VIRTUAL POSSIBILITIES: EXPLORING THE ROLE OF EMERGING TECHNOLOGIES IN WORK AND LEARNING ENVIRONMENTS

Vom Fachbereich Sozialwissenschaften
der Rheinland-Pfälzischen Technischen Universität, Campus Kaiserslautern
zur Verleihung des akademischen Grades
Doktor rerum naturalium (Dr. rer.nat.)
genehmigte

D i s s e r t a t i o n

vorgelegt von
Felix Hekele, MSc

Tag der Disputation:	Kaiserslautern, 21. März 2024
Dekan:	Prof. Dr. Michael Fröhlich
Vorsitzende/r:	Prof. Dr. Shanley E.M. Allen
Gutachter/in:	1. Prof. Dr. Thomas Lachmann
	2. apl. Prof. Dr. Daniela Czernochowski

DE 386

April 2024

“The reasonable man adapts himself to the world. The unreasonable one persists in trying to adapt the world to himself. Therefore, all progress depends on the unreasonable man.”
George Bernard Shaw (Aphorisms to *Man and Superman*)

“Maybe it's a dream, maybe nothing else is real
But it wouldn't mean a thing if I told you how I feel.”

Lyrics, “*Bad Apple!!*” by Masayoshi Minoshima

Statement of Authorship

Hiermit versichere ich,

- dass ich die vorgelegte Arbeit selbst angefertigt und alle benutzten Hilfsmittel in der Arbeit angegeben habe,
- dass ich diese Dissertation nicht schon als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht, und
- dass weder die gleiche noch eine andere Abhandlung der Dissertation bei einer anderen Universität oder einem anderen Fachbereich der Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau veröffentlicht wurde.

18. Dezember 2023, Kaiserslautern

Felix Hekele

Contents VII

Remarks

Chapter 3.1 has been published as “Relevance of the assessment mode in the digital assessment of processing speed” by Francisca Rodriguez, Jan Spilski, Andreas Schneider, Felix Hekele, Thomas Lachmann, Achim Ebert and Franca Rupprecht (2019, in Journal of Clinical and Experimental Neuropsychology). It will be referred to as *study 1* in the following text.

Chapter 3.2 has been published as “Physical and cognitive demands of work in building construction” by Francisca Rodriguez, Jan Spilski, Felix Hekele, Nils Beese and Thomas Lachmann (2020, in Engineering, Construction and Architectural Management). The content of the paper has not been changed. The references section in the paper did not follow APA guidelines in the published version and has been adapted to conform with the rest of the work. It will be referred to as *study 2* in the following text.

Chapter 4.1 has been published as “Remote vocational learning opportunities—A comparative eye-tracking investigation of educational 2D videos versus 360° videos for car mechanics” by Felix Hekele, Jan Spilski, Simon Bender and Thomas Lachmann (2022, in British Journal of Educational Technology). It will be referred to as *study 3* in the following text.

Chapter 4.2 has been published as “Spatial Sound in a 3D Virtual Environment: All Bark and No Bite?” by Radha Nila Meghanathan, Patrick Ruediger-Flore, Felix Hekele, Jan Spilski, Achim Ebert and Thomas Lachmann (2021, in Big Data and Cognitive Computing). The references section in the paper did not follow APA guidelines in the published version and has been adapted to conform with the rest of the work. The content of the paper has not been changed. It will be referred to as *study 4* in the following text.

Acknowledgements

Six years of research, projects, experiments, and conferences find their conclusion with the submission of this dissertation. None of this would have been possible without the incredible amount of effort of a whole lot of people, who I would like to thank in the following, without any particular order.

The first thanks have to go out to my immediate family, my parents Frank and Christine Hekele as well as my brother Fabian. Your emotional and motivational support over the many years of studies, whether in Salzburg, Wales, or now in Kaiserslautern has been nothing short of incredible. Thank you for the tireless support over phone or in person when everything was down in the gutter again. Thank you for enabling me to live my dream and believing in me, especially in those times when I did not believe in myself. Thank you for being there for me no matter the circumstances. Words can scarcely describe how much I own the three of you.

A thank you to both of my grandmothers, Waldtraut Hekele and Irma Stadler for the loving care and support, for listening to my rambling about science even if it was more often than not completely incomprehensible to you. I owe a lot to the both of you and feel your influence on many of my actions every day. Thank you for being the bridge from the ivory tower of Academia to my home in rural Bavaria.

Thank you to my supervisor Thomas Lachmann, for both giving me the freedom I needed to get to the point where I am now, but also for giving me the feeling that I am never alone. Your immense support and guidance, especially when writing my “own” first-authorship paper was immeasurable.

A big thanks also to Jan Spilski as a colleague and mentor at times, I deeply appreciate your help and support in project-related problems and in general. I was able to learn a great amount from you over the years and will keep the collaborative work in projects as well as the late evening coffee discussions in good memory.

Thank you as well to Udo Petruschkat, not only for the great collaborations but also the fun discussions and always keeping a positive atmosphere, especially when the sleepy academic had to be kept awake for the 8am appointments. I always enjoyed coming to Iserlohn and appreciate your assistance with our research. I wish you and your soon growing family all the best in the future.

Thank you to my colleagues and friends Nils Beese, Chris Allison and Omar Jubran for all the input, the “tech guy” talks, the many Mensa breaks – whether the buffet was there or not (soon™). Thank you for having my back, for the many science discussions, the shared coffee and research. Without it, I likely would have thrown the towel years ago. I can’t wait to see what each of you will achieve in the future.

Last but not least, a big thank you also to my co-authors, the many colleagues from the InKraFT, ConWearDi, KaMeRi and MERLOT projects and the many students from our cognitive science programme. Thank you for your support and hard work over the many years, whether “in person” or online. I always enjoyed the double-track drifting of project work and university research and wouldn’t want to miss the experience I gleaned from the dozens of studies, even if 2020 and 2021 did their very best (or worst) to keep us from conducting our research.

Danksagung

Mit der Einreichung dieser Doktorarbeit enden ganze sechs Jahre Forschung und Projektarbeit, Studien und Konferenzen. Diese Arbeit wäre ohne die große Unterstützung vieler Menschen nicht möglich gewesen, welchen im Folgenden herzlich gedankt werden soll. Dies erfolgt in keiner spezifischen Reihenfolge.

Der erste Dank gebührt meinen Eltern, Frank und Christine Hekele, sowie meinem Bruder Fabian, die mich über die vielen Jahre in Kaiserslautern, Wales als auch in Salzburg emotional und motivational unterstützt haben. Danke für die unermüdliche Unterstützung, wenn ich mich am Telefon mal wieder länger auslassen musste. Danke, dass ihr mir ermöglicht habt, meinem Traum zu folgen und auch dann an mich geglaubt habt, wenn ich es selbst nicht getan habe. Danke dafür, dass ihr immer für mich da wart und seid. Ich hoffe, ich kann euch dies in irgendeiner Form angemessen guttun, denn in Worte kann ich es kaum fassen.

Danke auch an meine beiden Großmütter, Waldtraut Hekele und Irma Stadler, für die herzliche Unterstützung, und ein immer offenes Ohr, auch wenn der Inhalt meiner Forschung vielleicht nicht mehr wirklich greifbar war. Ihr habt mich beide tief geprägt und eure unsichtbare Hand spiegelt sich auch in dieser Arbeit. Danke, dass ihr die Brücke aus dem universitären Elfenbeinturm zur Heimat nach Eggstätt immer aufrechterhalten habt.

Danke an meinen Betreuer Thomas Lachmann, dafür dass du mir die Freiheit gelassen hast, die ich brauchte und mir gleichzeitig doch auch das Gefühl gegeben hast, dass ich nie alleine dastehe. Besonders bei meiner ersten „eigenen“ Publikationen war dein Rat und Engagement von unschätzbarem Wert.

Danke auch an Jan Spilski, der mir als Kollege und Quasi-Mentor sehr geholfen hat und immer ein offenes Ohr bei Projektsorgen hatte. Ich konnte über die Jahre unheimlich viel von dir lernen, und werde sowohl die Zusammenarbeit in den Projekten, als auch die abendlichen Gespräche beim Mitternachtskaffee in guter Erinnerung behalten.

Danke an Herrn Udo Petruschkat, nicht nur für die tolle Zusammenarbeit in Iserlohn, aber auch die lustigen Gespräche und die immer gute Stimmung, wenn der verschlafene Akademiker um Acht auf der Matte stehen musste. Ich bin immer gerne nach Iserlohn gekommen und schätze deine Mithilfe bei unserer Forschung sehr. Ich wünsche dir und deiner Familie samt baldigen Zuwachs alles Gute für die Zukunft.

Danke an meine Freunde und Kollegen Nils Beese, Chris Allison und Omar Jubran für Rat und Tat, „Tech Guy“ Diskussionen, die unzähligen Mensa-Runden – ob mit oder ohne Buffet (soon™). Ohne euren Rückhalt, die Kaffeerunden und generellen Zusammenhalt hätte ich wahrscheinlich irgendwann aufgegeben. Ich bin sehr gespannt, was ihr in Zukunft erreichen werdet!

Danke auch an meine Ko-Autoren, die vielen Projektkollegen aus InKraFT, ConWearDi, KaMeRi, MERLOT sowie die vielen Studierenden unseres Cognitive Science Masterprogramms für die tatkräftige Zusammenarbeit über viele Jahre hinweg, ob online oder doch „in Person“. Ich habe sowohl die Projektarbeit als auch die universitäre Forschung stets genossen und möchte die Erfahrungen aus keiner einzelnen der Dutzenden Studien missen, auch wenn die Umstände insbesondere im Jahr 2020 und 2021 ihr bestes taten, um unseren Forschungsdrang zu mildern.

Glossary

AoI	Area of Interest
AR	Augmented Reality
ACC	Anterior Cingulate Cortex
CAMIL	Cognitive Affective Model of Immersive Learning (Makransky & Peterson, 2021)
CLT	Cognitive Load Theory (Sweller, 1988)
DoF	Degrees of Freedom
F-JAS	Fleishman - Job Analyse System (German Version by Kleinmann, Manzey, Schumacher, & Fleishman, 2010)
GTL	Gaze Trajectory Length
HCI	Human-Computer Interaction
HMD	Head-Mounted Display
IPQ	iGroup Presence Questionnaire (Schubert, Friedmann, & Regenbrecht, 1999; 2001)
L2	Secondary Language
MDQ	Mental Demands Questionnaire (Handel, 2016)
MR	Mixed Reality (alternatively: XR)
nVR	Non-interactive Virtual Reality
RODOS	Robot based Driving and Operation Simulator
TFF	Time to First Fixation
TLX	NASA Task Load Index (Hart & Staveland, 1988)
TMT	Trail-Making Test (Oswald & Roth, 1987)
VE	Virtual Environment
VR	Virtual Reality
VET	Vocational Education and Training

List of figures

List of tables

Chapter 1: General Introduction

Following the roots of their discipline, psychologists have been intrigued by the way humans perceive and interact with the world around them. The field has been trying to understand and classify human behaviour since its inception around 150 years ago. From the beginning, psychological experiments relied on technology to present stimuli and measure reaction times, such as the tachistoscope and myographs employed by Wilhelm Wundt's laboratory of experimental psychology in the second half of the 19th century (see Mandler, 2011, for a detailed overview). Briefly after, scientists began trying to measure eye movements, first with invasive mechanical setups, which were slowly replaced in the late 1940s with the first head-mounted eye-trackers (Cognolato, Atzori, & Müller, 2018). Since these humble beginnings, experimental psychologists and behavioural scientists have worked with increasingly complex devices on their quest to better understand the human behaviour and human cognition. Over time not only have the experimental settings grown more sophisticated (Brown, Reeves, & Sherwood, 2011), the speed with which technological advancements have been invented and adapted has been increasing. Gordon Moore predicted in 1965 that the number of components in an integrated circuit doubles roughly every two years (Moore, 1965; 1998). In what is often referred to as "Moore's Law" the number of components per circuit, which is closely related to the computing power, doubles every two years – in common understanding the assumption is that computers become twice as powerful every two years. Over the last 60 years since Moore's prediction technology has advanced in many ways with devices becoming smaller and more powerful, leading to wearable electronics in the modern age. For scientific research this progress has allowed experimenters to not only collect data such as skin conductance using small or even remote devices (Hoogeboom, Saeed, Noordzij, & Wilderom, 2021) but in some cases to leave lab settings behind and expand studies to the outside world, for example to investigate natural gaze behaviour in crowds (Hessels et al., 2022). The technological progress has not halted and at the time of writing virtual reality, augmented reality and mixed reality are all used in educational research, while still considered emerging technologies (Xiong, Hsiang, He, Zhan, & Wu, 2021) – which they have been for a long time. This stands in contrast to other emerging technologies such as mobile phones and home computers which became everyday utility devices.

The technological progress has become nigh omnipresent in our daily lives over the last 20-30 years. With the rise of the internet and the ever-increasing connectivity of humans not only with each other, but also with machines and devices. Machines have a previously rarely seen capability to influence behaviour and create entire careers and lifestyles - not just for a select few, but for the majority of the global population. In 2022, around 88,1% of households in Germany had access to at least one smartphone and 92% of households at least one computer (Federal Statistical Office, 2022), making these devices near ubiquitous for many people.

Can these devices be used to facilitate our ability to interact with and learn in an increasingly complex world? Young people who are trying to find their place in this world often feel overwhelmed by the sheer number of choices to make – which school to pick, which careers are there, which job do I want to learn – these are but some of the crucial questions which can lead to a lot of pressure on the individual (Lüdtke, Roberts, Trautwein, & Nagy, 2011; Chen, Liu, & Wen, 2023).

Information technology such as computer-based learning and, by an extend, social media often played and plays a role in different aspects of people's lives, including education. Research into social media has shown that it can be used to increase the connectedness and engagement as well as reduce isolation (Lee, Jang, Pena-y-Lillo, & Wang, 2020). For university students, there has been evidence by Phuthong (2021) that the capability for collaboration and perceived enjoyment can predict the use of social media as learning platforms. Freina and Ott (2015) see potential for integrating modern technology such as virtual reality into traditional education settings, a sentiment with has been by other work groups, especially in combination as a form of blended learning (Dziuban, Graham, Moskal, Norberg, & Sicilia, 2018). On the less positive side, a meta-analysis by Marino and colleagues (2018) compiled evidence for a significant correlation between social media use and personal distress. Social media is also used for educational purposes, where it provides learners with additional opportunities through courses and materials (Huang, Spector, & Yang, 2019) and prepares learners by training them in skills such as remote collaboration and effective communication which became a necessity through the Covid-19 pandemic (Bonsu, Bervell, Armah, Aheto, & Arkorful, 2021). The educational potential for virtual reality lies more with the possibility to individualize learning experiences for each learner as well as engaging them more in the learning process (Bacca, Baldiris, Fabregat, Kinshuk, & Graf, 2015). While information technology has revolutionized the way learners can access

and interact with educational material, it has also become clear that teachers and educators need to be trained to deal with these changes in the educational landscape (Burbules, 2016; Ichou, 2018).

With an increasingly technology-driven world, educational institutions are also confronted with several major challenges. Over the last decades the number of children who got diagnosed with learning disabilities has increased dramatically, which led to a greatly increased demand for individual care and support in the school system (Grigorenko et al., 2020). In the same timeframe, most countries in Europe have seen a shift in the educational preferences of young people with less people pursuing a vocational career and more young students enrolling in universities. In Germany, the number of young people in vocational education program decreased by nearly 15% over the last decade (Statistisches Bundesamt, 2023a) while number of students enrolling in university has increased from 230000 in 1998 to 398000 in 2022 (Statistisches Bundesamt, 2023b). Along this shift there has been growing need from the job market for highly specialized and well-trained students, which led to increased social inequality in education (Kromydas, 2017). At the same time, technology also holds the promise of opportunity, potentially also to decrease this inequality - if embedded properly within the educational system for everyone (Ichou, 2018). In some areas of education, virtual reality has begun to transform the educational process of vocational students (Freina & Ott, 2015) and medical students (Ekstrand, Jamal, Ngyuen, Kudryk, Mann, & Mendet, 2018), while some areas such as collaborative learning is still in its infancy (Bower, Lee, & Dalgarno, 2016). Martín-Gutiérrez and colleagues (2017) described virtual reality's high potential for inclusiveness of people with disabilities by creating accessible experiences. However, VR also still has shortcomings such as the lack of specialized educational content in many fields (Jensen & Konradsen, 2018) and availability issues due to the cost of acquisition and maintenance of the equipment (Martín-Gutiérrez, Mora, Añorbe-Díaz, & González-Marrero, 2017).

In front of this challenging landscape, the present work seeks to investigate ways to utilize emerging technology, mostly in form of virtual reality for their use in vocational educational contexts to enable as many people as possible to develop their potential. Since the technology is still in its infancy, a multi-angle approach was used: To start, different input methods will be compared in a classic lab-based study using a tablet with different input methods (Rodriguez et al., 2019). In recent years, there has been an increased focus on input methods especially in the field of human-computer interaction, and VR offers a particular range of interaction

through stimulating different sensory modalities (Dede, 2009). Study 1 aimed to lay the foundation for further research in the working group by investigating multiple modalities and placed the focus on the issue of comparability, which is often raised as shortcoming in educational research (Rickinson, 2001; Göransson & Nilholm, 2014; Buchner, Buntins, & Kerres, 2022).

The next two papers discussed in the present work are focussing on vocational education. In order to design adequate educational content, it is important to first understand the needs of the learners, e.g. by mapping their daily work demands to tailor educational content towards the learner. Study 2 aimed to investigate the demands construction workers face in their vocation via interviews and surveys (Rodriguez, Spilski, Hekele, Beese, & Lachmann, 2020). Construction work involves heavy physical labour as well as cognitive demands, and those demands need to be adequately represented and addressed in the development of suitable educational material (Kim et al., 2020). Study 3 takes a different approach to investigate the vocational education of car mechanics. Instead of interviews however, the aim of study 3 was the comparison of different presentation modalities for educational material (Hekele, Spilski, Bender, & Lachmann, 2022). As one of two virtual reality papers included in the present work it utilized instructional videos as a method to convey knowledge in either a 2D or 3D space with the goal to investigate the potential added value of virtual reality to video-based education.

The last study which is part of this work was fully based within a virtual environment. In study 4 the focus was put on the effect of audio cues and noise in virtual reality and its suspected impact on correlates of task performance (Meghanathan, Ruediger-Flore, Hekele, Spilski, Ebert, & Lachmann, 2021). The underlying goal was related to study 1 and 3, but instead of comparing different input or presentation modalities, study 4 investigated differences in task performance based on the amount of distraction within the virtual environment. The control over the visual and auditive input which VR-based technology can provide is seen as a major advantage of the technology in both experimental (Blascovich et al., 2002) and educational settings (Dede, 2009), but in future VR applications such as collaborative learning environments (Bujdosó, Novac, & Szimkovics, 2017) noise and other distractors are predicted to play a similarly major role to noisy learning environments in the “real” world (Ruotolo et al., 2013). After each paper is discussed individually, a larger picture of the overall research will be drawn and discussed in conjunction with an outlook for future prospects of VR applications in educational research.

Chapter 2: Literature Review

2.1. Demands in learning and work environments

2.1.1 Task-specific demands and cognitive load

Since the early days of the 20th century, researchers have been interested into factors which determine learning outcomes as well as factors which influence individual task performance (Taylor, 1911). Research into working memory and by extension cognitive load as a road to predict task performance has been a focal point of as a route to understand human behaviour especially in a learning context for decades (Sweller, 1988). In more recent times it also seems to have taken a centre stage in behavioural research, with over 10000 papers published within the last five years (Retrieved August 7, 2023, from https://www.scopus.com/). The causes for this strong focus are diverse but can at least partially be attributed in the scientific endeavour to better understand what type of stimuli are impacting human performance and then designing e.g. learning material which is better suited for humans (Buchner, Buntins, & Kerres, 2022; Skulmowski & Xu, 2022).

Buchner and colleagues (2022) gave a recent overview over the field of augmented reality research and concluded that the field suffers from several shortcomings. While there has been an increase in comparative research which compared the impact of different presentation media on learning outcome, the research is often focussed on the question whether AR works conceptually and therefore can more than not not applied to educational settings or replicated in general (Buchner, Buntins, & Kerres, 2022). For the field of virtual reality research in education, a review by Kavanagh and colleagues (2017) provided additional insights. They found similar issues for VR research and noted that many studies failed to build on pedagogical principles or by taking the needs of the learners into adequate account. Another shortcoming was the assumption of many reviewed articles that the mere inclusion of VR would increase learner motivation without additional work to ensure a better fit between the technology and the learner or the learning material (Kavanagh, Luxton, Reilly, Wuensche, & Plimmer, 2017). Without considering the needs of the many, the mere inclusion of additional technology doesn't have the impact on learning which people have come to expect from new technology (Ichou, 2018; Buchner, Buntins, & Kerres, 2022). One way to investigate the effectiveness of educational material is investigating the task-specific demands in work environments (Fleishman & Reilly, 1995) or by measuring and understanding the mental workload on the individual learner (Sweller, 2020).

In educational research, mental workload or cognitive load is traditionally measured either through behavioural measures such as task accuracy or error rates (Sweller et al., 2011) or a variety of physiological measures, most prominently eye-tracking for continuous pupillometry (Krejtz, Duchowski, Niedzielska, Biele, & Krejtz, 2018) or fixations on relevant learning objects (Korbach, Brünken, & Park, 2017). Additionally, questionnaires and interviews are often used to investigate the closely related concept of task complexity, such as the NASA Task Load Index (TLX, Hart & Steveland, 1988) or the Fleishman Job Analysis System (Fleishman & Reilly, 1995; Kleinmann, Manzey, Schumacher, & Fleishman 2010). The goal of estimating task complexity and the resulting load on the individual has been of major concern for the design of educational material. The advent of virtual and augmented reality and with it the creation of virtual 3D learning environments are seen as a major step forward in engaging students in novel ways to boost motivation and learning outcomes (Rosedale, 2017). Another potential for VR-based education materials is its ability to break down visual concepts in new ways which will enable educators to creative more accessible and inclusive learning materials for students (Martín-Gutiérrez, Mora, Añorbe-Díaz, & González-Marrero, 2017).

