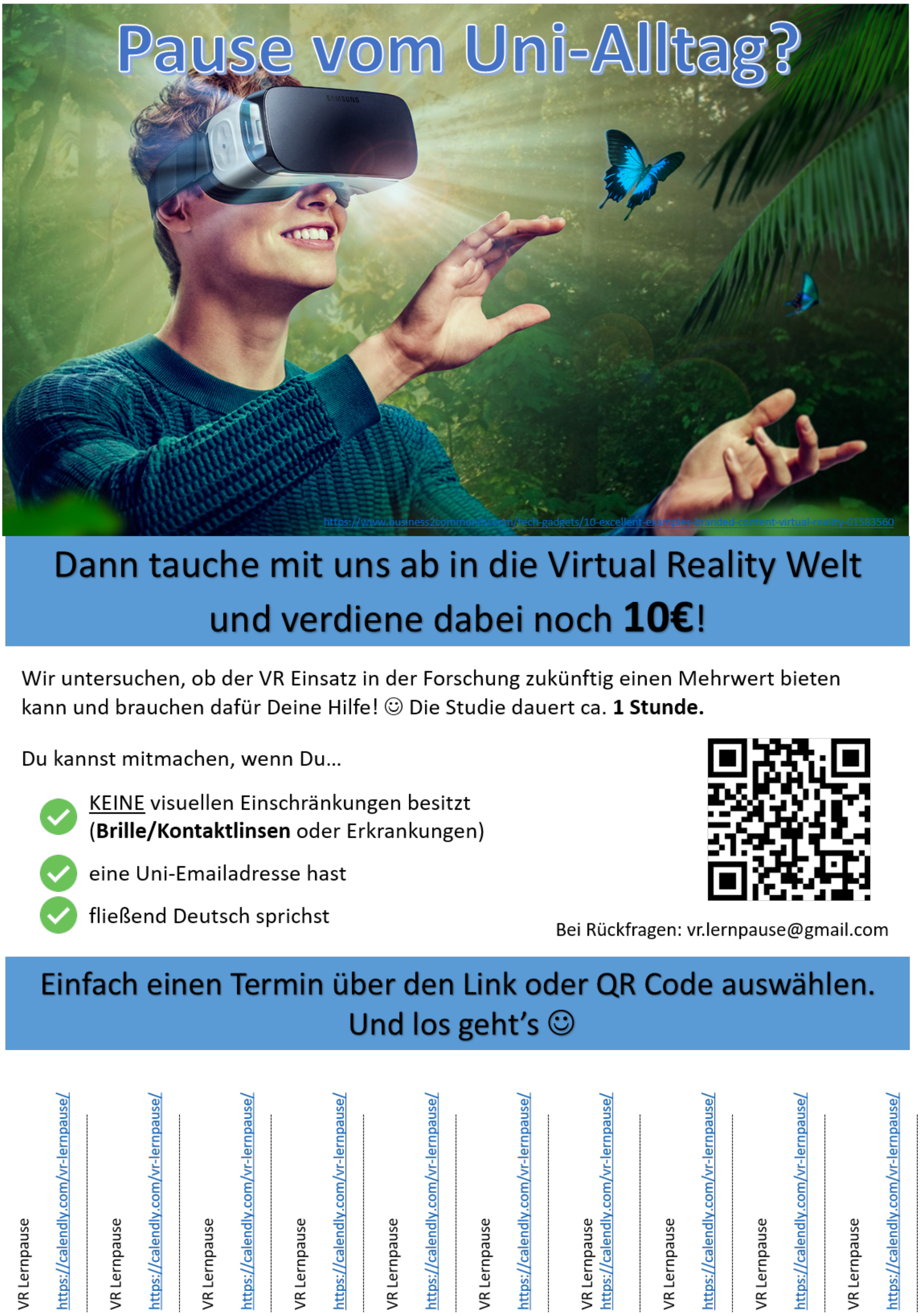
Gib Dein Bestes!

Variante [1] beendet.

Melde Dich nun bei der Versuchsleitung, um eine kurze Pause einzulegen und die Fragebögen auszufüllen.

(Weiter durch die Versuchsleitung)

Brochure



Pause vom Uni-Alltag?

https://www.business2community.com/tech-gadgets/10-excellent-examples-branded-content-virtual-reality-01583560

Dann tauche mit uns ab in die Virtual Reality Welt und verdiene dabei noch **10€**!