In the present work the term “cognitive load” will be used along related terms such as task-specific effort, mental effort and mental workload to reflect the wording used in the papers which the present works aims to integrate. Since the overarching goal of this work is the investigation of technology on learning in vocational education partially on learning, it will broadly define cognitive load in accordance with the work of Sweller (1988) and colleagues (2011). At the very basis, the term refers to the mental effort and resources required to process information during a particular task (Sweller, 1988). Furthering our understanding of cognitive load processes is crucial as it can affect task performance, problem-solving and thus learning outcomes in both experimental and real-world settings. The existence of a theoretical framework of human cognitive processes could guide the design of instructional materials and improve the ability to learn (Sweller, Ayres, & Kalyuga, 2011). In Sweller and colleagues’ (1998) cognitive load theory (CLT), the cognitive load of a task is viewed as a multi-faceted construct, out of which two will be introduced here. There is a source-specific component inherent to any task or material which is based on the complexity of the task’s information content as well as the interconnectivity of the information. This is referred to as intrinsic cognitive load and cannot easily be increased or decreased through external means such as different presentation modi (Sweller, 1988), but is dependent on the learner’s pre-

existing knowledge. Intrinsic load will be lower for experienced learners compared to novices (Kalyuga, 2005) since conceptually, the more organized knowledge or expertise a learner holds over a chunk of new information or the more straightforward a learning task is presented, the less intrinsic load they should experience (Kalyuga, 2011). Learning materials, or any form of information, can be presented in different ways which will affect the learner's performance – if the presented material is not explained properly, contains irrelevant information (Sweller, van Merriënboer, & Paas, 1998) or there are external distractors such as noise or other people speaking in the background (Choi, van Merriënboer, & Paas, 2014), the extraneous cognitive load increases. From a practical point of view extraneous load should be minimized or at least reduced as much as feasible to ensure a smooth learning experience (Kalyuga, 2011). The present work is primarily concerned with different forms of presentation to understand the demands work tasks and educational materials have on workers' and learners' respective task performance. Therefore, while the framework by Sweller and colleagues (1998) is used as foundation, study 1-4 will refer to either "workload", "demand", or "load" which each mostly correspond to the construct of extraneous load or more directly to physiological measures such as pupil size or attention-based fixation.

2.1.2. Task performance in the context of learning and working

From an early point in time, working memory research has shown that high cognitive load can result in lower task performance (see Baddeley & Hitch, 1974 for an overview). In Baddeley's original theory of mental workload (Baddeley, 1983) this reduction in performance is mostly attributed to attentional limitations in the working memory which can be caused by a variety of factors, such as stress (Stawski, Sliwinksy, & Smyth, 2006), different facets of individual differences such as learning preferences or motivation (Plass, Kalyuga, & Leutner, 2010) or task difficulty (see Duchowski et al., 2018; 2020). The presence of multiple tasks (Croizet et al., 2004) or high perceptual load (Greene, Maloney-Derham, & Mulligan, 2020) can also result in lower performance when limitations of cognitive resources are reached. It should however be noted that a low task-intrinsic load does not equal a low task difficulty by itself, for example a simple mathematical task in a language the participant is not fully fluent in, while conceptually simple, will still lead to high cognitive load (Sweller, Ayres, & Kalyuga, 2011). The last point is important to keep in mind, especially for the presentation and creation of educational content and

tasks – while the learner's attention might be captured by presenting information in novel ways such as displaying it in virtual reality, the content might get more difficult to process and thus lead to a higher (extraneous) workload and potentially worse learning outcomes (Croizet et al., 2004). Reversely, it also has been assumed that extraneous load can be reduced and learning outcome improved when task-relevant information is gathered from collaborators instead of other sources (Kirschner et al., 2018). Virtual reality environments could allow researchers to control for differences in the lab environment by maintaining comparable experiment conditions independent from the location or research group, which should foster the establishment of standardized, beneficial learning environments (Armougum, Orriols, Gaston-Bellegarde, Joie-La Marle, & Piolino, 2019). In recent reviews, educational research using AR (Buchner, Buntins, & Kerres, 2022) or VR (Kavanagh, Luxton, Reilly, Wuensche, & Plimmer, 2017) currently falls short of this expectation for standardisation, but progress in more specialized fields is made. Ferdous and colleagues (2019) reported significantly higher learning outcomes in physiotherapy students who used augmented reality in conjunction with a learning strategy, compared to students who only learned using AR objects. In the same paper, the authors also conducted a comparative usability analysis between traditional teaching and a tablet-based visualisation learning intervention. This analysis revealed that, while participants preferred the tablet condition and believed it had a positive impact on their learning outcomes, it rather increased their cognitive load (Ferdous et al., 2019). The authors assume that this increased load might be due to an unexpected usability issue while using the tablet. This also highlights one of the difficulties many studies utilizing emerging technologies such as VR and AR face – due to their still "emerging" status there are no widely accepted standard formats yet and educators and researchers often have to rely on custom software. This problem spans across most sections of educational technology and was for example also identified for virtual reality by Jensen and Konradsen (2018), which will be discussed in more detail in chapter 3 and 5.

Kirschner and colleagues (2018) bring a group aspect into cognitive load theory to bridge the gap from individual learning to collaborative learning. While learning traditionally includes several actors, such as an instructor or teacher and a learner, the authors pose that there is a necessary differentiation between non-collaborative instruction and collaborative learning. Due to the resulting interdependence (Tomasello & Gonzalez-Cabrera, 2017) between the learners the influence of the instructor but also instructional environment becomes more

important. Research has highlighted some advantages, but also shortcomings in computer-based learning environments, such as difficulties to communicate with other learners (Kirschner et al., 2018) and lack of realism and limited interactivity (Zizza et al., 2018). Due to technological advances in virtual reality, fully virtual environments have become feasible also for educational uses. These environments allow for extensive customization and manipulation of both the world features as well as the avatar which represents the learner while they are immersed in virtual reality (Blascovich et al., 2002; Zizza et al., 2018). A big advantage of virtual environment is their customization which enables educators to better suit individual learners' needs and ensure a good fit between learner and environment which in turn could improve learning outcomes. Newer models of virtual reality-based learning such as the cognitive affective model of immersive learning (CAMIL) by Makransky and Peterson (2021) predict that higher levels of immersion, such as those commonly found in immersive virtual reality, should lead to improved learning outcomes compared to other forms of education.

2.2. Virtual Reality

2.2.1. Psychological Research in Virtual Reality

Virtual reality has been described as the next disruptive technology as it has both technological impact and increasingly widespread availability to transform cultural and social activities as well as education (Freina & Ott, 2015; Rosedale, 2017). The potential it holds for psychological research is also seen as very high as it enables educators and scientists to alter concepts by bringing them from "the abstract into the tangible" (p.14, Slater & Sanchez-Vives, 2016)

One of the earlier, yet most influential research papers would be the work of Blascovich and collegues (2002). Two decades ago, they laid out an overview and potential use cases for immersive virtual reality for psychological research, discussing the impact of immersive virtual reality for social interaction and experimental setups set in virtual environments. The authors proposed a paradigm to use immersive virtual reality as a tool to reduce the trade-off between experimental control and realism (Blascovich et al., 2002). This trade-off has long been seen as a problem in the investigation of human behaviour since lab studies tend to place human subjects in highly controlled and thus artificial scenarios – whereas virtual reality allows for the same or even higher experimental control but combines it with higher ecological validity. Ex sito studies, which take place in the

"real world" are increasingly common in some fields of study with the advances in mobile technology, but often come with their own sets of drawbacks such as less experimental control (Hessels et al., 2022) and reduced validity and thus difficulties in replication (Brown, Reeves, & Sherwood, 2011). Blascovich and colleagues (2002) saw virtual reality as a solution to this trade-off as it allows researchers to maintain experimental control by controlling all aspects of the virtual environment.

Simultaneously VR can allow more natural user behaviour (Piumsomboon, Lee, Lindeman, & Billinghurst, 2017) or can even go beyond what would be commonly accepted in reality (Kyriakou & Hermon, 2019). In the last five years, several additional sensors have been integrated into virtual reality HMDs, allowing for the simultaneous collection of eye-tracking data including fixation and pupillometry data (Clay, König, & König, 2019) as well as extensive movement data such as trajectories (Haar, Sundar, & Faisal, 2021). Furthermore, researchers have begun to combine VR setups with electroencephalography (Baceviciute, Terkildsen, & Makransky, 2021) and electromyography (Blana, Kyriacou, Lambrecht, & Chadwick, 2016) for both research and medical purposes. At the time of writing, most of these multimodal approaches rely on custom setups and are not commercially available. The exception is eye-tracking, which not only has a high relevance for scientific research as will be highlighted in the present work, but also became a crucial technological component for the rendering of virtual environments. Without going into too many details, accurate gaze-tracking allows developers to create virtual environments which utilizes foveated rendering: a technique where instead of fully processing the entire environment, only the part where users fixate on is rendered, similar to how the brain processes the human visual field (see Patney et al., 2016 for a technical deep dive). For the research use case, foveated rendering reduces the cost of using virtual reality HMDs, since less powerful computers are necessary to run a virtual environment on an HMD with built-in eye-tracking. This was utilized in some explorative studies in the author's workgroup, and implications for further research will be discussed at a later stage.

Since the first forays into the potential of virtual reality, a variety of research has tried to define scenarios and experimental setups for the study of human behaviour. There are a few fields of VR research in psychology and cognitive science which will be briefly discussed here: The first methodology which is often used are comparative studies between virtual reality and other presentation forms such as videos or real-world interventions. This is commonly seen in educational research as virtual reality (and augmented reality) has been established as an effective

educational method which can enhance individual learning experience and increase learning outcomes compared to traditional learning methods (Dhimolea, Kaplan-Rakowski, & Lin, 2022; Esteves, Cardoso, & Gonçalves, 2023). Bahari (2021) identified increased immersion, higher engagement as well as the ability to create collaborative learning environments as key features for educational virtual reality research. In a L2 language acquisition paradigm, Legault and colleagues (2019) reported that virtual learning environments facilitated especially weaker L2 learners, resulting in higher learning outcomes compared to traditional word-pair learning. Zhang and Sternad (2021) compared learning in virtual reality and the real world with a physical throwing task and noted that participants' performance improved over time in both training conditions. However, performance in virtual reality started at a lower level, possibly due to the unfamiliarity of throwing an object in virtual reality but improved over the course of three training days to comparable success rates and degrees of errors to the real-world training group. The authors suspect that increasing the pre-task instructions in the virtual training condition could reduce this gap potentially (Zhang & Sternad, 2021), this would however decrease the comparability of the two conditions. Other literature suggests that differences between training modalities could decrease with a longer training period (see e.g. Hasson, Zhang, Abe, & Sternad, 2016) which might not be as practical for experimental research but yields promise for educational purposes. In other educational fields, the evidence about the effectiveness of virtual reality-based trainings and learning interventions is inconclusive with some researchers finding comparable effects or no difference between VR-based and traditional learning (Jensen & Konradsen, 2018; Moro et al., 2021) while others reported improvements in learning outcomes with virtual reality over traditional learning in the automotive sector (Chen, Luo, Fang, & Shieh, 2018). On the other hand, several reviews also see major disadvantages for VR applications in education. According to Buchner and colleagues (2022) the concerns with the use of augmented and virtual reality in the context of education are two-fold: From a usability perspective, the scalability is still highly problematic and most environments are created for very specific use cases, a conclusion also shared by Jensen and Konradsen (2018). The other concern lies in fear of cognitive overload of the participants if they are not familiar with the technology, which has also been reported by Albus and colleagues (2021). The research included in the present work has therefore taken as much care as possible to minimize the strain on the participant.

2.2.2. Presence and Immersion

Modern virtual reality setups utilizing head-mounted displays (HMD) are often referred to as “Immersive virtual reality” (IVR), to distinguish the concept more clearly from other virtual reality constellations such as Augmented reality (AR) and mixed reality (MR/XR) which is closer to the real world in the reality-virtuality continuum (Milgram, Takemura, Utsumi, & Kishino, 1995). Milgram and colleagues (1995) viewed the real world and virtual reality not as opposites, but as a continuum with AR sitting in between, as it enriches reality with additional information. Enriching the real world with additional information, overlays or other visual or auditive inputs is often also referred to as mixed reality, with augmented reality being seen as a subcategory of the same (Hönig, Milanes, Scaria, Phan, Bolas, & Ayanian, 2015). In contrast, virtual reality features an artificially generated, interactable environment which may or may not share the same characters as the real world (Milgram, Takemura, Utsumi, & Kishino, 1995). In a recent revisitation of this reality-virtuality continuum Skarbez and colleagues (2021) disagreed with the previously established definition and argued that, at the current technological level, virtual reality systems should be seen as mixed reality as the presenting device is still situated in the real world. In their definition, virtual environments are still real environments with virtual objects embedded in them (Skarbez, Smith, & Whitton, 2021), since the wearer is still aware of their surroundings to a certain amount. Conceptually participants are, at all times, both in the virtual world *and* the real world. Even if users feel completely immersed in a virtual world, they are still physically located in “our” shared reality. Mitzner and colleagues (2021) highlighted the possibility of conflicts between the real and virtual world, which could reduce presence if certain events were to happen such as a loud noise from “outside” the virtual environment. This does however not necessarily mean that experimental manipulations will not work, as showcased by clinical research into exposure therapy, where Ling and colleagues (2014) conclude in their meta-analysis that presence is an essential predictor for treatment success with an overall medium effect size across 33 studies.

While these definitions are more concerned with technical aspects, it is also important to consider the user’s perspective – do they perceive the real world or are their senses focussed on the virtual world around them? The field of virtual reality research introduced two concepts which can be used to evaluate the quality of VR environments - Immersion and presence. Immersion is a widely used term, so widely in fact that VR environments are often described as “immersive virtual

reality" (IVR; see Slater & Wilbur, 1997). Despite this common use, immersion proves harder to measure scientifically since there is an ongoing discussion about a commonly accepted definition of the construct (Slater & Wilbur, 1997; Grabarczyk & Pokropski, 2016). In general, immersive virtual reality experiences are defined by coherent mapping and design of the environment, feedback mechanisms to user actions as well as the rather diffuse notion of being part ("immersed in") of the VR environment (for a deeper discussion, see Grabarczyk, & Pokropski, 2016). Irrespective of a clear definition, we believe the notion of immersion is key in a vocational education context, since learners interacting with the environment should later on translate their virtually acquired skills into the "real" world (Kahlert, van de Camp, & Stiefelhagen, 2015). Presence on the other hand is commonly defined as the feeling of "being there" (Slater, Usoh, & Steed, 1994), indicating a subjective impression of being in the virtual world; whereas immersion describes the technology-related aspects which contribute to a particular VR environment, such as the resolution and vividness of visualization, sensory experiences and plausibility (Schubert, Friedman, & Regenbrecht, 2001). Presence can be measured using different types of instruments. A direct method of measurement is to ask participants to assess presented stimuli, scenarios or systems in hindsight, either qualitatively-verbally (e.g. structured interviews) or by means of questionnaires such as the iGroup presence questionnaire (IPQ; Schubert, Friedman, & Regenbrecht, 2001) or the Immersive Experience Questionnaire (IEQ; Rigby, Brumby, Gould, & Cox, 2019). Other, more indirect methods include behavioural measures (e.g. reaction times or task performance; see Agrewal, Simon, Bech, Baerentsen, & Forchammer, 2020) as well as physiological measures such as eye movement patterns (Wissmath, Stricker, Weibel, Siegenthaler, & Mast, 2010). A recent study by Hammond and colleagues (2023) tested levels of immersion using both self-report and behavioural measures and found that participants who reported higher levels of immersion had on average slower reaction times. In the present work, presence and immersion will be used not interchangeably but usually alongside one another.

Chapter 3: Task Performance and Demands Research

In this chapter, the research investigating primarily task performance will be described and discussed. Both of the main studies have been published as papers in 2019 and 2020, respectively. Study 1 aimed to investigate whether different stimuli assessment modes of an established behavioural measure in form of the trail making test are leading to comparable results.

Study 2 tried to assess demands on individual cognitive load via interviews and surveys in the context of modern work environments in three German companies and used the aggregated data to formulate suggestions for the design of construction work in the future.

Afterwards, related early-stage or otherwise unpublished research into mental workload and task performance from our work group will be described in some detail.

3.1. Relevance of the assessment mode in the digital assessment of processing speed (Study 1)

3.1.1. Introduction

Technological innovations and the increasing availability of online services are changing the world. This includes epidemiologic and psychometric assessments. Many medical facilities and research institutes started using digital assessment methods to assess cognitive abilities, such as processing speed, instead of the traditional paper and pencil versions. However, digital assessment methods may not necessarily lead to the same performance scores as the traditional paper and pencil assessments. Previously established norms may not be applicable and validity and reliability may differ.

The visual input of a digital display requires more perceptual and executive cognitive resources compared to paper, which strain working memory resources and can alter response accuracy (Noyes & Garland, 2008). Reading speed on digital displays is also slower and reading accuracy is sensitive to the visual context (Noyes & Garland, 2008). Even though studies and meta-analysis report strong correlations between paper and pencil and computer versions of psychometric assessments (Gwaltney, Shields, & Shiffman, 2008; Muehlhausen et al., 2015; Riva, Teruzzi, & Anolli, 2003; van Ballegooijen, Riper, Cuijpers, van Oppen, & Smit, 2016), some instruments show low inter-format reliability (Alfonsson, Maathz,

Hursti, & Eysenbach, 2014). For most of the digitized adaptations, there are no psychometric norms reported and concurrent validity with the original paper and pencil version is not established (Gates & Kochan, 2015; Zygouris & Tsolaki, 2014). Vora and colleagues (2016) investigated differences between computerized neurological testing and the respective paper and pencil tests in young adults and came to the conclusion that the computerized neurological tests are valid but research teams do not compare the performance scores to the norms of the paper and pencil version (Vora, Varghese, Weisenbach, & Bhatt, 2016). Hence, the comparability of performance score across modes is often unclear. Assessment mode may considerably alter performance scores in tests that aim to assess processing speed. Processing speed is an important clinical marker, for instance, for brain infarcts and generalized brain atrophy (Prins et al., 2005). Processing speed is also used as an indicator for disease severity in multiple sclerosis (Demaree, DeLuca, Gaudino, & Diamond, 1999), reading disability and Attention Deficit/Hyperactivity Disorder (Shanahan et al., 2006). Yet, assessing processing speed may be confounded by the assessment mode. Instruction wording and format, stimuli presentation, and response methods can cause differences between responses (Zygouris & Tsolaki, 2014). Digital assessments, for instance, may come with an increased perceptual load (Carpenter & Alloway, 2018). Bando, Asano, and Nozawa (2017) observed that reading from digital devices activates the parasympathetic nervous system more than reading from paper (Bando et al., 2017). Responses on digital versus paper versions differ also with regard to other factors such as individual response strategies (i.e., gaming effects), user expectations, use of one or more fingers, feedback on errors, and physical challenges (Jenkins, Lindsay, Eslambolchilar, Thornton, & Tales, 2016). Therefore, it is necessary to re-evaluate psychometric properties when developing a new, digital version of an assessment (Meade, Michels, & Lautenschlager, 2007).

An instrument that is widely used to assess processing speed is the Trail Making Test (TMT), in particular, the version in which participants have to connect numbers in ascending order as fast as possible (Misdraji & Gass, 2010). The TMT was originally developed for the Army Individual Test of General Ability to assess intelligence (Tombaugh, 2004) and was later adopted to be used in clinical settings (Reitan, 1958). Given its usability as a clinical screening tool (Martin, Hoffman, & Donders, 2003), it is important to ensure the validity of the TMT. The TMT, in which participants connect numbers in ascending order, requires mainly visuoperceptual abilities (Sanchez-Cubillo et al., 2009). Therefore, it may be sensitive to simple

aspects of the assessment mode. Previous studies tested digital versions of the TMT. In 1995, Salthouse and Fristoe developed a digital version of the TMT in which participants used the arrow keys on the keyboard to move the cursor. They observed that a longer completion time in the digital version that was attributable to the greater number of keystrokes (Salthouse & Fristoe, 1995). A study by Fellows, Dahmen, Cook, and Schmitter-Edgecombe (2016), on the other hand, reported moderate correlations between a tablet-based TMT with a pen and a paper and pencil version. The input method may play an important role, as another study that used the mouse as input device did not observe significantly different performance scores (Drapeau, Bastien- Toniazzo, Rous, & Carlier, 2007). These observations suggest that motor components of the digital TMT – such as the use of the finger or of a pen – may influence performance scores. Moreover, the different versions of the TMT differed in their design. For instance, the tablet-based TMT contained 20 number-circles whereas their paper and pencil version of the TMT contained 26 number-circles (Fellows et al., 2016). Differences in the design of those studies do not allow us to draw any conclusions on assessment mode effects. For that reason, it is essential to investigate to what extent the assessment mode affects validity of the TMT, especially when there are alterations in the design such as the visual stimuli or the handling of errors. If performance differs systematically between assessment modes, then the digital version needs its own psychometrics and validity testing.

Aim of our study was, therefore, to test for assessment mode effects within subjects while keeping the design of the TMT the same. We investigated whether results for assessing processing speed via (i) the paper and pencil version of the TMT, (ii) a tablet and pen version of the TMT, and (iii) a tablet and finger version of the TMT are comparable. For this purpose, we used the TMT developed by Oswald that contains the numbers 1 to 90 in a systematic 9 × 10 grid (Oswald & Roth, 1987) because it had four different versions available in which the respective numbers are at different locations and could thus systematically be tested on the tablet. This version differs from the version in the Army Individual Test Battery from 1944 in which 25 numbers are randomly distributed over the paper (Tombaugh, 2004) but is comparable to a version that Salthouse and his colleagues used with 49 numbers in a 7 × 7 grid (Salthouse et al., 2000). Our study investigated how performance in the three different assessment modes of the 90 number version of the TMT differed within-subjects by (a) having participants complete four versions

in each assessment mode and (b) counterbalancing the order of the assessment mode. In this way, the study also presents findings on the order of the assessment modes, learning effects over time, and the role of subjective preferences.

3.1.2. Methods

3.1.2.1. Study design

The study was approved by the ethics committee of the Department of Social Science of the University of Kaiserslautern and was carried out in conformity with the principles embodied in the Declaration of Helsinki. The study employed a 3 × 4 within-subject design that included three different assessment modes with four TMT versions each. A total of 30 students (73% male) from the University of Kaiserslautern participated in this study. Information about the study was communicated verbally in meetings, courses, and occasional encounters on the university campus. If someone was interested in participating, we gave him or her a date and time. All participants had a high level of education, as all of them were actively enrolled at the university. Of the participants, n = 22 were male and 8 were female. A total of 10 participants were 22–25 years old, 14 participants were 26–29 years old, 4 participants were 30–34 years old, and 2 participants were 34 years or older. All but one participant reported to be familiar with tablets.

3.1.2.2. Trail making test (TMT)

We used the TMT by Oswald and Roth (1987), which was developed as an alternative method for assessing intelligence assuming that perceptual information processing is a good indicator of intelligence (Oswald & Roth, 1987). This TMT contains 90 numbers that are systematically ordered in horizontal and vertical lines (9 × 10 grid, dimensions 19.5 × 22 cm). The task is to connect the 90 numbers as fast as possible in ascending order. The next number to be connected is always one of the eight adjacent numbers. There are four different versions (a, b, c, d) of this TMT. The difference between versions a, b, c, and d is the order of the numbers. The tablet version of the TMT was developed using HTML and JavaScript as a web-based application. It is therefore platform-independent and can be employed on every browser-enabled device. In our study, the application was running on an 8-in. screen of a Samsung Galaxy A touchscreen tablet. The tablet and the paper and pencil version of the TMT were scaled to 13 × 13 cm (that is almost 50% of original size) for them to be exactly the same size. The screen of the

tablet contained 90 empty circles that were filled with numbers corresponding to one of the four versions as soon as the participants pressed “Start”. A timer started from exactly that moment. To guarantee the comparability of the tablet version and the paper version, the participants always had to press a “stop” button as soon as they finished the task. A screenshot of the completed TMT was either saved on the tablet or collected by the investigator. We prepared three different setups each using versions a, b, c, and d: (i) the traditional paper and pencil version, (ii) a tablet and pencil version, and (iii) a tablet and finger version (3 × 4 design). Accordingly, each of the 30 participants completed the TMT 12 times in counterbalanced order.

3.1.2.3. Procedure

Participants who were interested in taking part in our experiment were invited to a testing session at our institute. Each session started with information about the study and the opportunity to ask questions. Participants then received instructions about the task. The instructions included the information to correct any error immediately by returning to the last correctly connected number. During the study, participants were sitting at a desk with a tablet lying flat on the table’s surface. A regular pen and a special pen for the tablet were placed next to the tablet. The participants were informed beforehand about which setup they had to complete. For the paper version, the TMT was placed in front of the participant and the participant immediately pressed “Start” on the tablet timer and began completing the TMT. For both tablet versions, the participants had to press “Start” to start the timer and to begin completing the TMT. The time that the tablet screen took to refresh after pressing the “Start” button was comparable to the time that the participants needed to move their hand from the tablet after pressing the “Start” button to the paper. For every version, the participant had to press “Stop” as soon as he or she finished the task. The tablet-based TMT picture was then saved on the tablet and the paper TMT picture was handed to the investigator. This process was repeated 12 times in a predetermined, counterbalanced order of assessment mode and TMT version. After the last TMT, participants were asked to fill out a short questionnaire on paper that included questions on age, gender, familiarity with tablets, perceived ease/control/comfort (from 1 = low to 5 = high) in completing the different setups, and setup preference. Following the testing sessions, the investigators counted the errors on the TMT pictures. When a participant skipped a number or connected the wrong number without a visible trace of him or her going back to the correct number, then the investigator marked an error. As the study

investigated within-subject effects, the order in which the participants completed the three different setups of the TMT – (i) paper and pencil version, (ii) tablet and pen version, (iii) tablet and finger version – was counterbalanced between participants. To guarantee that throughout the entire study, each setup was completed as first, second, or third by the same number of participants, we predetermined the following order: 10 participants doing (i), (iii), (ii); 10 participants doing (iii), (ii), (i); and 10 participants doing (ii), (i), (iii). Each participant completed each setup of the TMT four times (e.g., four times (i), four times (iii), and four times (ii)). The reason for this procedure was to make sure that every participant completed all of the TMT versions a, b, c, and d for each setup. We counterbalanced the order of these versions between the setups and between the participants.
In this way, we can test whether the assessment mode effects are stable across different versions.

3.1.2.4. Statistical analysis

Statistical analyses were performed using STATA 15 and employed an alpha level for statistical significance of .05 (two-tailed). Power was estimated retrospectively using means via the Stata command “power repeated” specifying between- and within-effect calculation for 30 participants over three conditions that are each
repeated each four times. Results indicated a power of 1.000 when alpha is 0.05. TMT completion times were normally distributed and there were no missings. With respect to errors, we had three missings out of 357 TMT tests because the screenshots of the respective TMTs did not allow us to draw firm conclusions on the number of errors that were made. Comparisons between assessment modes and the different versions of the TMT (a, b, c, and d) were carried out using Analyses of Variance (ANOVA). As errors were not normally distributed, we repeated the analysis using Pearson’s Chi-square test. To test for the influence of age, gender, setup order, number of errors, and number of uncorrected errors on performance in the TMT, linear regression analyses were used. Within-subject effects of assessment mode were analyzed using repeated measures ANOVA as well as mixed-model analyses. We chose to calculate mixed-models because it allows us to model performance in the TMT in the different assessment modes over time. The mixed-models included, in Model 1, fixed effects for a trial number for each participant and random effects for age and gender and, in Model 2, in addition, nested effects

for assessment mode. Random effects for age and gender were included to adjust for variances in performance caused by age and gender. For errors, we used mixed-effects ordered logistic regression analysis as errors were not normally distributed. In the last step, we analyzed the responses on the questionnaires by calculating means. We created a binary variable (yes/no) representing whether the trial was completed in the preferred assessment mode or not. In this way, we were able to analyze differences in performance in the TMT with respect to preference for a specific assessment mode (using ANOVA).

3.1.3. Results

3.1.3.1. Average completion times

The average time to complete the TMT was 62.59s (*SD* = 9.14s). The completion time did not differ significantly between the different versions a, b, c, and d of the TMT (ANOVA $F(3,356) = 0.64$, p = .592, see also Figure 1). However, participants completed the TMT significantly faster in the tablet and pen version (*M* = 59.07s, *SD* = 10.88s) than in the paper and pencil (*M* = 64.16s, *SD* = 13.10s) or the tablet and finger version (*M* = 64.57s, *SD* = 13.87s; ANOVA $F(2,357) = 7.01$, $p = .001$, Bonferroni comparison: Tablet and pen vs. Paper and pencil −5.092, $p = .006$, Tablet and pen vs. Tablet and finger 5.500, $p = .003$, Paper and pencil vs. Tablet and finger 0.408, $p = 1.000$).

3.1.3.2. Average number of errors

The overall mean error was 0.62 (*SD* = 0.58). Out of the 357 valid trials, 215 trials (60.22%) did not contain any errors, 95 trials (26.61%) contained one error, 27 trials contained two errors (7.56%), 11 trials (3.08%) contained three errors, five trials (1.40%) contained four errors, and four trials (1.12%) contained five errors. The mean number of uncorrected errors (i.e., the participant did not go back to the last correct number) was 0.09 (*SD* = 0.15). Out of the 357 valid trials, 29 trials (8.12%) contained uncorrected errors of which 25 trials (7.00%) had one uncorrected error, three trials (0.84%) had two uncorrected errors, and one trial (0.28%) had four uncorrected errors. All the other errors reported before were self-corrected. Participants made a similar amount of errors across assessment modes (ANOVA $F(2, 354) = 1.31$, $p = .270$; Pearson's $\chi^2 = 15.47$, $p = .116$) and TMT versions a, b, c, and d (ANOVA $F(3,353) = 0.08$, $p = .972$; Pearson's $\chi^2 = 20.26$, $p = .162$).

3.1.3.3. Order of assessments on completion times and errors

The order of the assessment modes did not lead to significant different completion times (ANOVA $F(2, 357) = 2.56$, $p = .079$) but possibly to significantly different
errors (ANOVA $F(2, 354) = 5.55$, $p = .004$; Pearson's $\chi^2 = 17.31$, $p = 0.068$). Descriptive inspection of the data suggested that if participants completed the tablet and finger version first, they had more errors on average ($M = 0.80$, $SD = 1.18$) than if they completed the paper and pencil version ($M = 0.67$, $SD = 0.99$) or the tablet and pen version ($M = 0.39$, $SD = 0.67$) first. Statistical analysis of the errors in the first trial only revealed no significance (ANOVA $F(2, 27) = 0.63$, $p = .539$; Pearson's $\chi^2 = 4.70$, $p = 0.319$).

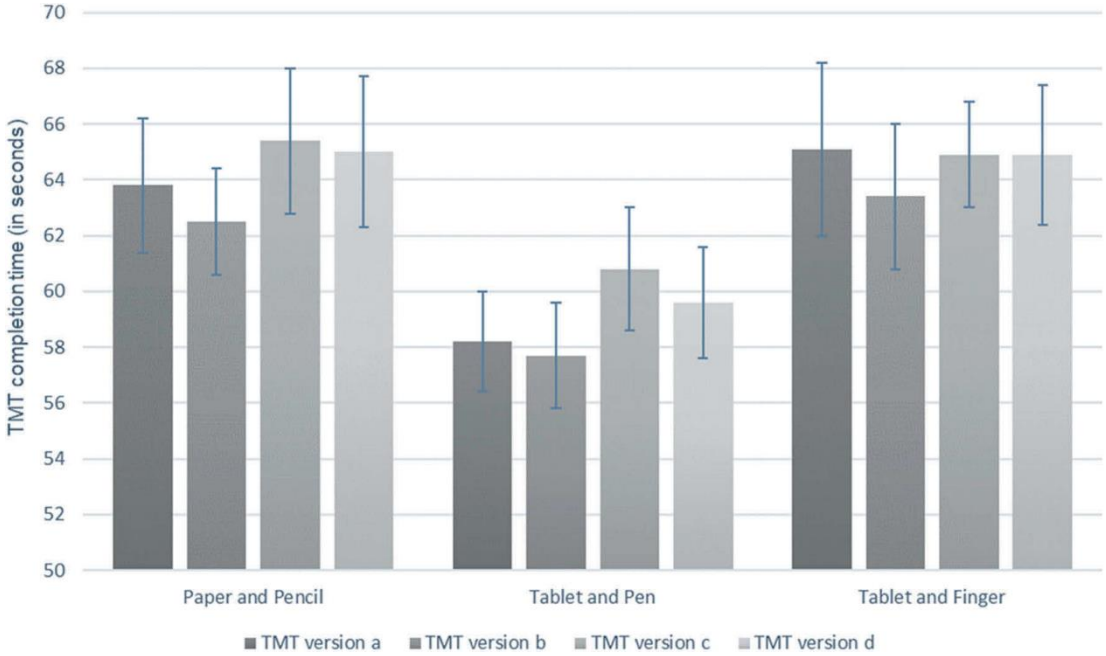

Figure 1: Mean completion time by assessment mode in the four different versions a, b, c, and d of the Trail-Making Test (TMT), separated by the three assessment modes paper and pencil version, tablet and pen version, and tablet and finger version (N = 30). Error bars reflect the standard error.

3.1.3.4. Multivariate influence on completion time

Age, gender, setup order, and errors on completion times

The influence of age, gender, setup order, number of errors, and number of uncorrected errors was analyzed using linear regression analyses. Results with respect to each person's average TMT completion time indicated no significant effects ($p > .20$, see Table 1). Linear regression analyses on each trial's completion time indicated that older age (e.g., participants in their 30s), fewer errors, and the assessment mode tablet and pen were associated with significantly faster per completion times (see Table 1). Whether somebody corrected the error did not significantly alter completion time.

Within- and between-subject effects of assessment mode on completion times via repeated measures ANOVA

Repeated measures ANOVA revealed significant between-subject effects as well as significant within-subject effects for the trial number indicating a learning effect (model 1, see Table 2). With each participant having completed the TMT 12 times, the trial number (from one to twelve) refers to the first, second, third, etc., time that the participant completed the TMT during the experiment. Including nested effects for the assessment mode in the model (model 2) revealed significant main effects for the assessment mode and the trial number, indicating that participants were able to complete the tablet and pen version faster and that the completion time varies between assessment modes depending on the trial number (see Table 2). Eta-squared was 0.889 (df 116) for the model, 0.064 (df 2) for the assessment mode, 0.558 (df 11) for the trial number, and 0.180 (df 18) for the interaction effect between assessment mode and trial number. Pairwise comparison of the estimates suggests the following contrasts: Tablet and pen vs. Paper and pencil −5.092, 95% *CI* [−6.419, −3.764], Tablet and pen vs. Tablet and finger 5.500, 95% *CI* [4.172, 6.828], Paper and pencil vs. Tablet and finger 0.408, 95% *CI* [−0.919, 1.736].

Independent variable on each person's average TMT completion time	*b*	*SE*	*t*	*p*
Independent variable on each person's average TMT completion time				
Age	−2.51	2.09	−1.20	.241
Gender (female)	2.09	4.29	0.49	.630
Order ((iii), (ii), (i) [a])	2.67	4.22	0.63	.540
Order ((ii), (i), (iii) [a])	2.70	4.45	0.61	.550
Mean errors	3.56	3.98	0.90	.380
Mean uncorrected errors	6.35	15.21	0.42	.680
Independent variable on each trial completion time				
Age	−2.67	0.78	−3.42	.001
Gender (female)	2.36	1.54	1.53	.126
Order ((iii), (ii), (i) [a])	3.08	1.62	1.90	.059
Order ((ii), (i), (iii) [a])	2.35	1.63	1.44	.151
Assessment mode (Ref: paper and pencil) tablet and pen	−4.88	1.60	−3.04	.003
Assessment mode (Ref: paper and pencil) tablet and finger	0.46	1.61	0.29	.773
Number of errors	1.65	0.82	2.00	.046
Number of uncorrected errors	1.97	2.14	0.92	.359

[a]Assessment mode order with (i), (iii), (ii) being the reference group; (i) = paper and pencil version; (ii) = tablet and pen version; (iii) = tablet and finger version; *b*, coefficient; *p*, level of significance; Ref., reference group; *SE*, standard error; *t*, t statistics from Analysis of variance test.

Table 1: Linear regression analyses for the impact of age, gender, order of assessment modes, errors, and uncorrected errors on each person's average Trail-Making Test (TMT) completion time and per trial completion time (N = 30).

Within- and between-subject effects of assessment mode on completion times via mixed-models

We analyzed the effect of assessment mode over the trials in more detail using mixed-effects models (with random effects for age and gender). Results indicated that TMT completion times were always significantly faster in the tablet and pen version than in the paper and pencil version (see Table 3). Learning effects disappeared after trial eight, and completion time in the tablet and finger version were not significantly different from the paper and pencil version (see Table 4). The predicted mean completion times from the mixed-effect models were *M* = 65.15s (*SE* = 1.33) in the paper and pencil version, *M* = 65.55 s (*SE* = 1.33) in the tablet and finger version, and *M* = 60.05s (*SE* = 1.33) in the tablet and pen version; confirming that participants completed the tablet and pen version about 5s faster than the other versions. Analyses with respect to the number of errors indicated significant between-subject effects but no effects for trial number or assessment mode in a repeated measures ANOVA (see Table 4) and mixed-effects ordered logistic regression analysis (results not shown).

Models	*df*	*F*	*p*
Model 1			
Participant	29	22.20	<.001
Trial number	11	32.51	<.001
Model 2			
Setup	2	2.91	<.001
Participant over setup[a]	85		
Trial number	11	27.88	<.001
Setup # trial number	18	2.97	<.001

#, interaction; [a]Between subject error terms; *df*, degrees of freedom; *F*, F statistics from the repeated measures ANOVA; *p*, level of significance. Model 1 includes fixed effects for trial number for each participant and random effects for age and gender and model two includes model 1 together with nested effects for assessment mode.

Table 2: Repeated measures analysis of variance (ANOVA) on the impact of trial number and assessment mode on Trail-Making Test (TMT) completion time (N = 30).

3.1.3.5. Subjective preferences of the assessment modes

The responses on the questionnaires indicated that 70.0% (*n* = 21) of the participants preferred the tablet and pen version, while three participants preferred the paper and pencil version, three participants the tablet and finger version, and three participants had no preference. Similarly, n = 20 participants (68.9%) felt that they had a better performance in the tablet and pen version than in the other versions. However, participants reported slightly higher levels of control in the traditional paper and pencil version (*M* = 4.30, *SD* = 0.88) compared to (*M* = 4.00, *SD* = 0.98) in the tablet and pen version and (*M* = 3.97, *SD* = 1.22) in the tablet and finger version (*p* > .05). Even though not significantly different, there was a trend of

participants reporting more comfort in the paper and pencil version (M = 4.03, SD = 0.85) compared to the tablet and pen version (M = 3.57, SD = 1.19) and the tablet and finger version (M = 3.90, SD = 1.21) and more ease of performance in the paper and pencil version (M = 4.30, SD = 0.84) compared to the tablet and pen version (M = 4.07, SD = 1.01) and the tablet and finger version (M = 4.00, = 1.02, p > .05). The TMT completion time was significantly faster in the assessment mode that the participant preferred (M = 58.81 s, SD = 13.75 vs. M = 64.22s, SD = 9.69, ANOVA $F(1, 358) = 13.75$, $p < .001$).

Fixed Effects [§]	*b*	*CI 95*	*z*	*p*
Assessment mode (Ref: paper/pencil)				
Tablet/pen	−20.55	−28.73; −12.73	−4.93	<.001
Tablet/finger	−3.77	−10.83; 3.30	−1.04	.296
Trial number	−2.77	−3.41; −2.13	−8.45	<.001
Trial number (Ref: 1)				
2	−15.13	−19.92; −10.34	−6.19	<.001
3	−16.16	−20.84; −11.48	−6.76	<.001
4	−15.59	−20.25; −10.93	−6.55	<.001
5	−17.07	−24.37; −9.79	−4.59	<.001
6	−16.61	−23.79; −9.42	−4.53	<.001
7	−11.64	−18.79; −4.48	−3.19	.001
8	−11.87	−19.04; −4.69	−3.24	.001
9	−4.61	−9.27; 0.05	−1.94	.053
10	−2.44	−7.12; 2.24	−1.02	.307
11	−1.45	−6.26; 3.32	−0.60	.548
12	omit.			
Assessment mode (Ref: paper/pencil) # trial number (Ref: 1)				
Tablet/pen # 2	13.2	6.16; 20.24	3.67	<.001
Tablet/pen # 3	11.7	4.66; 18.74	3.26	.001
Tablet/pen # 4	14.3	7.26; 21.34	3.98	<.001
Tablet/pen # 5	23.44	10.01; 36.87	3.42	.001
Tablet/pen # 6	22.44	9.01; 35.87	3.27	.001
Tablet/pen # 7	18.34	4.91; 31.77	2.68	.007
Tablet/pen # 8	18.84	5.4; 32.27	2.75	.006
Tablet/pen # 9	13.52	1.35; 25.69	2.18	.029
Tablet/pen # 10	14.02	1.85; 26.19	2.26	.024
Tablet/pen # 11	16.92	4.75; 29.09	2.73	.006
Tablet/pen # 12	18.82	6.65; 30.99	3.03	.002
Tablet/finger # 2	10.00	2.96; 17.04	2.78	.005
Tablet/finger # 3	1.60	−5.44; 8.64	0.45	.656
Tablet/finger # 4	5.10	−1.94; 12.14	1.42	.156
Tablet/finger # 5	13.31	1.51; 25.49	2.15	.032
Tablet/finger # 6	9.22	−2.95; 21.39	1.48	.138
Tablet/finger # 7	5.32	−6.85; 17.49	0.86	.392
Tablet/finger # 8	5.02	−7.15; 17.19	0.81	.419
Tablet/finger # 9	−2.82	−14.21; 8.56	−0.49	.627
Tablet/finger # 10	4.28	−7.11; 15.66	0.74	.461
Tablet/finger # 11	0.38	−11.01; 11.76	0.07	.948
Tablet/finger # 12	−1.32	−12.71; 10.06	−0.23	.820

§, random-effects for gender b = 111.84, age b = 12.87; *CI* 95, 95% confidence interval; omit., omitted because of collinearity, *b*, coefficient; *p*, level of significance; Ref, reference category; tablet/finger, tablet and finger version; paper/pencil, paper and pencil version; tablet/pen, tablet and pen version; *z*, z statistics from the repeated measures ANOVA.

Table 3: Fixed effects estimates of the mixed-effects model on the impact of assessment mode on Trail-making Test (TMT) completion times (N = 30).