Wir untersuchen, ob der VR Einsatz in der Forschung zukünftig einen Mehrwert bieten kann und brauchen dafür Deine Hilfe! ☺ Die Studie dauert ca. **1 Stunde.**

Du kannst mitmachen, wenn Du…

- ✓ KEINE visuellen Einschränkungen besitzt (**Brille/Kontaktlinsen** oder Erkrankungen)
- ✓ eine Uni-Emailadresse hast
- ✓ fließend Deutsch sprichst

Bei Rückfragen: vr.lernpause@gmail.com

Einfach einen Termin über den Link oder QR Code auswählen. Und los geht's ☺

VR Lernpause
https://calendly.com/vr-lernpause/

VR Lernpause
https://calendly.com/vr-lernpause/

VR Lernpause
https://calendly.com/vr-lernpause/

VR Lernpause
https://calendly.com/vr-lernpause/

VR Lernpause
https://calendly.com/vr-lernpause/

VR Lernpause
https://calendly.com/vr-lernpause/

VR Lernpause
https://calendly.com/vr-lernpause/

VR Lernpause
https://calendly.com/vr-lernpause/

VR Lernpause
https://calendly.com/vr-lernpause/

VR Lernpause
https://calendly.com/vr-lernpause/

Study Case Report Form

CASE REPORT FORM

Studienprotokoll für Cogntive Load Messung mit VR Brille
Sommersemester 2019
Universität Konstanz
Studienleitung: Ariane Rahn

Utilizing a Virtual Environment to Measure Cognitive Load using Eye Tracking Technology

Probanden-Code: ________________________

Gruppenzugehörigkeit: ________________________

Datum Untersuchungstag: __________________

Sonstige Notizen:

Uhrzeit	Aufgabe	Erledigt
	Vorbereitung Labor PC: Pupil Capture starten (Desktop) Steam VR → VR Ansicht anzeigen Unity Hub starten (Windows-Menü) -> Projekt: Eyetracking VR Study Laptop: Digit Memory Test öffnen (ohne Audio) Materialen bereit? (VR Brille, Fragebögen, Makeup Entferner, Wasser) Schild an Tür?	Erledigt ❒ Nein ❒ Ja
_ _:_ _	Ankunft am Labor, Begrüßung Jacke etc abnehmen Wasser anbieten	Erledigt ❒ Nein ❒ Ja
	Abfrage Ausschlusskriterien - Uni-Emailadresse - Keine visuelle Einschränkung - Fließend Deutsch - Kein Make-Up?	Erledigt ❒ Nein ❒ Ja
	Einverständniserklärung 2x **VL:** eine rauslegen für Rückgabe am Ende; eine in Ablage für erledigte Aufgaben Lockere, angenehme, vertraute, ruhige Atmosphäre schaffen. Auf Fragen und Sorgen eingehen. ***Handy ausschalten*** → *darf nicht genutzt werden*	Erledigt (2x) ❒ Nein ❒ Ja Erledigt ❒ Nein ❒ Ja Erledigt ❒ Nein ❒ Ja
	Wie wach fühlen Sie sich gerade? □ hellwach □ wach □ etwas abgeschlagen □ müde □ sehr müde	Erledigt ❒ Nein ❒ Ja
_ _:_ _	Digit Span Memory Test _______ Score	Erledigt ❒ Nein ❒ Ja
	N-Back Task	
	Proband vorbereiten - Sitzplatz + Handposition - VR Brille aufsetzen - VR Controller aktivieren (Button unterhalb Touchpad) - VR Controller In richtiger Hand? → Handzeichen auf Controller - VR Brille kalibrieren durch Anweisungen *„Sitzt die Brille gut? Alles gut?*	Erledigt ❒ Nein ❒ Ja
_ _:_ _	**Programme starten** - Unity: „Play“, dann sofort „R“ - Unity: in Feld „Game“ klicken - Bed. 1 – 6 auswählen: _________ **Bedingung ausgewählt** - Programm startet mit „Rechtsklick“ (oder Leertaste)	Erledigt ❒ Nein ❒ Ja
	Phase 1 **Übung + Testphase** *VL: Fragebögen vorbereiten/beschriften*	Erledigt ❒ Nein ❒ Ja