3.1.4. Discussion

The study investigated whether results for assessing processing speed via (i) the traditional paper and pencil version of the TMT, (ii) a tablet and pen version of the TMT, and (iii) a tablet and finger version of the

Models	*df*	*F*	*p*
Model 1			
Participant	29	5.63	<.001
Trial number	11	0.79	.654
Model 2			
Setup	2	0.71	.493
Participant over setup [a]	85		
Trial number	11	0.84	.599
Setup # trial number	18	0.81	.694

[a]Between subject error terms; *df*, degrees of freedom; *F*, F statistics from the repeated measures ANOVA; *p*, level of significance. Model one includes fixed effects for trial number for each participant and random effects for age and gender and model two includes model 1 together with nested effects for assessment mode.

Table 4: Repeated measures analysis of variance (ANOVA) on the impact of trial number and assessment mode on errors in the TMT (N = 30).

TMT are comparable. The three versions were counterbalanced in a within-subject study design. Results indicate that performance in the tablet and pen version was significantly faster (by about 5 s) compared to the tablet and finger version and compared to the traditional paper and pencil version. Participants did not only perform best with the tablet and pen version but they also subjectively preferred this version. Only one previous study seems to have compared different digital modes of the TMT: Performance on the touch screen did not differ significantly from performance with the mouse (Canini et al., 2014). Accordingly, the superior performance that we observed in our study appears to be related to the use of the pen on the touch screen. It is surprising that the tablet and pen version resulted in faster completion of the TMT than either using the tablet and finger or the traditional paper version. With regard to the finger, this observation is unexpected because a study that compared shape tracing by pen and by finger on a tablet has shown that using the finger was fastest (Zabramski, 2011).Maybe the faster reaction times that Zabramski (2011) observed with the finger compared to the pen on the tablet have to do with the nature of the task, which was a shape-tracing task. As performance on the TMT is attributable to visual search and motor speed together (Crowe, 1998), using the finger may severely hinder the visual search. Holding the hand right in front of the display makes it difficult to see the next number that has to be connected and, hence, slows down the overall performance in the TMT. Moreover, Zambramski and Stuerzlinger (2012) have shown that the pen is the least and the finger is the most error-prone entry method, which could contribute to faster completion times. With regard to the traditional paper version, however, it is unclear how the tablet surface facilitates performance compared to completing the TMT on paper. Nonetheless, this is not a new finding: Gerth and colleagues investigated handwriting performance in children and observed faster writing velocity on tablets on all tasks (Gerth et al., 2016b). Their findings suggest that the pen is sliding more on the smoother surface of the tablet than on the paper (Gerth et al., 2016a). However, differences in the pen itself such as material and flexibility of the top as well as characteristics of the tablet such as pressure sensitivity and operating system may affect performance in the tablet and pen version to an extent that is not yet known.

Further, our results suggest that errors increased the completion time irrespective of whether the participants corrected the errors. Usually one would expect that the simple action of correcting an error would also slow completion time. It is possible that the awareness of having made a mistake affected the

participant and thus slowed down performance. Making errors are associated with an increase in skin conductance and a heart rate deceleration, which is correlated with a post-error slowing of performance (Hajcak,McDonald, & Simons, 2003). It seems as if the fact that the brain recognizes an error evokes a physical reaction that slows subsequent performance. The brain area dealing with error processing is the anterior cingulate cortex (ACC) and increased ACC activity has shown to influence reaction time in face of errors (Mulert, Gallinat, Dorn, Herrmann, & Winterer, 2003). Holroyd and Coles (2002) emphasize the relevance of error processing by the dorsal anterior cingulate in the early stages of learning (Holroyd & Coles, 2002). Error information processing can thus influence decision-making of further actions independently of whether the error was actually corrected. Our results suggest that this error processing then leads to TMT completion time that are similar whether the error was corrected or not.

Another important finding of our study is learning effects. We observed learning effects until trial eight, suggesting that the threshold for arriving at the best performance levels occur after the eighth trial. Previous studies already reported practice effects in variants of the Trail Making Test during serial assessment (Buck, Atkinson, & Ryan, 2008). Interestingly, we observed that learning effects were less pronounced in the tablet and finger version. Whether this has to do with the restrictions to the visual field is unclear. The lack of familiarity of using a finger to draw instead of a pen might have also hampered the learning process. Habits of using the wrist, hand, and fingers determine learning speed (Krakauer, Mazzoni, Ghazizadeh, Ravindran, & Shadmehr, 2006). Researchers and practitioners should keep that in mind when applying digital versions of psychometric tests because the input method (mouse, keyboard, pen, finger, laser pointer, etc.) of the digital version might influence the psychometric results.

Despite the counterbalanced within-subject design, our study is subject to some limitations. First, our sample comprised young, healthy participants. We cannot draw any conclusions on clinical populations and older individuals. Second, we reduced the size of the paper version of the TMT to make it the same size of the digital version. Hence, we were not able to compare the originally sized version with the digital version. Reducing the size of the TMT also reduced the size of the circle around the numbers, which may then have contributed to more errors and slower completion times in the tablet and finger version. Third, we do not know whether other factors such as light or physical strength may have affected our results.

Fourth, we are not sure to what extent the fact that the participants had to press the “Start” button might have affected the cognitive processes related to completing either version of the test.

Digitalizing psychometric assessment tools, also referred to as psychoinformatics, is a new, very useful trend in the medical field, cognitive science, and psychology. However, to ensure that data are comparable to traditional assessment methods, the administration of digital tests has to be validated, and standardized psychometric data have to be obtained (Gates & Kochan, 2015). Currently, there is not sufficient data on validity available (Zygouris & Tsolaki, 2014). The results of our study point out how trivial it is to test even small alterations in the design like the use of a pen instead of the finger. Other factors, such as the operating system level, are also crucial (Montag, Duke, & Markowetz, 2016) and more research are necessary to identify critical aspect of digital design in psychoinformatics. This is important for digital tests are to be used in the clinical setting. Moreover, the clinical use of new digital tests should also encompass the development of their own set of norms.

3.2. Physical and cognitive demands of work in building construction (Study 2)

3.2.1. Introduction

3.2.1.1. Context of the Study

Construction work is an essential component of our society, providing us with houses and apartments, school buildings and hospitals, roads, airports and other important infrastructure. In that respect, the construction industry makes up an ever-increasing amount of the gross domestic product (GDP). In Germany, the contribution of the construction sector to the GDP increased from €26bn in 1991 to almost €44bn in 2018 (Trading Economics, 2019). In the USA, the amount increased from US$440bn in 2002 to US$740bn in 2017 (Federal Reserve Bank of St Louis, 2018). With a current total return of €125bn in Germany (Statista, 2018) and a total revenue of US$129bn in the USA (Statista, 2019), the growth of the total factor productivity in the construction industry worldwide has been extremely slow (Abdel-Wahab and Vogl, 2011). One factor that contributes to a limited productivity growth is that workers can face only a certain quantity of demands in a single workday.

Workers in the construction industry face high physical demands at work (Eaves et al., 2016). Apart from physical demands, there are five other types of demands at work such as the cognitive, psychomotor, sensory/perceptual and social/interpersonal demands (Fleishman and Mumford, 1991). However, less research has been done on them so that it is difficult to identify any information on the extent and the causes of these types of demands (see Section 1.3 for more details). Nonetheless, it is important to know about these types of demands in construction work because, in consequence, they can affect workers' health. The health of construction workers is sensitive to their work context. A study of 4,958 German construction workers used data from occupational health examinations in 1986 and 1992 and observed a threefold increase in disability over this period (Arndt et al., 1996). In fact, construction workers of all age groups exhibit symptoms of musculoskeletal health conditions and the length of time in the job ultimately increases the risk for such symptoms (Eaves et al., 2016). Compared to white-collar workers, construction workers are more likely to have hearing deficiencies, obstructive lung diseases, increased body mass index and musculoskeletal abnormalities (Arndt et al., 1996). Already 54 percent of the young apprentice construction workers report job-related health symptoms; a longer time in the construction industry was then additionally associated with knee and wrist

symptoms (Merlino et al., 2003). Health problems can be triggered by physical demands in the construction industry, as a number of studies have observed, but they can also be trigged by cognitive demands.

Globally, the most debatable issue with regard to work demands is that besides physical demands other types of demands (e.g. psychosocial demands) are relevant for health and wellbeing. Evidence from office workers has shown that high demands such as time pressure and high workload, especially in the context of low control (job demand-control theory (Karasek et al., 1981)) or low resources (job demands-resources model (Bakker et al., 2010)), affect health in a negative way. These types of studies also revealed that high cognitive and psychosocial demands at work come with a greater risk for mental health problems (Then et al., 2014; Seidler et al., 2014), such as symptoms of depression and burnout (Hakanen et al., 2008; Nahrgang et al., 2011). Increased demands sometimes also lead to physical symptoms without that the person actually having an underlying disease (somatic symptoms; Nomura et al., 2007). We are not able to identify any evidence with regard to construction workers. Hence, at this point, it is not possible to draw any conclusion on these demands compared to other industries. To estimate the actual risk among construction workers, we first need to gain a better understanding of the types of demands that construction workers are faced with on a daily basis.

Learning more about the level of cognitive demands in construction work is important for another reason. New digital technologies in construction work such as building information modelling, digital time schedules, digitization of services, workflows and construction planning promise technical and economic benefits for stakeholders in the construction industry (Rüßmann et al., 2015) but comes with completely new demands for construction workers. Technical requirements, building regulations and administrative procedures are frequently extended and updated (Visscher and Meijer, 2007), which ultimately requires workers to constantly develop new skills to meet the new demands in the industry (Pichyangkul et al., 2015; Thayaparan et al., 2010). These structural changes pose additional cognitive demands on the construction worker. It is possible, but not yet known, that compared to other types of jobs, an increase in cognitive demands may lead to a significant deterioration in health as construction workers face already high demands in the physical domains. Whereas big construction companies have the resources to provide adequate training for their workforce, little attention has been paid to smaller firms that employ the majority of the construction workers (Dainty et al., 2005). For instance in Germany, 89 percent of construction

companies have less than 20 employees (Federation of the German Construction Industry (Hauptverband der Deutschen Bauindustrie, 2019). Given the digital changes in the work environment, it is likely that these new demands are relevant for construction workers in small companies as well, but the extent is unknown so far. Our goal was to record explicitly the requirements of construction workers in smaller companies, as they employ significantly the majority of people in Germany.

3.2.1.2. Aim of the Study

As it is unclear what demands, in addition to physical demands, workers are faced with in the current construction industry, this study's aim was to assess the physical and cognitive demands in construction work and outline the amount and extent of these demands that workers are faced with on a daily basis. Accordingly, a large part of the analysis is explorative because, without previous evidence, we cannot yet formulate a specific hypothesis on the expected results. As construction work is considered a physically demanding job, we hypothesize that physical demands are higher than cognitive demands. In addition, the study aimed at assessing stress symptoms in construction workers. We hypothesize that higher demands are associated with more stress symptoms. We conducted structured interviews and a survey in three smaller construction companies in Germany to assess the extent of physical demands and a variety of cognitive demands as experienced by construction workers. Small companies were selected on purpose, as they reflect the majority of the construction companies in Germany.

The paper continues with a literature review and a description of the methods (Section 2). The results are presented in Section 3 of the paper, which will begin with the results on physical demands as assessed via the Fleishman job analysis and the NASA Task Load Index (NASA-TLX) (Section 3.1), followed by the results on cognitive demands as assessed via the Fleishman job analysis, the NASA-TLX and the Mental Work Demands questionnaire (Section 3.2). In Section 3.3, the results on stress symptoms including, first, the results on somatic symptoms and symptoms of depression as assessed via the Patient Health Questionnaire (PHQ), second, the results on personal and work burnout as assessed via the Copenhagen Burnout Inventory (CBI), and third, associations between demands and stress symptoms are reported. In Section 4, the results are then discussed in context with previous studies and with current standards of practice, and practical implications are derived.

3.2.1.3. Literature review

Most of the previous studies on demands in construction work focus on physical demands. Research shows, for instance, that construction workers exceed the thresholds for energetic workload regularly (Boschman et al., 2011). Statistics from the US Department of Labor Employment and Training Administration show that construction jobs are more physical and ergonomically challenging than other jobs (Schneider et al., 1998). Based on the relevance of high physical demands, most current studies focus on improved assessment of these demands (e.g. Yang et al., 2018) and interventions to reduce these demands in construction work (e.g. Alabdulkarim and Nussbaum, 2019). Research on other types of demands is less readily available. One study asked construction supervisors to indicate demands in their workplace: they reported that the highest demands are a lack of competent staff (75 percent), inadequate staffing (up to 69 percent), inadequate communication (68 percent) and multiple regulations (37 percent), in addition to the rapid change of tasks, holding conversations, financial management, complex decisions and paperwork-caused stress (Boschman et al., 2011). Another study from Germany shows that construction supervisors view task interruptions (47 percent) and interference (21 percent) as important stressors (Stadler and Bayerisches Landesamt für Gesundheit, 2012). The most recent studies that we could identify regarding this topic were a study by Leung et al. (2015), which observed associations between job stress and safety behavior, and a study by Chen et al. (2016), which did a prototype testing of a wearable electroencephalography safety helmet to assess mental workload of construction workers. Overall, there seems to be a lack of information in the literature on other types of demands such as cognitive demands. One reason for the lack of research findings could be that most researchers use cognitive task analysis techniques, in which workers are asked to articulate the challenges they experience (Roth, 2008). This method does not allow us to derive information on the level of cognitive demands. The lack of information motivated us to conduct the present study.

Knowing about different types of demands in construction work is important because they can affect workers' health. As evidence from construction work is scarce, we look at studies with other occupational groups: research on job demands (other than physical demands) has demonstrated significant associations with deterioration of health. A survey from the Japanese labor union has shown a significant relationship between high job demands (in terms of excessive amounts of work and working fast/hard) and health symptoms (Sakano et al., 1995). An online

survey of construction managers investigated “job demands,” defined as the sum of work quantity, time pressure and concentration, and observed associations between higher levels of job demands and psychological strain (Bowen et al., 2014). A systematic review of studies with construction workers concluded that high quantitative job demands are a risk factor for musculoskeletal disorders (Sobeih et al., 2006). Each of these studies, however, used a sum score of work quantity, time pressure, concentration and similar to define job demands. None of them distinguished between different types of demands. Only few studies examined associations between certain types of demands in construction work and poor health, each of them focusing on stress symptoms. One study has shown that high work speed and quantity are associated with symptoms of depression (Boschman et al., 2013), and another study has shown that job schedule irregularities and number of hours worked per week are associated with burnout (Lingard and Francis, 2005). We were not able to identify further studies investigating such associations. Compared to the large number of studies on physical demands and health, the number of studies on other types of demands in construction work is so small that it is difficult to draw sound conclusions. As it is largely unclear what types of demands, in addition to physical demands, workers are faced with in construction work, the major aim of this study was to assess the physical and cognitive demands in construction work and outline the amount and extent of these demands that workers are faced with on a daily basis. In addition, we provide some information on stress symptoms.

3.2.2. Methods

3.2.2.1. Study Group

Three construction companies in Germany specialized in plastering, interior fittings and drywall construction agreed to participate in the project. The number of people employed in these companies were n = 27, n = 50 and n = 80. We intentionally selected small companies as they make up the majority of construction companies in Germany (Federation of the German Construction Industry (Hauptverband der Deutschen Bauindustrie), 2019) and therefore represent the general trend in the field. We continuously contacted small construction companies in the area until we had the agreement of three companies to participate in the study. All construction workers of the three companies were asked whether they wanted to participate in the study. Inclusion criteria were working regularly in the company and receiving a salary from them. Employees had

the chance to participate in a structured interview on 73 different job demands (Fleishman job analysis survey) and/or a survey that comprised job demands in general (NASA-TLX) as well as the details on cognitive demands at work (Mental Work Demands Questionnaire). Our research design contained both, the interview and the survey, for two reasons. First, the Fleishman job analysis was too complex for construction workers to answer in a survey style so that we decided to conduct the Fleishman job analysis in the form of an interview. Second, we wanted to explore the general level of demands, all the different types of demands and the details of the mental demands at work using several assessment tools. As the interview as well as the survey each took about 1 h to complete, we conducted them at different points in time to avoid stress. Participation was voluntary. A total of 35 construction workers (n = 11 Company 1, n = 12 Company 2, n = 12 Company 3) agreed to participate in the interview. In addition, a total of 30 construction workers (n = 6 Company 1, n = 12 Company 2, n = 12 Company 3) agreed to complete the survey.

3.2.2.2 Study Design

To assess the amount and extent of physical and cognitive demands that construction workers are faced with on a daily basis, we conducted structured interviews and a structured survey that took place during an ordinary working day. All employees that were actively working as a skilled construction worker (e.g. electrician, painter, plasterer and drywall installer) in any of the three companies were eligible to be included in this study. A member of the study team visited the workers either at their place of work or in an office. The workers were first informed about the purpose and content of the study and had the chance to ask questions. If they agreed to participate, the interview or survey would begin. The study was approved by the ethics committee of the University of Kaiserslautern.

3.2.2.2.1 Interviews

The purpose of the interviews was to assess work demands according to the well-established Fleishman job analysis survey. The interviews were conducted face-to-face, separately with each individual construction worker directly at the worksite during regular working hours. This was usually a construction site inside a building and, in some cases, on exterior scaffolding or in an office. The interviewer came to the construction worker's location and asked whether the construction worker was interested in answering the questions. A total of 35 workers (34 male

and 1 female) who were actively working on an apartment or office building (electricians, floorers, painters, plasterers and drywall installers) took a break from their work activities to answer the questions. Of those who were willing to give information on their demographics, the age range was 19–49 years old (Mean: 33.2 and SD: 9.8), 80 percent had a secondary school certificate and 20 percent had completed at least some college. The interview was comprised of the German version of the Fleishman job analysis survey (Kleinmann et al., 2010). The Fleishman job analysis survey is an ability requirement taxonomy with meaningful description of 73 job activities (Fleishman and Mumford, 1991) such as “gross body coordination”, “gross body equilibrium”, “dynamic flexibility”, “reliability”, “friendliness”, “auditory attention” among others (see Appendix I for complete set of job demands in the Fleishman job analysis). It was originally developed for evaluating physical abilities required by jobs (Fleishman, 1979), but was then modified and used as one of the basic components of the O*NET ability taxonomy and measurement system from the US Department of Labor, Employment & Training Administration (Peterson et al., 2001). We chose this assessment method because it assesses a variety of job demands and it has since been validated in many studies (Hogan et al., 1980; Fleishman and Mumford, 1991; Cunnigham et al., 1996). The construction workers were interviewed individually by the interviewer who read out the demands in the Fleishman job analysis one by one and asked the construction worker to rate it on a scale of 1 (not at all) to 7 (extremely demanding) by giving the examples of the ratings delineated in the manual. Usually, the demands were discussed based on the description in the manual until the construction worker felt confident to give a reliable rating.

3.2.2.2.2 Survey

The purpose of the survey was to assess cognitive demands in construction work more in depth. The survey was conducted in the headquarters of the construction companies either before or after a regular workday. All construction workers were invited to come to a meeting room where the researcher informed them about the purpose and the content of the survey. Construction workers that were interested in completing the survey were given either a paper booklet or a tablet with the survey to complete based on their preferences. The survey comprised the NASA-TLX for assessing the overall level of work demands and the Mental Work Demands questionnaire for assessing detailed intellectual demands. The NASA-TLX comprises six questions on perceived workload due to “physical,” “mental,”

"temporal," "performance," "effort," and "frustration"-related demands on a scale of 0 (very low) to 100 (very high). It was originally developed to assess workload in laboratory conditions (Hart and Staveland, 1988). We chose this assessment because it is an easy-to-use and reliable measure to assess task difficulty in different settings and is used intensively worldwide (Hart, 2006). Participants completed the German version of the NASA-TLX, as it is usually presented on paper. They indicated their ratings from 0 (very low) to 100 (very high) by making a cross with a pencil or, in the tablet version, with a finger touch. To assess the details of the mental demands that construction workers face at work, we used the Mental Work Demands questionnaire that contains 18 questions originally derived from the O*NET survey instruments, which was developed by an interdisciplinary expert team (Handel, 2016). The 18 questions describe intellectual demands at work such as "updating and using relevant knowledge," "giving advice and consultation," and "developing objectives and strategies" (see Table 1 for the complete set of job demands in the Mental Work Demands questionnaire) on a scale of 1 (very low) to 7 (very high). These intellectual demands are combined in three indices of mental demands: verbal demands, executive cognitive demands (Then et al., 2013) and information processing demands (Then et al., 2017). These indices were originally developed for the purpose of operationalizing an intellectually demanding work environment (Then et al., 2017; Then et al., 2013). We chose this assessment because it allows the detailed assessment of cognitive demands. Participants indicated their ratings from 1 (very low) to 7 (very high) by making a cross with a pencil or, in the tablet version, with a finger touch. To assess stress symptoms, the survey comprised assessments

	M	CI 95%	SD
Verbal demands	3.93	3.44–4.42	1.29
Getting information	3.86	3.19–4.53	1.77
Evaluating Information[b]	4.04	3.51–4.56	1.37
Processing Information[b]	3.63	3.03–4.23	1.52
Analyzing information[b]	3.22	2.66–3.78	1.42
Interpreting information	3.21	2.64–3.79	1.49
Updating and using relevant knowledge[b]	5.07	4.47–5.66	5.07
Interpreting the meaning of information[b]	3.56	2.83–4.28	1.85
Providing consultation and advice	4.46	3.74–5.19	1.89
Executive cognitive demands	3.5	2.88–4.12	1.67
Negotiating with others	4.07	3.36–4.77	1.91
Resolving conflicts[b]	4.14	3.39–4.89	1.99
Developing objectives and strategies	2.93	2.20–3.65	1.88
Scheduling work and activities	3.38	2.62–4.14	2.03
Coordinating the work and activities of others	3.18	2.50–3.86	1.76
Guiding, directing and motivating subordinates	3.43	2.55–4.31	2.28
Information processing demands[b]	3.89	3.40–4.39	1.34
Identifying objects, etc.	4.10	3.43–4.78	1.79
Documenting Information	4.11	3.44–4.78	1.72
Estimating quantities	3.50	2.77–4.23	1.96
Communicating	4.48	3.76–5.21	1.92

Notes: CI, confidence interval; M, mean; SD, standard deviation. [a]$n = 34–38$ as not all participants answered all questions; [b]also included in information processing index

Table 1: Mean level of demands in the mental demands questionnaire[a]

for somatic symptoms, depression symptoms and burnout. Somatic symptoms were assessed using the 15-item PHQ on somatic symptoms (PHQ-15). We chose the PHQ because it is a short, self-administered questionnaire designed for use in primary care and non-psychiatric settings to diagnose patients according to the International Classification of Diseases or the Diagnostic and Statistical Manual of Mental Disorders (Gilbody et al., 2007). The PHQ-15 comprises 15 questions that measure somatic symptoms (Kroenke et al., 2002). It has shown strong validity in primary care settings (Interian et al., 2006) and in the general population (Kocalevent et al., 2013). Depression symptoms were assessed using the PHQ Depression Scale (PHQ-9), a short nine-item diagnostic measure for depression (Kroenke et al., 2001) that has shown strong clinical validity (Martin et al., 2006; Manea et al., 2012). The critical cut-off score for mild depression is 5 points and for clinically relevant depression is 10 points (Kroenke et al., 2001). Burnout was assessed using the CBI for personal and work-related burnout. The CBI was originally developed by Tage Kristensen and colleagues at the National Institute of Occupational Health in Copenhagen (Kristensen, Hannerz, Hogh and Borg, 2005) and has been validated internationally (Kristensen, Borritz, Villadsen and Christensen, 2005; Winwood and Winefield, 2004). The critical cut-off score for burnout is 50 points (Kristensen, Hannerz, Hogh and Borg, 2005). We used the original versions of each of those assessment tools in our survey. Participants indicated their ratings on the original
scales of each of these assessment tools by making a cross with a pencil or, in the tablet version, with a finger touch.