	NASA TLX	Erledigt ❒ Nein ❒ Ja
	PASA	Erledigt ❒ Nein ❒ Ja
	SAM	Erledigt ❒ Nein ❒ Ja
	Proband vorbereiten - Sitzplatz + Handposition - VR Brille - VR Brille kalibrieren	Erledigt ❒ Nein ❒ Ja
_ _:_ _	**Phase 2 (VL: Leertaste drücken)** **Übung + Testphase** *VL: Fragebögen vorbereiten/beschriften*	Erledigt ❒ Nein ❒ Ja
	NASA TLX	Erledigt ❒ Nein ❒ Ja
	PASA	Erledigt ❒ Nein ❒ Ja
	SAM	Erledigt ❒ Nein ❒ Ja
	Proband vorbereiten - Sitzplatz + Handposition - VR Brille - VR Brille kalibrieren	Erledigt ❒ Nein ❒ Ja
_ _:_ _	**Phase 3 (VL: Leertaste drücken)** **Übung + Testphase** *VL: Fragebögen vorbereiten/beschriften*	Erledigt ❒ Nein ❒ Ja
	Aufnahme beenden: Unity: erneut auf „Play“ drücken (Aufnahme stoppt automatisch)	Erledigt ❒ Nein ❒ Ja
	NASA TLX	Erledigt ❒ Nein ❒ Ja
	PASA	Erledigt ❒ Nein ❒ Ja
	SAM	Erledigt ❒ Nein ❒ Ja
Abschluss		
_ _:_ _	**Demographics**	Erledigt ❒ Nein ❒ Ja
	Verabschiedung - Einverständniserklärung geben - Geld **quittiert** ausgezahlt - Fragen klären / Interesse an Ergebnissen	Erledigt ❒ Nein ❒ Ja ❒ Nein ❒ Ja
	Nachbereitung Labor - Alle Unterlagen beschriftet? - Materialien gebündelt verstaut, Daten auf USB Stick? - Brille desinfizieren - Labor für nächsten Proband vorbereiten - Controller aufladen	Erledigt ❒ Nein ❒ Ja ❒ Nein ❒ Ja

Demographics

Fragebogen mit allgemeinen Informationen zur Person

VPN-Code: ___________ Datum: ____________________
(dd.mm.yyyy)

(von Versuchsleitung ausgefüllt)

Alter: ________ Jahre
Geschlecht: □ Männlich □ weiblich □ divers

Primäre Tätigkeit/Beruf: □ Studierende/r im Fach: ____
□ andere, und zwar: ____

Welche Hand ist bei Ihnen die dominante?

□ Links □ Rechts □ nicht eindeutig

Leiden Sie unter irgendwelchen chronischen oder akuten körperlichen Erkrankungen?

□ Nein
Wenn ja, welche? ____

Leiden Sie chronisch oder akut unter diagnostizierten psychischen Störungen?

□ Nein
Wenn ja, welche? ____

Nehmen Sie zur Zeit Medikamente ein, sowohl ärztlich verordnete als auch andere? (Kontrazeptiva/die Pille, Schmerzmittel, Ritalin, etc. gelten als Medikamente)

□ Nein □ Ja
Wenn ja, welche? ____

Haben Sie Erfahrung mit der Nutzung einer Virtual Reality Umgebung (heutige Studie ausgenommen)?

□ Nein
□ Einmalig
□ Selten
□ Regelmäßig
□ Sehr oft

Wenn ja, in welchem Kontext nutzen Sie eine Virtual Reality Umgebung?

□ Spaß / Unterhaltung □ Beruflicher Kontext □ Akademischer Kontext

Traten bei Ihnen bei der heutigen VR Nutzung Beschwerden von Übelkeit auf?
□ Nein □ Einmalig □ Selten □ Regelmäßig □ Sehr oft

Haben Sie bereits an einem Kurzzeit-Gedächtnis Test teilgenommen?

□ Nein

□ Ja
Wenn ja, welche(r)? Beschreiben Sie bitte kurz.

____

____

____

____

Haben Sie schon einmal an den heutigen Kurzzeitgedächtnis-Tests teilgenommen (z.B. in anderen Studien)?

□ Nein
□ Ja, an dem Digit Span Memory Test (der Zahlentest zu Anfang der Studie)
□ Ja, an dem n-back Test (der Buchstabentest in VR Umgebung)

Waren Sie motiviert, die Tests bestmöglich zu bearbeiten?

Ich war sehr motiviert	□
Ich war eher motiviert	□
Ich war eher nicht motiviert	□
Ich war gar nicht motiviert	□