A total of n = 30 employees from the construction companies (construction site manager, electricians, floorers, painters, plasterers and drywall installers) agreed to participate in the survey. The age range was 18–60 years old with a mean age of 34.79 (SD: 11.83) of the construction workers. Of all the participants, only one was female. A total of n = 16 construction workers (57.1 percent) had completed less schooling than a secondary school certificate and n = 7 (25.0 percent) completed some form of secondary schooling. Another n = 5 construction workers (17.9 percent) had acquired a university-qualifying school certificate and completed at least some college.

3.2.2.3 Statistical analyses

All statistical analyses employed an α level for statistical significance of 0.05 (two-tailed) and were performed using Stata (version 15). Means and standard

deviations were calculated using standard procedures in Stata. Comparison between construction workers and office workers were calculated by using Kruskal–Wallis test if the scales were ordinal or not normally distributed (Chi^2 was obtained as a significance indicator in the Kruskal–Wallis test). Correlations between demands at work and stress symptoms were calculated by using Kendall's τ rank correlation, as the scales were ordinary and our sample was relatively small.

3.2.3. Results

As the aim of this study was to assess physical and cognitive demands in construction work, we present the results in the corresponding order. In addition, we present results on stress symptoms in construction workers and their association with these demands.

3.2.3.1 Physical Demands

Information on the extent of physical demands in construction work was obtained via the Fleishman job analysis in great detail and via the NASA-TLX in the form of an overall score. Results from the Fleishman job analysis interviews suggest that, on a scale of 1 (very low) to 7 (very high), the average level of physical demands among construction workers was 4.84 (SD: 0.73), indicating a high (but not very high) overall level of physical demands. The highest levels were reported for gross body coordination and gross body equilibrium and the lowest levels for dynamic flexibility and stamina. Details of all physical demands are shown in Table I. Results from the survey using the NASA-TLX suggest that on a scale of 0 (very low) to 100 (very high) the average level of physical demands was 68.61 (SD: 33.07), indicating a slightly more than moderate level with large variance.

3.2.3.2 Cognitive Demands

Information on the extent of cognitive demands in construction work were obtained via the Fleishman job analysis and the Mental Work Demands questionnaire in great detail and via the NASA-TLX in the form of overall scores for temporal demands, effort-related demands, mental demands, frustration-related demands and performance-related demands, respectively. Cognitive demands at work were obtained on a scale of 1 (very low) to 7 (very high) in interviews using the Fleishman job analysis and in a survey using the Mental Work Demands questionnaire, respectively, and the NASA-TLX on a scale of 0 (very low) to 100 (very high). Results from the Fleishman job analysis interviews indicate that the average

level of psychomotor demands were 4.63 (SD: 0.60), social demands were 4.42 (SD: 0.79), cognitive demands were 3.75 (SD: 0.54) and sensory–perceptual demands were 3.67 (SD: 0.79), suggesting high, but not very high, cognitive demands in construction work. The highest levels were reported for reliability, friendliness, assertiveness and motivation and the lowest levels for originality and auditory attention (see Table II).

The results from the survey with the NASA-TLX on a scale of 0 (very low) to 100 (very high) indicate that the average level of temporal demands were 77.78 (SD: 21.09), effort-related demands were 77.78 (SD: 23.40), mental demands were 66.25 (SD 29.43), frustration-related demands were 47.78 (SD: 26.75) and performance-related demands were 31.67 (SD: 27.22), suggesting that temporal and effort-related demands were relatively high and performance-related demands were relatively low. Results from the Mental Work Demands questionnaire on a scale of 1 (very low) to 7 (very high) indicate that the average level of verbal demands among construction workers were 3.93 (SD: 1.29), information processing demands were 3.89 (SD: 1.34) and executive cognitive demands were 3.5 (SD: 1.67), suggesting moderate levels of intellectual work demands (see Figures 1 and 2). Details are shown in Table II. The highest levels were reported for "updating and using relevant knowledge" and "giving advice and consultation" and the lowest levels for "developing objectives and strategies".

3.2.3.3 Stress Symptoms

Information on the extent of stress symptoms in construction work and their association with physical and cognitive demands were obtained via the PHQ and the CBI. Stress symptoms included an overall stress level, somatic symptoms, symptoms of depression, personal burnout and work-related burnout. Results on stress indicate that, on a scale of 0–10, construction workers reported to experience on average a

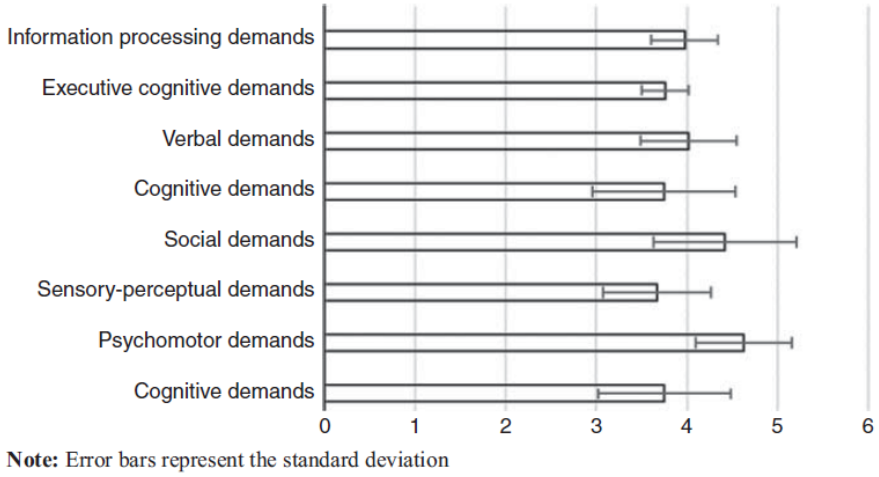

Figure 1: Mean level of demands at work according to the Mental Demands Questionnaire and the Fleishman job analysis survey

stress level of 6.72 (SD: 1.98), suggesting a relatively high stress level. Results on somatic symptoms from the survey indicate that 6.9 percent of the construction workers (2 out of 29) had any type of somatic symptoms (e.g. headache, dizziness and an upset stomach or others). The average number of somatic symptoms was 0.89 (SD: 1.37). The results on depression indicate that the average level of depression symptoms was 4.17 (SD: 3.04). The findings suggest that 40.0 percent (n¼12 out of 30) of the construction workers had a mild level of depression symptoms and 6.7 percent (n¼2) had clinically relevant symptoms of depression. The results on symptoms of burnout comprise the dimensions personal burnout and work burnout. The average level of personal burnout was 41.93 (SD: 16.91). A total of 39.4 percent of construction workers (13 out of 33) scored over the critical score for personal burnout. The average level of work burnout was 35.16 (SD: 15.61). A total of 12.5 percent of construction workers (4 out of 32) scored over the critical score for work burnout. Overall, 60.7 percent of the construction workers did not have any symptoms (this number does not include individuals with missing data).

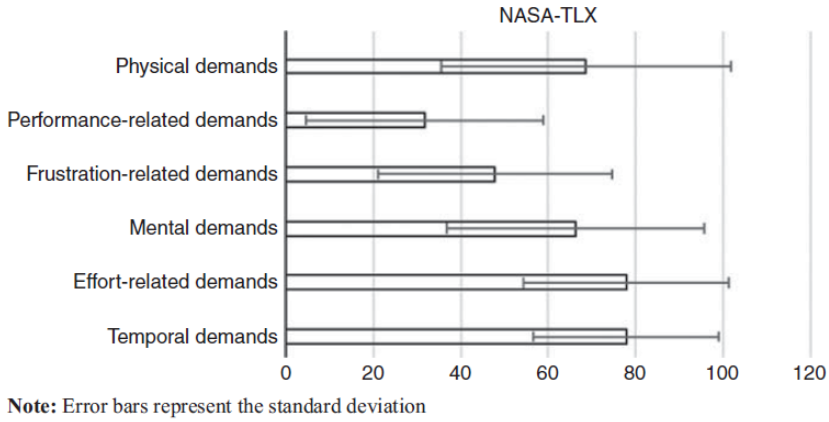

Figure 2: Mean level of demands at work according to the NASA Task Load Index (NASA-TLX)

Investigating associations between demands and stress symptoms is difficult given the small variance in the demands among construction workers and the limited number of construction workers who answered all of these questions (n¼16). Comparisons of those with stress symptoms and those without symptoms suggest that more stress, higher effort-related demands and higher physical demands were associated with stress symptoms (see Figures 3 and 4, and Table 2). No comparisons were statistically significant in the Kruskal–Wallis test (see Table III). Statistical findings from Kendall rank correlation suggest that higher psychomotor demands correlate with more somatic symptoms (see Table 3).

3.2.4. Discussion

The aim of this study was to assess physical and cognitive demands in construction work. Results from interviews (n¼35) and a survey (n¼30) in three construction companies in Germany suggest that the extent of physical demands is high but not very high and the extent of cognitive demands is also in a moderately high range, especially for verbal and information processing-related demands. In this sense, we cannot confirm our hypothesis that physical demands are higher than cognitive demands. Major aspects driving cognitive demands that construction workers face are updating and using job-related knowledge, in addition to giving advice and consultation. In each construction site, workers encounter different settings, buildings and construction plans and they are expected to adapt to each of the different new work sites accordingly. New innovations, technological

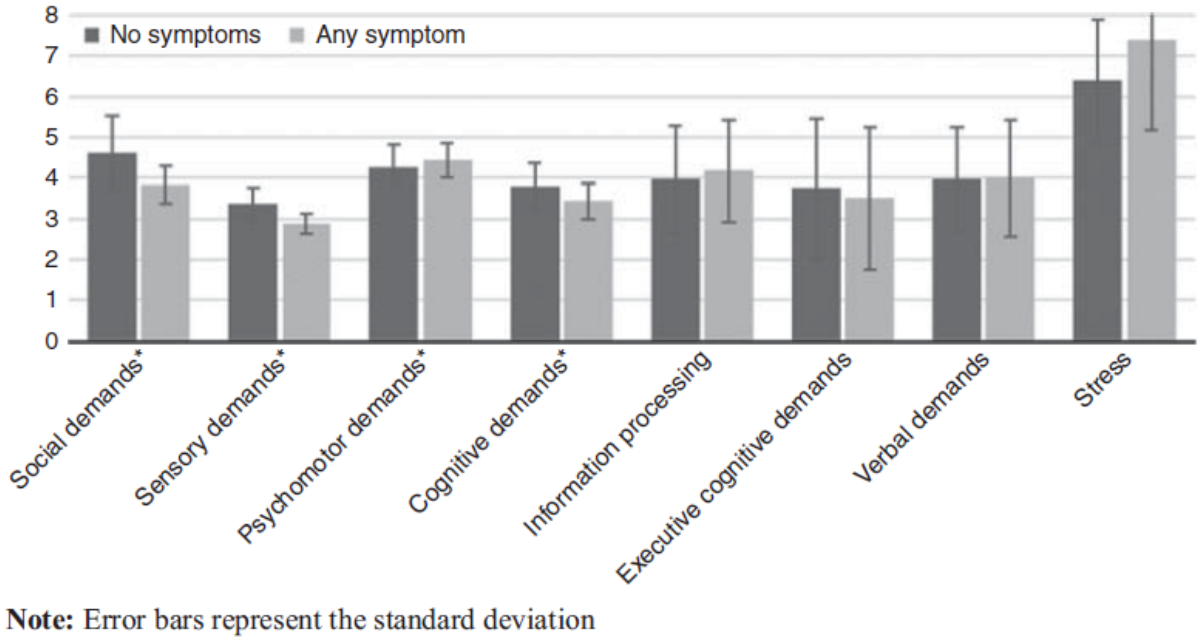

*Figure 3: Mean level of demands at work for participants with and without symptoms, assessed via the Mental Demands Questionnaire and, for *, via the Fleishman job analysis survey in a scale from 1 (very low) to 7 (extremely high), except for stress, which was assessed on a scale from 1 (very low) to 10 (extremely high)*

developments and new regulations concerning materials, safety and building construction need to be adapted to existing work processes (Sexton and Barrett, 2003; Kaplinski, 2018; Rutešić et al., 2015) and this might lead to additional cognitive demands. Another underestimated component of a construction worker's job seems to be consultation. Consultation is a service that workers provide when they inspect a building, make decisions of the relevant work steps and identify risks and problems instantaneously – a service that is, however, not frequently taken into account. We were not able to identify any scientific publication dealing with this issue, suggesting a considerable research gap. A practical implication of this finding

is that construction workers in small-sized companies seem to provide a service that is not acknowledged. If this is part of the regular project delivery system, this service needs to be included in the planning of the construction site.

Based on our findings, it seems clear that construction work is much more than just a physical job. Construction workers face at least moderately high demands in all types of demands (see Table I and Figure 1). Typically, every occupation has at least one area in which demands are low. For example, office workers or military intelligence personnel have low gross motor demands (Knapp et al., 1991). To get an impression of the uncommonly wide-ranging demand levels that construction workers face, the information of the Fleishman job survey in the O*NET database can help (www.onetonline.org). Compared to clerks, construction workers have a lower level in verbal communication but a more than double the level in psychomotor, physical and sensory demands, a more than two scale-points (Scale 1–7) higher level in fluency of ideas, memory, spatial orientation and imagination and a more than one scale-point higher level in seven other cognitive domains than clerks. With ongoing digitization of the sector, these cognitive demands may further increase, placing even higher demands on construction workers.

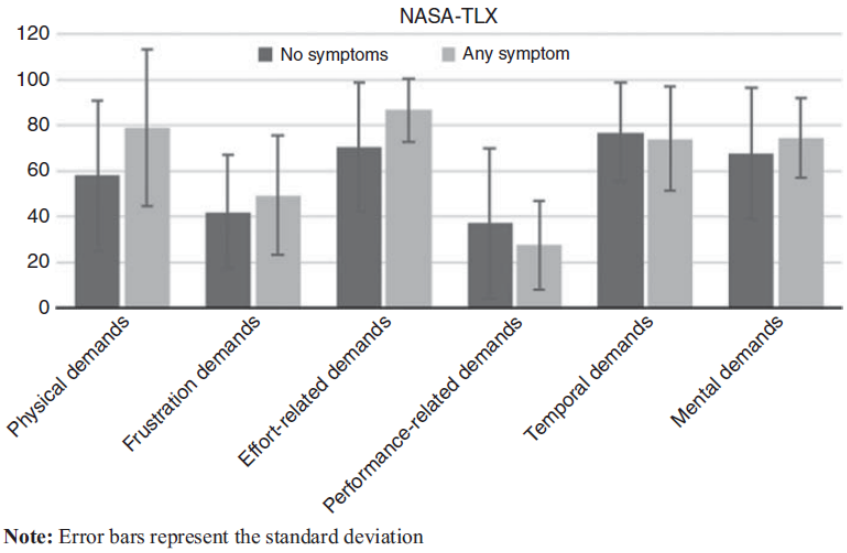

Figure 4: Mean level of demands at work according to the NASA Task Load Index (NASA-TLX) for participants with and without symptoms, assessed on a scale from 0 (not at all) to 100 (very much)

Our results suggest that construction workers endure a multi-component strain everyday at work. This could be one explanation for why every fourth construction worker in our sample had work-related burnout symptoms and every third had depressive symptoms, as many previous studies have shown that job demands are associated with a greater risk for burnout (Seidler et al., 2014) and depression (Mausner-Dorsch and Eaton, 2000). Policies to prevent cognitive overload in construction work are urgently needed. Yet, research findings that

provide the necessary evidence for such policies are not available. Although our sample size was rather small, we observed a positive association between temporal and effort-related demands and stress on burnout symptoms in construction workers. These findings confirm to some extent our hypothesis that higher demands are associated with more stress symptoms. Temporal demands like time pressure are a source of accidents at work (Gravseth et al., 2006). Time pressure increases physiological and psychological stress reactions (Wahlström et al., 2002; Wofford, 2001) and thus can lead to mental illnesses (Ilies et al., 2010). Time pressure is also associated with both lower back and upper extremity symptoms, especially in combination with high cognitive demands or interpersonal demands (Huang et al., 2003). Thus, the benefits from increasing the pace of work might be offset by generating work accidents, decreasing motivation to work and causing productivity losses (Nepal et al., 2006). A practical implication is that each demand, including cognitive demands, must be allocated a minimum amount of time in the planning of the construction site. Only in this way, time pressure and the consequences can be prevented. Digital work process planning tools are available already. However, planners often allocate time slots for the practical aspects of the work task and do not yet consider the cognitive effort necessary for the construction worker to complete the work package.

	No symptoms		Any symptom				
	M	SD	*M*	SD	χ^2	*p*	*n*[c]
Physical demands[a]	4.75	0.53	5.00	0.68	0.013	0.909	10
Social demands[a]	4.62	0.90	3.83	0.46	1.870	0.172	10
Sensory demands[a]	3.36	0.39	2.89	0.25	2.195	0.139	10
Psychomotor demands[a]	4.27	0.54	4.43	0.42	0.117	0.732	10
Cognitive demands[a]	3.79	0.57	3.43	0.45	1.052	0.305	10
Information processing	3.96	1.32	4.18	1.25	0.176	0.675	27
Executive cognitive demands	3.76	1.71	3.50	1.73	0.220	0.639	27
Verbal demands	3.96	1.30	4.00	1.42	0.000	1.000	27
Stress	6.40	1.50	7.40	2.22	0.889	0.346	25
Physical demands[b]	58.0	32.9	79.2	34.4	3.203	0.074	16
Frustration demands[b]	42.0	25.1	49.2	26.2	0.026	0.871	16
Effort-related demands[b]	70.5	28.0	86.7	13.7	1.424	0.233	16
Performance-related demands[b]	37.0	33.0	27.5	19.4	0.003	0.957	16
Temporal demands[b]	77.0	21.8	74.2	23.1	0.009	0.923	16
Mental demands[b]	67.7	28.9	74.4	17.4	0.047	0.828	23

Notes: CI, confidence interval; M, mean; SD, standard deviation; χ^2, measure as obtained via the Kruskal–Wallis equality of populations rank test. [a]According to the Fleishman job analysis survey; [b]according to the NASA Task Load Index (TLX); [c]variance because not all participants answered all questions

Table 2: Mean level of demands by individuals with and without stress symptoms

Effort-related demands are known for similar effects on workers' health. Effort-reward imbalance at work is a widely known concept that has been shown to contribute to poor mental health (Siegrist and Dragano, 2008) and also plays a causal role in psychological distress and physical complaints in the future (Shimazu and Jonge, 2009). A high level of effort, as experienced by construction workers, is thus a risk for poor health if they do not receive the appropriate reward

for it, which could be monetary gain, career opportunities, job security, early retirement or other types of rewards (Federal Health Monitoring Information System, 2019; Jacobi, 2009). Therefore, it is essential that companies learn to recognize high demands in the work context and monitor them together with the effect that demands have on their workers.

Our research findings give rise to a number of practical implications. First, work steps need to be reevaluated to include cognitive demands in order to avoid a cognitive overload of the construction workers. Cognitive demands need to be added to the planning of work packages. Second, the planning of work activities on the construction site should include time slots for applying knowledge and for consultation. Third, to avoid time pressure, project schedules should allow for flexibility and time constraints in the contracts of construction projects should be eluded. Fourth, services that construction workers supply such as consultation or the use of knowledge of standards and regulations should be officially recognized as such. Fifth, given the recent developments in the sector, enhanced and continuous training is necessary to guarantee that construction workers have the skills to deal with every challenge at work. Furthermore, public policies could help to ensure worker well-being. We recommend a new public policy that regulates the use of new technological devices on construction site. Specifically, new technological devices should only be used after proper testing regarding its usability and the demands that using the device adds to the work activities. We also recommend public policies on workers' health to expand the minimum safety standards to include cognitive and social demands. Risk assessment should always include these demands, which – although not always obvious for construction work – are a crucial factor for well-being. In addition, it is necessary to make resource materials available to small-

	Depression symptoms		Somatic symptoms		Work-related burnout	
	T	*p*	*T*	*p*	*T*	*p*
Physical demands[a]	0.119	0.689	0.396	0.193	0.164	0.519
Social demands[a]	−0.305	0.236	−0.479	0.095	−0.385	0.131
Sensory demands[a]	−0.359	0.156	−0.506	0.078	−0.248	0.342
Psychomotor demands[a]	−0.019	1.000	0.586	0.039	0.419	0.096
Cognitive demands[a]	0.169	0.528	−0.027	1.000	−0.133	0.635
Information processing	0.134	0.348	0.034	0.841	0.167	0.226
Executive cognitive demands	0.104	0.472	−0.034	0.841	0.148	0.289
Verbal demands	0.119	0.402	−0.082	0.609	0.022	0.889
Stress	0.250	0.104	−0.013	0.979	0.261	0.083
Physical demands[b]	0.008	1.000	0.384	0.066	0.286	0.141
Frustration demands[b]	−0.118	0.557	0.244	0.249	0.192	0.332
Effort-related demands[b]	0.056	0.801	0.189	0.381	0.032	0.899
Performance-related demands[b]	−0.111	0.585	0.237	0.268	0.000	1.000
Temporal demands[b]	−0.125	0.531	0.065	0.783	0.056	0.801
Mental demands[b]	0.061	0.704	0.147	0.404	0.155	0.304

Notes: *p*, level of significance; *T*, Kendall's tau-b correlation coefficient. [a]According to the Fleishman job analysis survey; [b]according to the NASA Task Load Index (TLX)

Table 3: Correlation coefficients for the association between demands at works and stress symptoms

and medium-size companies that help them to guarantee worker well-being in the long run.

The current study has some limitations. First, the number of participants in the study was relatively small and some of them did not answer all the questions. Psychological theories infer that those workers who fear negative consequences in particular (for instance, because their stress level is too high) are reluctant to participate (Graham et al., 1994).

Missing data from construction workers that feel very stressed due to their work may have led to an underestimation of the consequences of high work demands. Some workers would simply skip entire sections of the questionnaire resulting in missing data in some questionnaires for these individuals. This is unsurprising given that the response rate in the O*NET survey was only 16 percent (Campion et al., 1999). Therefore, we would like to encourage further studies to replicate our findings. Second, the study did not include the organizational structure or interpersonal aspects of work, which of course can have a major impact on the workers' productivity and well-being. Finally, the demand ratings are subjective, which could be a limitation but previous studies have shown that both hold predictive power (Sonnega et al., 2017). In fact, experiments have shown that there is a strong relationship between demands and ratings on effort (Hogan et al., 1980). Some scientists even consider personal perceptions more important because they encompass implicit information that cannot be observed (Jahedi and Méndez, 2014).

3.2.5. Conclusion

The aim of this study was to assess physical and cognitive demands in construction work and outline the amount and extent of these demands that workers are faced with on a daily basis. The results shows that construction work is demanding in physical terms as well as in perceptual, psychomotor, social and cognitive terms. Using and updating specialized knowledge, giving advice and providing consultation, friendliness, assertiveness and reliability are underestimated demands in the everyday work environment of construction workers. However, these demands are relevant to the workers' productivity and well-being. Further studies need to investigate the details of the relationship between the multi-component demands profile in construction work and the effects that it has on mental health. A practical implication is that those working in the construction industry must gain awareness of these different types of demands.

Managers should recognize demands in the cognitive and psychosocial domain when planning work processes and when shaping the organizational structure in the construction industry. A lack of consideration can lead to an overload of the individual worker or to constraints in the time schedule that might collapse of the time management of the project. In the long run, a lack of consideration may also lead to a mismatch between construction workers' performance and their health in the future. As employers have to protect workers from the risks arising from construction work, it is important to include cognitive demands in the occupational risk assessment as required by law (Herbig et al., 2012). Yet, at present, there is no standard application of this in the construction sector. An implication for public policy is therefore to provide resource materials that help small- and medium-sized companies to realize that. Occupational risk assessments of this type are constrained by the fact that certain methods, such as the cognitive task analysis, are limited and do not acknowledge the full range of cognitive and psychosocial demands. We strongly encourage those working in the construction industry to get together with researchers to develop urgently needed measures that include mental, social and cognitive demands in risk assessments, especially given the increasing digitization of the industry. Increasing levels of demands for construction workers in small companies (e.g. digital skills, expertise in standards and regulations and provided consultation services) can also have economic implications, as skilled labor valuable and costs in building construction may rise.

3.3. Interim Summary

Study 1 investigated whether different assessments using pen and paper, pen and tablet and finger and tablet lead to comparable measures of processing speed. The main results include improved performance and subjective preference in the tablet and pencil assessment mode compared to tablet and finger and paper and pen modes. Implications for digitalisation of tests have been discussed as well.

Study 2 aimed to use interviews and surveys to assess the physical and cognitive demands in construction work. Several demands were identified and potential ways to implement them into the increasingly digitalized work practice have been discussed.

In-depth discussion of these results by embedding them into the larger picture of the present work will take place in chapter 5.

3.4. Further Related Research

In addition to study 1 and study 2 two further experiments will be briefly described here. These experiments did either not get finished due to the Covid19 pandemic or are in the preparation of getting published. Since this work is partially still ongoing, it will only be mentioned briefly here, however their inclusion is relevant due to their added value to the overall scope of the research presented in this work.

The data for the first study is currently being analysed. It concerns a secondary explorative study conducted in conjunction with two colleagues who investigated the impact of music on task performance, using stimuli they validated previously (Hofbauer & Rodriguez, 2023). Participants were instructed to complete several cognitive tasks related to working memory - immediate and delayed word list recall and the trail-making test (TMT, Oswald & Roth, 1987) which was also used in study 1– while listening to either fast or slow music with positive or negative valence in a within-subjects design. A total of fifty participants was recruited and tested. For the duration of the experiment, participants wore a mobile Tobii Glasses 2 eye-tracker (Tobii Pro, Danderyd, Sweden). The main research question was concerned whether the type of music had an impact on task performance, measured either by performance in the recall tasks or the completion speed and error rate in the TMT. These behavioural measures are currently being analyzed by our colleagues. As an explorative addition, gaze and continuous pupillometry data was collected as well. The gaze data involves fixation durations and number of fixations but has proven challenging to analyze due to the semi-structured setup of the experiment. The pupil data was less affected by these shortcomings. The main assumption was that pupil dilation should relate to the behavioural data as a marker of cognitive load (Wierda, Van Rijn, Taatgen, & Martens, 2012; Van der Wel & Van Steenbergen, 2018) and optimally mirror the results from the behavioural study. Analyses are ongoing, but first insights reveal task-specific differences in pupil dilation which correspond to the task difficulty – delayed recall for instance eliciting an increased average pupil size compared to immediate recall of the same word list. This experiment is relevant for the present work since it uses similar principles to study 1 but incorporates more in-depth eye-

tracking analyses. However, due to the explorative nature of this data and technical problems which occurred during testing but were only discovered during data analysis, this research is still ongoing.

The second study which is of interest follows a similar research question to study 2, relying partially on the same F-JAS structured interviews, in this case to investigate the mental and physical effort required in car mechanics tasks. In addition to the interviews, three scenarios with different tasks relevant for car mechanics - changing a tire, disassembling an engine part and error search for a faulty car engine – were created in close collaboration with a vocational school. To investigate how demanding these tasks were, a dual approach was taken – first, participants at different stages of their career as car mechanics evaluated them in a structured interview with the same dimensions described in study 2 (see Appendix table 1 & 2 for full factor structure). Second, participants from the vocational school were recruited who either had no background in car mechanics at all or who were in their first or second year of a car mechanics apprenticeship. These participants were instructed to complete the described tasks while wearing a mobile Tobii Glasses 2 eye-tracker (Tobii Pro, Danderyd, Sweden) which would record their gaze behaviour and fixation patterns. Participants received detailed verbal and written instructions before each task but were otherwise free to complete it at their own speed.

The first results from a pilot study were promising especially for the error search, indicating differences in task completion speed but more interestingly also in their gaze pattern – with participants without any expertise using a random search pattern, participants with some expertise focussing on the car battery and brand of car and an expert who was recruited as reference finding the error within seconds, fixating solely on specific parts of the car battery (see Appendix Figure 1). Unfortunately, due to several factors, most severely the outbreak of Covid19 and resulting lockdowns, this line of research remained unfinished. Nevertheless, it represents an important piece in the overall picture discussed here, because it bridges the gap between study 2 and the following research, which uses one of the three tasks mentioned here (disassembly of an engine part) as part of the educational material.

Chapter 4: Virtual Reality Research

This chapter is concerned with the second research area of interest and contains two studies, which were published in 2021 and 2022, respectively. Study 3 was built on a similar principle as study 1 and aimed to provide an early glimpse into the potential of using virtual reality technology in learning environments by comparing it to more traditional 2D educational videos. Study 4 was conducted as a full virtual reality experiment and combined a classic visual search paradigm with different audio cue modalities and introduced visual noise as an additional independent variable.

Aside of these two main studies, several further related early-stage or otherwise unpublished research utilizing mixed reality and virtual reality from our work group will be described in some detail.

4.1. Remote Vocational Learning Opportunities

4.1.1. Introduction

The rapid advancement of the development of virtual reality (VR) technology quickly expanded beyond the entertainment sector and now holds promise to revolutionise the education sector as well. While a growing body of VR projects exist for primary and secondary education (eg, Morales et al., 2013; Nobrega & Rozenfeld, 2019) as well as university level education (Dyer et al., 2018), less focus has been placed on the use of VR in vocational education. While this varies between countries, in Germany the vocational sector remains one of the main employers for young adults. The German dual vocational education system (VET) interweaves hands-on practice in the industry with decentralised education. The education is mostly reliant on frontal teaching in dedicated VET schools mixed with practical hands-on experience at the place of work (Baethge et al., 2007). Contrary to universities and trade schools, vocational institutions are often far smaller in size, with nearly half of the institutions employing <50 people (Autorengruppe Bildungsberichterstattung, 2018). The small size of the individual institutions combined with the variety of vocational jobs on the market pose major challenges to the development and introduction of expensive equipment such as VR-ready learning environments. The push for digitalisation does not necessitate the exclusive use of VR technology however and alternatives should also be considered.

With the greatly accelerated push for remote learning opportunities caused by the COVID-19 outbreak, it is crucial to bring modern teaching techniques already present in other educational systems to vocational education as well. In this paper, the focus will be more specifically on the material generated for the use of vocational education in car mechanics. Given the smaller size of many institutions and the variety of topics and skills covered in vocational training, additional points need to be considered, such as alternatives to fully immersive VR environments.

The development and maintenance of fully interactive VR environments proves difficult given the variety of tasks a car mechanic has to learn over the course of their three-year long education. As one potential solution to the above, non-interactive VR environments generated from 360° video recordings of lessons should be considered. We understand non-interactive VR as an environment requiring only 3° of freedom (3-DoF, head movement on the x and y as well as body turning) without controllers, for example a 360° video viewed on a Samsung Gear VR headset. On the other end of the spectrum are fully interactive 6-DoF environments featuring environmental manipulation such as surgical training programs (e.g. Frederiksen et al., 2020) or human-in-the-loop driving simulators (Reinhard et al., 2019). Due to higher degrees of freedom provided by controllers and room-scale tracking, 6-DoF environments are usually created with interactivity in mind, which holds great promise in future interactive vocational learning environments, since experienced immersion will be higher (Slater, 2003, 2018). However, a major downside is the complexity of content creation leading to a current scarcity in content (Jensen & Konradsen, 2018).

Previous research highlighted the potential of educational 360° videos to increase student engagement (Violante et al., 2019) and leading to similar learning outcomes as traditional face-to-face or 2D video-based education (Ulrich et al., 2021). At the point of writing, the technology to have multiple points of view within a 360° video is still mostly hypothetical, although this could change in the near future (see e.g. Jeong et al., 2020) Compared to fully interactive VR environment, 360° video recorded material is more cost effective for vocational institutions since they can simply be played on aforementioned 3DoF head-mounted display (HMDs) such as the Samsung Gear, Oculus Quest or the Google Cardboard instead of requiring expensive room-based 6-DoF setups.

4.1.1.1. Immersion and presence

In addition to the level of interaction, other concepts need to be considered in VR research as well. The first is the level of immersion and sense of presence in

virtual environments. While these concepts are often used interchangeably (Grassini & Laumann, 2020), immersion is commonly seen as more related to technological aspects of VR, such as the field of view or realistic audio environments (Slater, 2003, 2018). Slater (2018) noted that immersion should be seen on a scale instead of a binary immersive/non-immersive decision, where contingency between actions of the participant and its effect in the world correlate – for example, being able to manipulate an object or observe it from multiple angles (see Huang et al., 2021) or maintaining eye-contact with an interactive avatar (see Albus et al., 2021) should lead to higher feelings of immersion. Presence on the other hand is a mental state based on a more subjective experience—the 'feeling of being in' a virtual environment (Grassini & Laumann, 2020; Schubert, Friedmann, & Regenbrecht, 1999, 2001). This subjective experience is commonly measured using physiological correlates such as EEG, fMRI or the galvanic skin response, or a variety of questionnaires (for an overview, see Grassini & Laumann, 2020). In the context of this study, subjective sense of presence was measured by using one of the standard questionnaires, the iGroup presence questionnaire (IPQ, Schubert et al., 2001). The sense of presence measured by the IPQ is primarily focussed on the individual's perception of being part of the virtual environment (Grassini & Laumann, 2020), and the present paper will use of the term 'sense of presence' in the same way.

4.1.1.2. Novelty effect in VR

One relatively recent notion in VR-based research, especially in the context of education, is the novelty effect of the utilised software and hardware. At the time of writing, a significant portion of students in both vocational facilities as well as the university reported having little to no prior experience with VR, with direct exposure usually limited to large events such as fairs. This led to high motivation to inquire about VR as well as participate in any VR-related research, irrespective of content. Whether this high motivation leads to a sustainable effect of VR-based learning is under debate. Merchant and colleagues (2014) reviewed a total of 69 studies utilising different forms of virtual learning environments in higher education and found that while virtual environments increased learning outcome, the effects started to deteriorate with repeated exposure to the virtual environment. Huang (2020) noted that this could be attributed to the diminishing novelty effect and added a first line of research where longitudinal learning in 3-DoF and 6-DoF VR was investigated. Notable results from Huang (2020) include that, while novelty alone did not increase learning outcome, it decayed slower than presumed by

Merchant and colleagues (2014) and novelty also positively influenced learner's motivation. In a recent study, Huang and colleagues (2021) reported that student engagement and immersion remained high throughout several sessions, with task performance being comparable between groups using 3-DoF ('moderate immersion') and 6-DoF ('higher immersion') setups. This highlights the viability of less interactive presentation of content such as 360° videos for learners.

4.1.1.3. Cognitive Load

The impact of VR on cognitive load has been under investigation since the technology began to become more widespread. Cognitive load as a construct can be measured in a variety of ways, with correlates ranging the subjectively experienced load measured by questionnaires such as the NASA task load index (TLX, Hart, 2006; Hart & Staveland, 1988) to physiological correlates such as electrodermal activity and changes in pupil size. For an overview of measures for cognitive load specifically in VR see Armougum et al. (2019). Due to the novelty and complexity of VR-based systems in educational contexts correlates introduced by VR could be highly relevant for cognitive load research as well since the danger of cognitive overload needs to be considered (Albus et al., 2021). In a surgical training context, Frederiksen and colleagues (2020) found that novices in an immersive virtual environment using a VR headset were subject to significantly higher cognitive load and consequently worse task performance compared to training with a conventional virtual setup using a screen and joysticks.

When investigating differences in expertise for a spatial navigation task, however, the physiological and subjective measures of cognitive load were mostly affected by expertise, with no differences between real-world and virtual navigation (Armougum et al., 2019).

A looming question is whether the introduction of VR content could prove to be a beneficial addition for vocational learners and learners in general. Empirical evidence in this respect is mixed. There is a general agreement about the potential of VR technology as a viable catalyst to improve education by digitalisation (Freina & Ott, 2015). A recent meta-analysis by Moro and colleagues (2021) found comparable effects of VR training to traditional methods in four reviewed studies (Moro et al., 2021). Chen and colleagues (2018) investigated the relationship between the use of VR technology and different factors of learning in automotive vocational students and found positive effects of VR on both learning satisfaction and learning outcome. On the contrary, a recent review by Jensen and Konradsen

(2018) about the state of VR in education indicated that, in the majority of reviewed papers, the use of VR environments yielded no advantage in either the cognitive or psychomotor domain. Jensen and Konradsen highlighted that one of the biggest shortcomings of educational VR remains the scarcity of content. This shortcoming also extends to the German vocational education system, where traditionally craftsmen with years of professional experience are teaching the next generation with traditional methods such as frontal classes and hands-on lessons, mixed with some older educational 2D videos (Baethge et al., 2007). At the time of writing, no affordable 'out-of-the-box' solution for generating VR content on a large scale exists outside of research projects and single educational projects. However, as noted before, recent research by Huang (2020), Huang and colleagues (2021) and the meta-analysis of Moro and colleagues (2021) showcased the viability of 3DoF setups for learning environments, which would be much easier to implement compared to full VR environments which often are made from scratch. Apart from the learning outcome, it is however also important to look at other factors of the learning experience, such as where the learner shifts their attention to.

Due to recent technological advances, eye-tracking became feasible in VR setups. This allows the comparison of participants' gaze pattern in the real versus the virtual world as well as a more direct measurement of attention through gaze tracking. So far, there is only a limited body of research investigating the effects of different learning modalities using eye-tracking.

Reichenberger et al. (2020) note a possible for gaze tracking as a measure for attention in clinical treatments for social anxiety disorders, whereas Cheng and Huang (2012) investigated the possibility of VR to foster joint attention in children with pervasive developmental disorder. In educational contexts, a study by Meppelink and Bol (2015) concluded that participants who were not knowledgeable about a topic showed higher recall of information related to that topic when they spent more time (total fixation duration) on picture-based information compared to text-based only. Therefore, for this study, we recruited participants with little to no previous knowledge on the educational content presented.

Whereas, however, the learning material used in Meppelink and Bol (2015) was website-based, this study focussed on the differences between traditional 2D-video-Based versus 3D-VR- video instruction.

4.1.1.4. Research question & hypotheses

The learning material used in this study was primarily visual with added verbal instructions from the instructor. Learners are therefore expected to pay attention on the relevant objects and areas in the instructional video in both the 2D video and nVR conditions by fixating on them. According to Yarbus (1967), both the average duration of single fixations as well as the number of fixations on each object or area of interest are ways to measure human visual attention. By an extension of this, the sum of the durations of all fixations, or total fixation duration, on a single area of interest reflects the amount of time an individual actively paid attention to an object or person in a video. Bhoir et al. (2015) and Jeelani et al. (2018) used these measurements, among others, to investigate the attention of construction workers and the handling of personalised safety instructions, respectively. Instead of presenting learners with different versions of video-based content, this study will focus on strictly on differences stemming from video presentation—whether instructional content leads to differences in visual attention and learning outcome when it is presented as a 360° video on an HMD compared to the same content presented as a 2D video on a tablet. Previous research highlighted different factors influencing learning outcome and learner engagement such as novelty (Huang, 2020, 2021), immersion and sense of presence (see Grassini & Laumann, 2020), this study aims to investigate learner's attention in different video media.

The primary research question this study aimed to answer was whether the presentation modality of a non-interactive 3DoF 360° video (nVR) compared to a 2D video of the same content influences the learners' visual attention. It is predicted that there are differences in attentional focus, indicated by differences in total fixation duration on areas of interest.

Additionally, this study will investigate whether the presentation of content in form of an educational 360° video on an HMD will improve the learning outcome compared to presenting the same content in form of a 2D video on a tablet. If that is the case, participants watching educational nVR content using a VR headset will show increased task-relevant performance, indicated by a higher test score in a standardised knowledge recall test compared to participants who watch the same video in 2D using a conventional tablet.

Lastly, since an immersive 360° video is presented on an HMD in one condition, differences in immersion and sense of presence are predicted. This serves as a manipulation check and will be indicated by higher self-report scores on a

standardised presence and immersion questionnaire, with participants in the nVR condition reporting increased sense of presence compared to the participants of the 2D video group.

4.1.2. Method

4.1.2.1. Participants

A total of 50 participants, all students at the University of Kaiserslautern, Germany, aged 20–34 (*M* = 25.28; *SD* = 2.53), were invited to be tested individually in a VR cabin and randomly assigned to either the 2D video or the nVR group. None of the participants reported prior knowledge or background of car mechanics. No participant reported dizziness or other symptoms during the experiment, which would have warranted the termination of the experiment.

Two participants from the 2D video group had to be excluded due to technical problems or equipment failure during data collection. Therefore, the final sample consisted of 48 participants: 23 (4 female) in the 2D video group, aged 20–30 (*M* = 25.13; *SD* = 2.14) and 25 (9 female) in the nVR group, aged 21–34 (*M* = 25.36; *SD* = 2.96). All participants had normal or corrected-to-normal vision. Participants received course credit as compensation for their participation.

4.1.2.2. Material

4.1.2.2.1. Questionnaires

A set of standardised questionnaires were deployed to assess different aspects of participants' introspection and a standardised test was used to assess task-related performance. Cognitive load was measured using the NASA TLX (Hart, 2006; Hart & Staveland, 1988). This questionnaire consists of six items measuring different aspects of cognitive load experienced while performing different tasks, for example: ‘How mentally demanding was the task?’ Each item is rated on a 21-point scale ranging from 0 to 20. In this study, the TLX was used to investigate potential differences in the mental effort participants experienced in the 2D video versus the nVR group. The TLX was chosen due to its widespreaduse as a quick, standardised measure of subjective cognitive load (see e.g, Armougum et al., 2019).

To assess the sense of presence, the German version of the IPQ (Schubert et al., 2001) was administered after the educational video in printed form. Due to the non-interactive nature of both video conditions, the version used in this study contained 13 out of the original 14 items, with the item ‘How much did your

experience in the virtual environment seem consistent with your real-world experience?' excluded. All questions were answered by participants on a six-point scale ranging from 'Completely disagree' to 'Completely agree' or contextual variations thereof.

Knowledge acquisition was assessed using a 16-item single-choice test which was created in cooperation with a vocational institute in North Rhine-Westphalia, Germany. The test was held in close similarity to written vocational exams conducted as part of the curriculum for car mechanics and was designed to assess acquired task specific knowledge based on the information presented in the instructional video. Each item had a single correct (target) and three incorrect answers (lures). One point per correctly marked answer and per correctly unmarked answer was given, leading to a minimum of 32 and a maximum of 64 points. Before the study, the test was given to seven naïve participants with no prior knowledge of car mechanics, and it was evident that the test needed mechanical education to reach a sufficient score.

4.1.2.2.2. Video material & presentation

The 2D and the 360° video were both recorded in a mechanics garage of the partnered vocational institute for the purpose of this study. The videos were recorded using a double camera setup with a 360° camera on a tripod and a conventional camera fixated below to create equal points of view for both the 2D and 360° videos. While working, the instructor regularly faces the camera to provide instructions and in addition also verbalises the current step he is working on. The content of the video was identical between the 2D and 360° versions, with the difference that the 360° video allowed to survey the entire workshop, although no other activity or distraction took place during recording. Content represents one full exempt of the vocational education curriculum, showing the disassembly and reassembly of an exemplary Otto engine to switch out the intake bridge. An experienced vocational instructor demonstrates all work steps one-by-one while giving additional verbal instructions, such as the advice on tool usage. Total video length was 9 min, 52 s for both videos.

The presented 2D video was recorded in a resolution of 3,840 × 2,160 pixels at 30 frames per second. It was presented on a fifth generation Microsoft Surface Pro tablet (Microsoft Corp., Redmond, WA) with a resolution of 2,736 × 1,824 at an aspect ratio of 3:2. The tablet was placed in front of the participant ~75 cm away from the face with the screen at an angle of 65°.

For the nVR condition, a 360° video was recorded in a resolution of 7,680 × 4,320 at 60 frames per second, downsized to 3,840 × 1,920/60 for presentation during the study. The HTC Vive HMD (High Tech Computer Corporation, New Taipei, Taiwan) used for presentation was a tethered 2016 version and connected to a laptop with a NVIDIA GeForce GTX 1,060 (Santa Clara, CA) via HDMI cable. The HMD had a resolution of 1,080 × 1,200 pixels per eye (2,160 × 1,200 combined) at a refresh rate of 90 Hz with a 110° field of view.

4.1.2.3. Apparatus

For measuring eye movements in the 2D-video group, Tobii Glasses 2 (Tobii Pro, Danderyd, Sweden) were used. This binocular eye-tracker records eye movements from both eyes with a sampling rate of 100 Hz using two cameras and six infrared illuminators per eye.

The recorded gaze patterns are later superimposed on a video recording captured by an integrated scene camera. The camera recorded 25 frames per second with a resolution of 1,920 × 1,080 pixels and a FoV of 90°. For measuring eye movements in the 3D-video group, the VR implemented eye-tracker Tobii Pro VR Integration, a retrofitted 2016 version of the HTC Vive, was used. The integrated eye-tracker recorded both eyes with a sampling rate of 120 Hz using one eye-tracking sensor and 10 infrared illuminators per eye. The trackable field of view for eye movements is 110°.

Eye-Tracking data in the 2D condition was recorded using the Tobii Glasses Controller v1.108 (Tobii AB, Danderyd, Sweden); 360° data collection was conducted in Tobii Pro Lab 360VR v.118. All eye-tracking data analyses in both groups were conducted using Tobii Pro Lab 360VR v1.118. Statistical analyses were conducted primarily using IBM SPSS 25 (IBM, Armonk, New York). Calibration was done via Tobii Pro Lab VR's integrated five-point calibration for participants wearing the Tobii Pro VR Integration, while participants wearing the Tobii Glasses 2 completed the single-point calibration provided by Tobii Glasses Controller's single-point calibration. Calibration was successfully completed by all participants described in this study.

4.1.2.4. Design

A single-factor between-subjects design was used (2D video-based vs. non-interactive 360° [nVR] instruction). Participants were randomly assigned to either experimental condition.

4.1.2.5. Procedure

Participants were individually invited to the lab. Upon arrival, participants were greeted by two experimenters, shown the lab before receiving information about the purpose and the procedure of the study from the experimenters. There was sufficient time given to ask questions. Afterwards, the participants were then asked to give informed consent and to fill out a brief questionnaire concerning their demographic information as well as their previous experience with modern technology including VR, and car mechanics. Participants of the 3D-video group were then shown the VR equipment, while participants of the 2D-video group were shown the head-mounted eye-tracker before being both groups were seated and received task-specific instructions. These instructions did not differ between the groups. When the participant had understood the instructions and any open questions were adequately answered, they were asked to put on the VR headset or eye-tracker and calibration was initiated.

After completing calibration, the participant watched the video assigned to his or her experimental group. Time-stamped markers were recorded when participants started and finished the video – this happened automatically in the VR group and remotely in the 2D video condition. Since for the 2D-video group time-stamped recordings were collected from all participants, accurate markers were added on an individual basis for analysis. When participants finished watching the educational video, recording of the eye-tracking was stopped. The experimenters then helped the participants to take off the eye-tracker or the VR HMD before handing them the second set of questionnaires. Participants were then fully debriefed and received course credit at the end of the study. This study was approved by the Ethics committee of the University of Kaiserslautern (Application #72018).

4.1.3. Data Analysis

4.1.3.1. Selection of areas of interest

The video material presented in this study was recorded in the context of a larger project on digitalisation in the vocational education of car mechanics. The videos showcase a standardized step-by-step guide on how to disassemble a common type of engine, where each step is shown and explained by an expert. The content of the videos would usually be demonstrated in person by vocational instructors; therefore, the areas of interest were able to be pre-defined based on expert opinions from these vocational instructors.

The scene of the video is static in both conditions, while head movement is possible in the 3D video the participants are instructed to focus on the content, which is placed directly in front of them when the stimulus presentation starts. There were four relevant stimuli in total: The engine block in the centre of the video, the instructor who was positioned either to the right or left of the engine and hence represented by two areas of interest, a table with tools on the right side as well as an assembly trolley to the left where spare parts were placed. The engine was mounted on a carriage and remained stationary throughout the recording and was defined as the first primary area of interest. To account for movement throughout the video, two counterbalanced areas of interests were defined for the instructor – Instructor (Left) and Instructor (Right). At the beginning of the video, the instructor stands to the right of the engine but moves over the course of the video between both spots.

In line with the expert opinions of the collaborating vocational instructors, it was assumed participants would focus primarily on the instructor and the engine. The table with tools was seen as potentially relevant due to the instructor's explanation which tools are utilised in each work step. The assembly trolley was included as a secondary area of interest since the instructor places spare parts throughout the disassembly on it, although this happens rather infrequently.

After initial data screening, one of these five areas of interest was excluded, which was the assembly trolley. No participant spent more than a few seconds on it and further analyses suggested most fixations on the trolley were visiting fixations or glances, with few full fixations overall. Thus, the four other areas of interest were analysed in this study: The instructor in two positions to account for movement over the course of the video, the engine, and the table. An exemplary frame of the 360° video, which shows all areas of interest as well as the full 360° video environment, can be seen in Figure 1. The areas of interest and field of view in the 2D group were similar to the lower picture in Figure 1.

4.1.3.2. Eye-tracking data acquisition

In the 2D video group, automatic mapping on snapshot basis as provided by Tobii Pro Lab was conducted individually for every participant using a single keyframe chosen from each recording. The keyframe for each participant was marked with pre-defined areas of interest (AoI) based on the relevant objects in the presented video as described above. To conduct AoI-based analyses, each participants' recording was rendered over the chosen keyframe. Gaze data and pupil

data were also recorded but not analysed for this study. In addition to the automatic mapping, participant recordings were checked at least once per second to identify potential automated mislabelling of data.

For the 3D-video group, the data acquisition was conducted similarly with the same selection of areas of interest defined a priori but in addition the 360° video was pre-rendered before conducting the experiment since there were no individual differences in viewing angle or distance due to the use of a stationary placement of the participant in the video. In effect, this pre-rendering enables automatic mapping of participant gaze patterns onto areas of interest. Similar to the 2D video group, recorded eye-movements of each participant were superimposed within pre-determined temporal markers onto the 360° video. As with the 2D recordings, participant recordings were checked at least once per second to identify potential automated mislabelling of data.

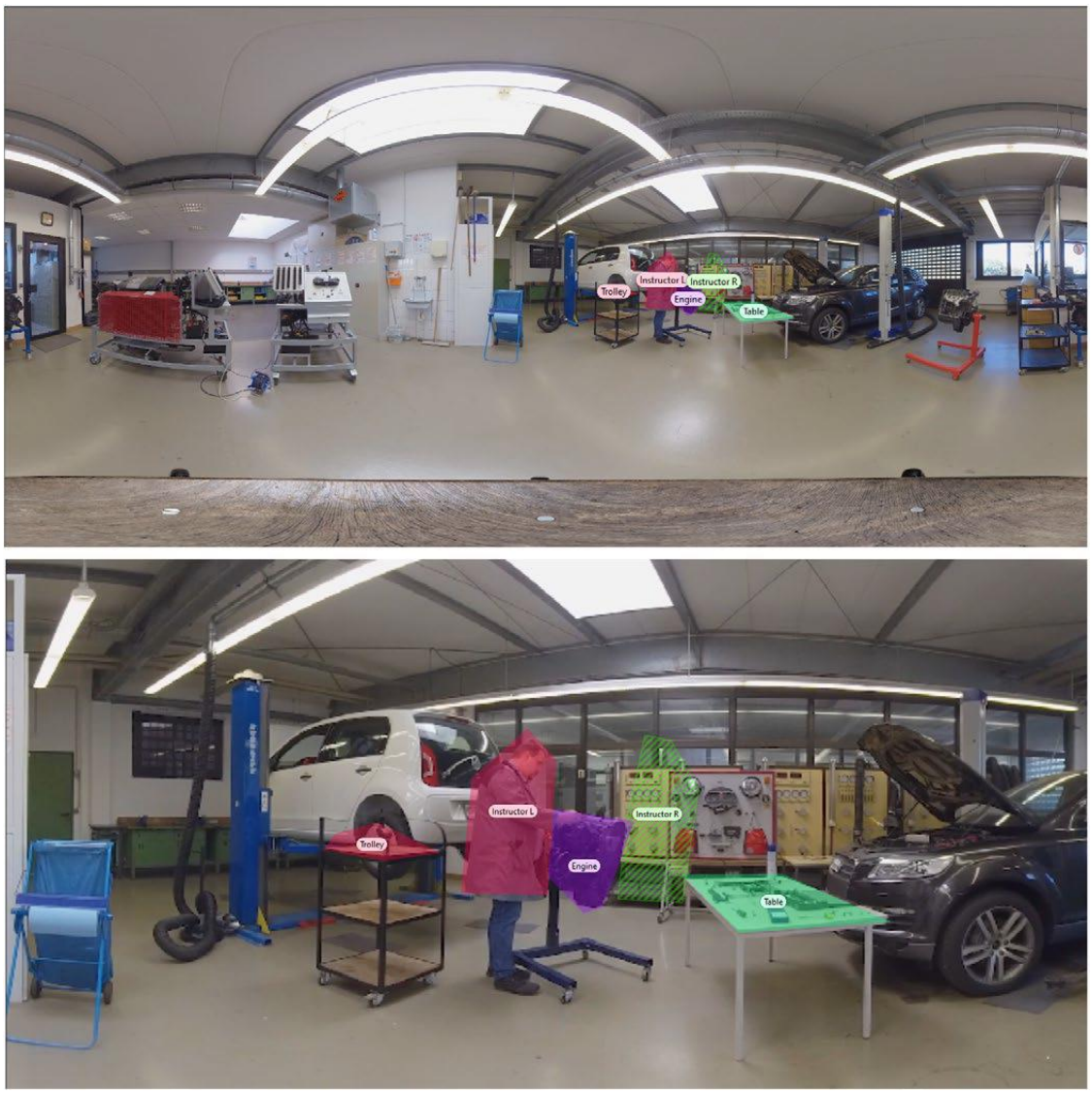

Figure 1: Snapshot of the 360° recording with added areas of interest in full 360° view (top) and focal area (bottom)

4.1.3.3. Data quality

The overall quality of the collected data was satisfactory in both groups. One participant had to be excluded due to a potential calibration issue which resulted in no AoI-based fixations despite otherwise normal gaze patterns. One participant had to be excluded due to a battery-related problem which resulted in an incomplete recording. All other recordings from both groups were within the expected parameters. In two participant cases, recordings were split into two or three parts by the automatic mapping algorithm, further investigation revealed no anomalous eye-tracking data, both recordings in question were complete. It is presumed that these interruptions might have been caused by participants closing or averting their eyes for a few seconds.

It should be noted that presentation of the 360° video was hard-coded in a self-running environment within the Tobii Pro Lab software, whereas the 2D-video was manually started by one experimenter, after manually starting the recording of the eye-tracking glasses. Therefore, while video length was held constant between groups, the recording duration slightly varied between the groups and within the 2D video group. Overall data matching was working as intended with a single keyframe allowing gaze mapping on average 97.9% of recording duration in the 2D group, and 99.9% in the VR group due to the automatic rendering as described above. Figure 2 shows the distribution of recording durations for the 2D group (N = 23) and the VR group (N = 25).

The classification of recorded eye-tracking data was done in accordance with Tobii Pro's white paper on the fixation filter in Tobii Pro Lab (Olsen, 2012; Tobii Technology, 2012). This study used the Tobii I-VT fixation filter with two changes—the addition of a gap fill-in interpolation with a maximum gap length of 75 ms to account for losses of data due to technical issues of up to three frames (see Olsen, 2012). The second change was setting the threshold for maximum gaze velocity calculation from 30°/s to 20°/s since participants in neither experimental condition were expected to make sudden eye-movements (see Olsen, 2012). This change was made to decrease false-positive merging of adjacent fixations. All other settings were kept in line with Tobii recommendations, notably fixations shorter than 60 ms were excluded from analysis and adjacent fixations were merged when they were <0,5° apart (Tobii Technology, 2012).

4.1.4. Results

4.1.4.1. Manipulation check: Sense of presence

There was one primary presumption which had to be accounted for. HMD-based VR is seen as one of the most immersive experimental modalities, therefore participants in the 3D-video group, using HMD, are expected to report more immersion and feelings of presence (Schubert et al., 1999, 2001) compared to the 2D-video group using tablets. The manipulation check hypothesis was therefore that there are differences in perceived presence, indicated by higher self-report scores on the IPQ, with participants in the 3D-video group reporting increased presence compared to the 2D-video group. To test this hypothesis, a univariate analysis of variance (ANOVA) with the factor 'Group' and the dependent variable 'IPQ Grand Mean' was conducted. Levene's test was not significant ($F(1, 46) = 1.54$, $p = 0.220$), indicating equal error variance equal across groups. The IPQ Grand Mean was calculated as the grand average of all 13 used items of the adapted German version of the IPQ used in this study. An unrotated factorial analysis was conducted and results were mostly in accordance with the pre-existing three factor structure in the original questionnaire. There was a strong tendency towards a single factor solution which explained 44.69% of total variance.

To investigate levels of immersion and presence reported by participants, this single factor solution with an Eigenvalue of 6.26 was used. The ANOVA revealed a significant difference ($F(1, 46) = 14{,}75$, $p < 0.001$, $\eta^2 = 0.243$) in perceived immersion and presence, with the nVR group reporting significantly higher scores on the IPQ ($M = 4.03$, $SD = 0.65$) compared to the 2D group ($M = 3.18$, $SD = 0.87$). The results are displayed in Figure 3.

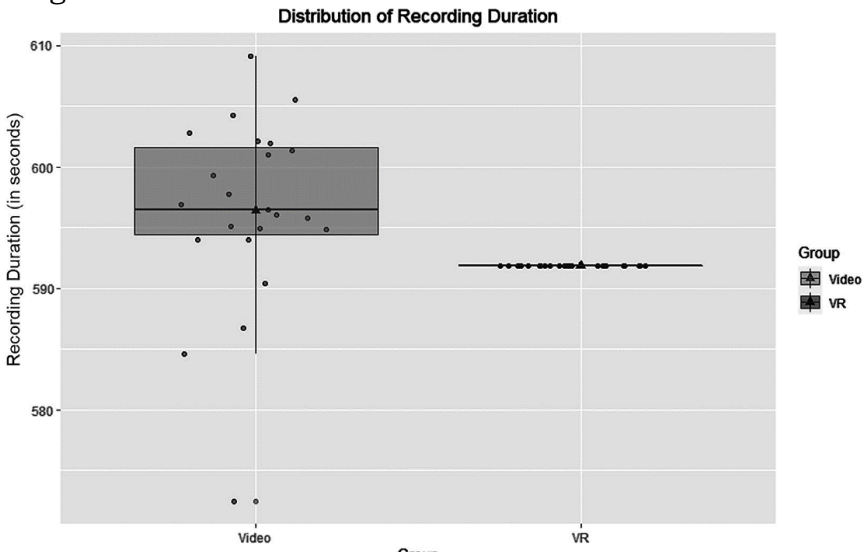

Figure 2: Distribution of recording durations, separated by group. Group mean indicated by triangle marker. Error bars indicate standard deviation.

4.1.4.2. Hypothesis 1: Differences in AoI-based total fixation duration

The primary research question this paper aims to answer was concerned with the attentional focus, indicated by differences in total fixation duration on relevant areas of interest, between the two experimental groups. It was assumed that there should be attentional differences between HMD-based 360° and traditional video-based learning. For the purpose of this study, the total fixation duration within pre-defined areas of interest is seen as representative for attentional focus (see eg, Bhoir et al., 2015; Jeelani et al., 2018; Yarbus, 1967). As the literature so far provides evidence in either direction and the viability of eye-tracking in educational VR is relatively novel, this hypothesis was undirected. For every area of interest, a univariate ANOVA with the variable 'Total Fixation Duration' was conducted. The Total Fixation Duration is the sum of all individual fixations over a given time of interest. In this study, the time of interest was the full duration of the video (T = 592 s) for each group. As detailed above, the areas of interest under consideration are the instructor, the engine block and the table with tools. Levene's tests were conducted for each ANOVA, with the total fixation duration data for the instructor not being normally distributed between groups ($F(1, 46) = 4.612$, $p = 0.037$). Therefore, a Kruskal–Wallis test was conducted for the instructor data. Both the

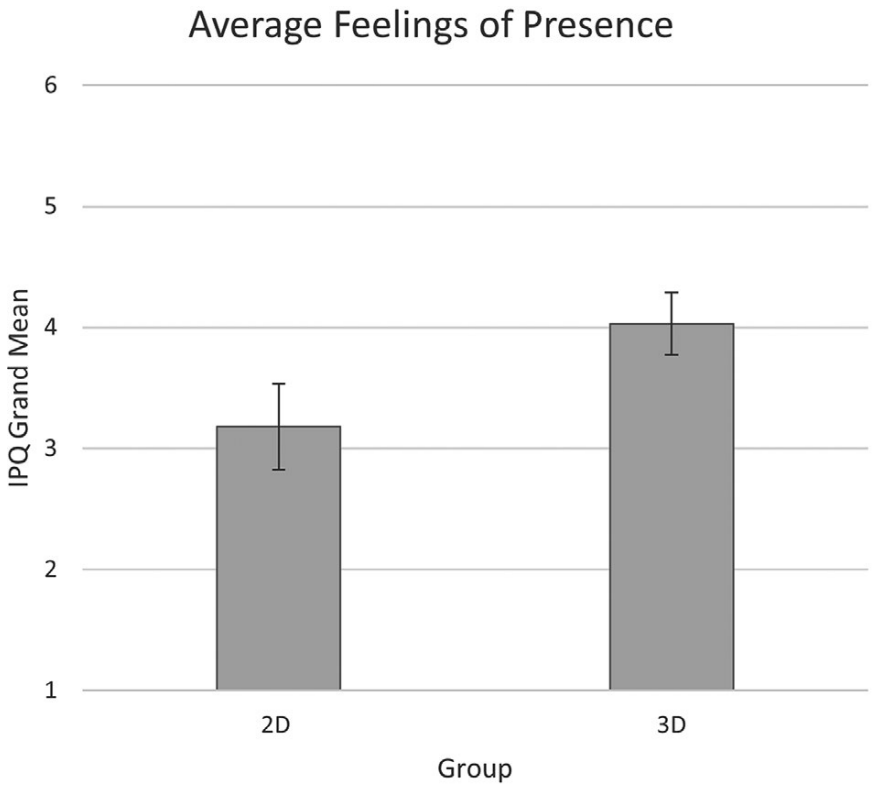

Figure 3: Averaged Grand Sum IPQ scores, separated by group. Error bars indicate 95% confidence intervals.

engine ($F(1, 46) = 3.230$, $p = 0.079$) and table ($F(1, 46) = 0.932$, $p = 0.339$) fixation data did not differ significantly from normal distribution, hence no non-parametric tests were necessary.

Part 1: Attentional focus on the instructor

To investigate whether the groups showed differences in their attentional focus on the instructor, a univariate ANOVA with the factor 'Group' and the dependent variable 'Total Fixation Duration: Instructor' was conducted. There was a significant main effect ($F(1, 46) = 89.75$, $p < 0.001$, $\eta^2 = 0.661$) in total fixation time on the instructor, with the 3D-video group fixating more on the instructor (M = 98.55 s, SD = 34.24 s) compared to the 2D-video group (M = 24.93 s, SD = 15.28 s). The results can be seen in Figure 4. A Kruskal–Wallis Test was conducted to account for the not normally distributed data. The previous results could be confirmed, there was a significant difference in total fixation duration on the instructor, $H(1) = 34.231$, $p < 0.001$. Participants in the 3D Video group (Mdn = 96.78s) fixated significantly longer on the instructor compared to participants in the 2D Video group (Mdn = 18.40s).

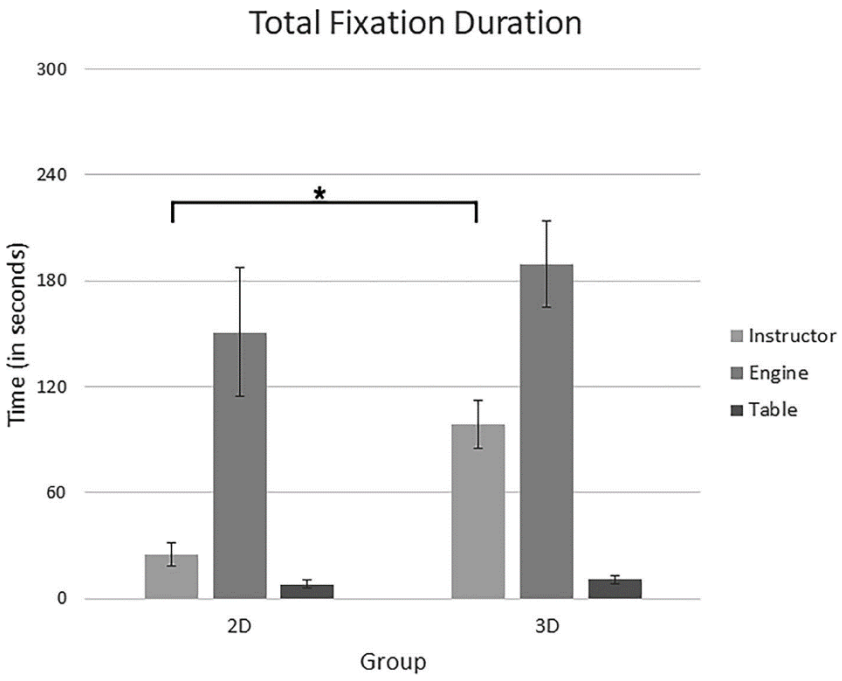

Figure 4: Sum of all fixations on areas of interest, in seconds, separated by group. Error bars indicate 95% confidence intervals.

Part 2: Attentional focus on the engine

The second area of interest was the engine block itself. A univariate ANOVA on 'Total Fixation Duration: Engine' revealed no difference between participants in the 2D (M = 150.95 s, SD = 89.81 s) and the 3D-video (M = 189.53 s, SD = 62.43 s) group in their attentional focus on the engine, $F(1, 46) = 3.03$, $p = 0.089$, $\eta^2 = 0.062$. The results can be seen in Figure 4.

Part 3: Attentional focus on the table

Due to the research question of this study, the presented eye-tracking data primarily focused on the engine and instructor. The table has been included as a minor area of interest since there is educational context related to it presented in the video. A univariate ANOVA on 'Total Fixation Duration: Table' revealed no difference between participants in the 2D (M = 8.11 s, SD = 6.09 s) and the 3D-video (M = 10.70 s, SD = 5.28 s) group in their attentional focus on the engine, $F(1, 46) = 2.50$, $p = 0.121$, $\eta^2 = 0.051$. Results can be seen in Figure 4.

In summary, Hypothesis 1 could be partially confirmed, with participants in the nVR group fixating more on the instructor, but not the engine or table compared to the 2D video group.

4.1.4.3. Hypothesis 2: Learning outcome

The first experimental hypothesis involved the learning outcome. Based on the literature review, it was predicted that participants in the 360° group should perform better compared to participants in the video condition. Performance was measured by a standardised 16 item exam. Each of the 16 items had between 1 and 4 correct responses with each correctly marked or correctly unmarked response awarding one point. Thus, the highest achievable score was 64 points.

To investigate whether the participants in the experimental conditions showed differences in the learning outcome, a univariate ANOVA with the factor 'Group' and the dependent variable 'Learning Outcome' revealed no significant differences between groups. However, Levene's test revealed significant differences ($F(1, 46) = 5.26$, $p = 0.026$) in the error variances between the two experimental groups, hence a

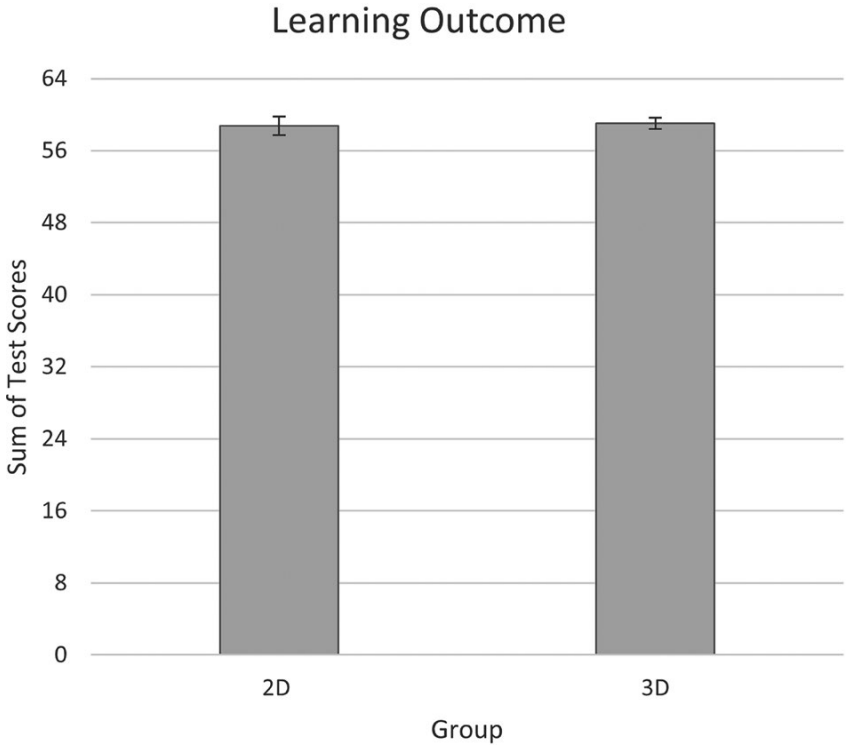

Figure 5: Comparison of test scores on the 16-item knowledge test, separated by group. Error bars indicate 95% confidence intervals.

Kruskal–Wallis test with the grouping variable 'Group' and the dependent variable 'Learning Outcome' was conducted. No significant difference was found, $H(1) = 0.063$, $p = 0.801$, the learning outcome of participants in the nVR group ($Mdn = 59.00$) did not differ from participants in the 2D video group ($Mdn = 59.00$). The means and 95% confidence intervals for each group are displayed in Figure 5, whereas the distribution of the individual values for each group is visualised in Figure 6.

In short, Hypothesis 2 could not be confirmed, no differences in learning outcome could be attributed to the learning modality.

4.1.4.4. Additional analyses: Cognitive load

In addition to the results above participants were asked to fill out the NASA TLX after watching the educational video in either group as subjective measure for experienced cognitive load.

Previous literature (see e.g, Armougum et al., 2019) found a decrease in cognitive load when participants grew familiar with navigation in VR compared to unfamiliar participants. This study recruited primarily participants who were unfamiliar with VR, therefore the TLX was used as a measurement to compare the

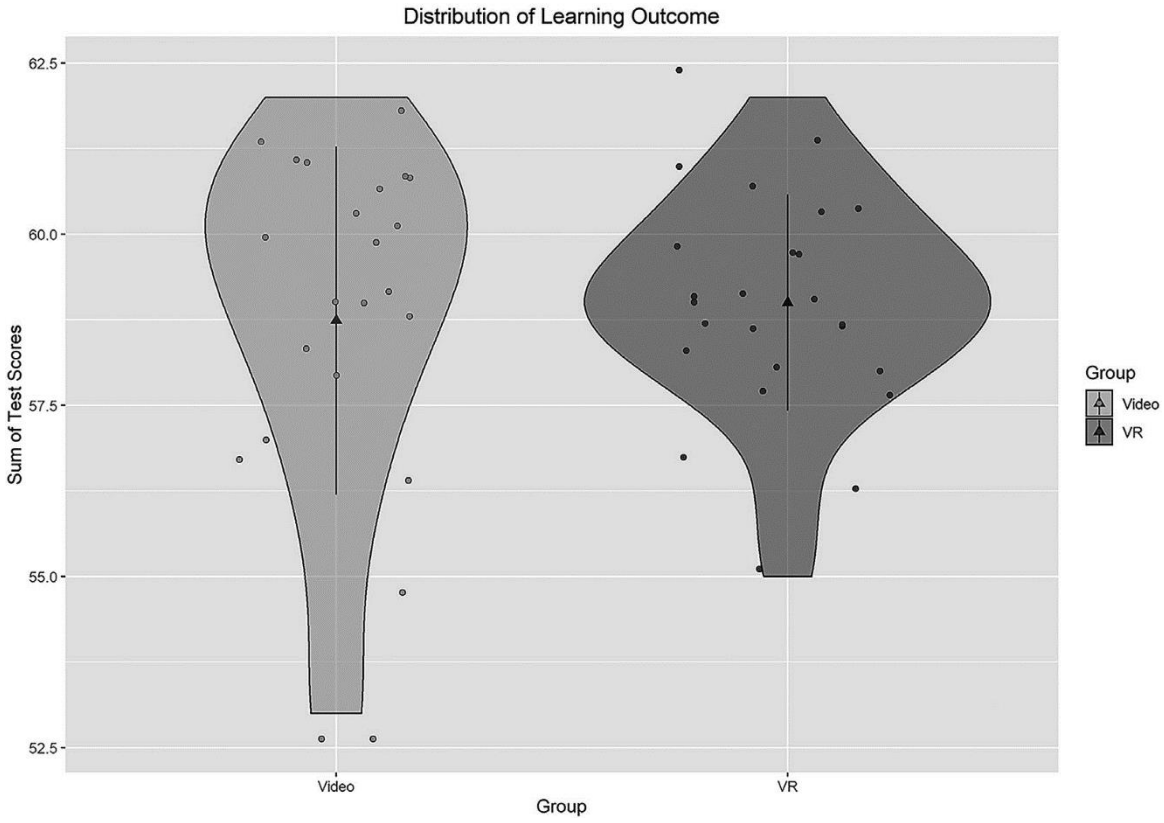

Figure 6: Distribution of test score averages on the 16-item knowledge test, separated by ghroup. Group mean indicated by triangle marker. Error bars indicate standard deviation.

subjective load of watching a 3D-video on an HMD compared to a more common task such as watching a 2D video on a tablet. To investigate whether the participants in the
experimental conditions showed differences in the subjective cognitive load, a univariate ANOVA with the factor 'Group' and the dependent variable 'Average Cognitive Load' was conducted.

Levene's test indicated no differences ($F(1, 46) < 1$, $p = 0.463$) in the error variances between the two experimental groups. The ANOVA did not reveal significant differences ($F(1, 46) = 1.477$, $p = 0.231$, $\eta^2 = 0.031$) in subjective load between participants in the 3D-video group ($M = 8{,}49$, $SD = 3{,}22$) and participants in the 2D-video group ($M = 7{,}34$, $SD = 3{,}34$). Results can be seen in Figure 7.

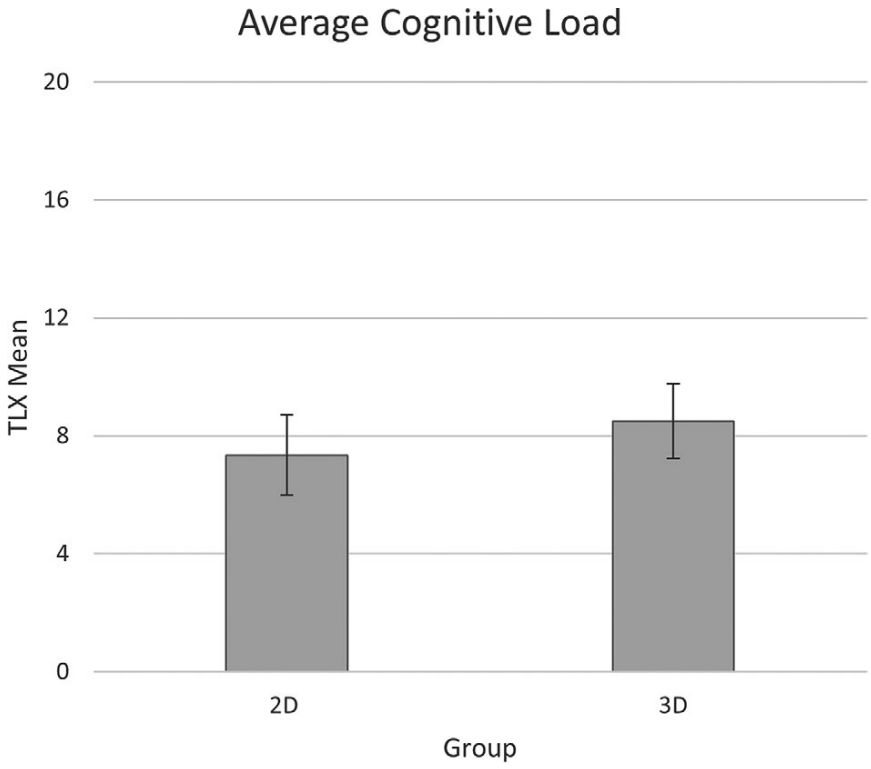

Figure 7: Averaged NASA TLX scores, separated by group. Error bars indicate 95% confidence intervals.

4.1.5. Discussion

This study investigated differences in vocational education using instructions through either a 2D video presented via tablet or a 360° video presented via HMD and found no effect of the presentation modality on test performance, but differences in total fixation duration. Therefore, the primary hypothesis was able to be partially confirmed since the effect was only significant for the instructor. The second hypothesis, which predicted a higher learning outcome in the 360° nVR video condition compared to the 2D-video, must be rejected since no difference was found. The study revealed several differences in attention between the 2D-video and nVR groups: participants in the nVR group fixated on the instructor both significantly more and significantly longer than the 2D-video group. This led to a

significantly higher total fixation duration on the instructor in the nVR group. A similar pattern was not seen in engine or table fixation patterns.

Regarding Hypothesis 2, the results were contrary to the expected outcome. Participants were able to learn content from an unfamiliar field with both a tablet and a HMD in the same way: participants in both groups performed surprisingly well based on their backgrounds and the assumed complexity of the learning material. Regardless of if they had watched the 2D or nVR video, they scored very high on the knowledge recall test. However, these findings are in line with the meta-analysis from Moro and colleagues (2021) and a recent study by Ulrich et al. (2021) found similar increases in learning outcome for education done via 360° videos, via 2D videos and via face-to-face training, which is in line with the results of this study as well.

Speaking from a strictly performance-based perspective, the results of this study show that tablets with 2D-video content are just as effective for instruction as 360° nVR videos presented on an HMD. Since tablets are much more readily available and cost-effective compared to setups utilising HMDs, 2D video-based vocational education could be a viable short-term solution until fully interactive VR content becomes accessible to a broader audience. As predicted, participants watching the 360° VR video rated the video experience as more engaging and felt more present - indicated by higher self-report scores on the IPQ scales – compared to participants in the tablet group.

Participants in the 360° video condition showed increased attentional focus on the instructor as indicated by a higher number of fixations and a longer total fixation duration on the instructor compared to the 2D-video group. This is in line with previous research (Albus et al., 2021; Birmingham et al., 2008; Rubo & Gamer, 2021), which found increased attention in VR for socially relevant stimuli, which is the case for the instructor in the presented content. Chen et al. (2018) found higher learning satisfaction using VR in automotive vocational education settings, as well as higher learning outcome. While this study did not find any increase in learning outcome compared to 2D videos, the presented results do indicate that there is higher engagement with the learning material even in non-interactive VR scenarios, which could lead to higher learning satisfaction. However, this was not investigated in this study and should be confirmed in future research.

In a short poll conducted after the primary experiments, both groups rated the importance of new educational technology as high, an opinion shared by many VR research groups and tech companies. As answer to an open question about

problems and challenges regarding the usage of VR in education many participants stated that they themselves, as the next generation of teachers, are not well prepared because it was lacking in their own education. Another issue mentioned by several participants was that there is a lack of content for VR-based education (see also Jensen & Konradsen, 2018). While this still poses a challenge for many vocational institutions, companies are emerging, which aim to provide content for different levels of education. The recent surge in remote teaching due to the COVID-19 pandemic highlighted the need for educational concepts, which are both engaging and remote. VR can fill this gap, and many companies such as edify (https://www.edify.ac/) or Labster (https://www.labster.com/) strive to create virtual laboratories or other collaborative spaces for use with VR. Others such as the WDR, a German media company provide free classroom materials (https://www1.wdr.de/schule/digital/unterrichtsmaterial/virtualreality-100.html) tailored for school children, which utilise 360° videos that could be played also on a tablet or smartphone. Without a doubt, the next years will see an increase in available materials for all levels of education, which makes the scientific investigation of their effectiveness for the learner even more relevant.

4.1.5.1. Notes on data quality in HMD-based and mobile eye-tracking research

Due to the novelty of the deployed design, it is necessary to elaborate on the utilized eye-tracking devices and their comparability between HDM-integrated and standard mobile eye-tracking devices. This is largely based on the Tobii environment used in this study. The principles can, however, also be applied to other eye-tracking systems.

On a surface level, VR provides two major experimental advantages over standard experimental settings. First, the participants' visual field is fully controllable, leading to higher experimental control and participant engagement (Bacca et al., 2015; Fox et al., 2009). Second, any head or body movement is counterbalanced by the HMD technology by default. This leads to a higher quality of the data compared to mobile eye-tracking systems. However, we will focus more on the temporal dimension of eye-tracking.

One possible flaw in this study lies in the differences between the deployed eye-tracking systems. While both systems were developed by the same company, there are differences between HMD integrated eye-tracking and standard mobile or stationary eye-trackers. One crucial difference lies in the automatic data segmentation and rendering in VR. Since the participations field of view is not only

fully observable, but also controllable by and shared across all participants, problems of data loss that might occurred in the recording and analysis of mobile eye-tracking data do not occur in VR at all. This is visible in both in the higher homogeneity of data collected by the VR-based system and also in the temporal dimension, since video presentation will be matched frame-by-frame between participants, as can be seen in Figure 2. Mobile eye-tracking introduces some significant differences compared to the traditional stationary eye-tracking paradigms. Whereas the latter requires a stationary participant, often with a headrest or even head fixation, mobile eye-tracking will often involve the participant moving their head or entire body over the course of the recording. This poses a significant challenge for area-of-interest based research, since the visual field and thus the areas of interest might move over the course of the experiment. On the other hand, mobile eye-tracking enables the investigation of learning in more natural settings since the learner can remain in the normal learning environment and is obstructed less by the recording gear, which makes it much more viable for field studies conducted in many fields of educational research.

In this study, participants were seated with their attention focussed on a screen in front of them. Therefore, some of the problems described above could be alleviated. However, while the participants were seated a fixed distance from the tablet (60 cm), screen size differences between participants due to body or head movement during data recording could not fully be accounted for despite instructions to move as little as possible.

While great care has been applied in the structuring, filtering and the analyses of all data, there are still more chances for human error to occur in manual analysis compared to fully automated, standardised analysis. However, we are confident that these potential oversights, while problematic, should have no major impact on the present results.

4.1.5.2. Limitations

This study utilised a single knowledge test administered briefly after the educational content. To strengthen potential insights into long-term knowledge acquisition, a longer delay or a form pre-/post-design would have been preferable. While this study incorporated a delay period of 15 minutes, this period might have proven insufficient to have a significantly effect on knowledge recall. On the contrary, it could be hypothesized that the learning abilities of high-performing participants such as university students are more suited to short-term recall such

as the one applied in this study - even without prior knowledge on the subject. This is, however, strictly hypothetical and warrants further investigation.

However, all participants were pre-scanned for knowledge of car mechanics and confirmed to have no former professional experience. Furthermore, the deployed knowledge test was designed by a vocational education school as part of a collaborative project and was kept close to exams taken by mechanics-in-Training as part of their vocational education. The difficulty of the knowledge test was deemed sufficiently high to exclude random guessing as a successful test strategy. In this light, the apparent ceiling effect seems surprising. However, it should be noted that the subjects in this study were undergraduate students enrolled at a public university compared to the usual target group of such exams who rarely have university-level education backgrounds. As no comparative study was conducted with entry-level vocational students, generalisation at this point is difficult and further research with the target group is necessary. While this study incorporated a delay period of 15 minutes, this period might have proven insufficient to have a significantly effect on knowledge recall. On the contrary, it could be hypothesized that the learning abilities of high-performing participants such as university students are more suited to short-term recall such as the one applied in this study – even without prior knowledge on the subject. This is, however, strictly hypothetical and warrants further investigation.

4.1.5.3. Implications for further research & outlook

One crucial area of education where VR holds promise is in transferring knowledge from the theoretical to the practical domain. Previous research indicated that transfer learning is possible utilising VR environments, for instance for driving performance (Wallet et al., 2009) and in school educational knowledge acquisition (Meyer et al., 2019; Parong & Mayer, 2018). So far, however, there are only a few studies which investigate learning effects of VR on real world scenarios. At present, the common approaches seem more focussed on knowledge recall than actual task performance. This is possibly due to the increased complexity in study designs that would need to utilise both virtual and real-world facilities. Nevertheless, it would be a promising approach to combine VR-based training scenarios with the follow-up real-world task performance. This could yield promising insights into the capability of VR to prepare vocational students for their later jobs in the real world.

In addition, this study showcases the possibility to apply eye-tracking methodology to VR research, which opens a variety of possibilities for further

research on attention, cognitive load and other correlates of eye-tracking data which can now be collected within VR. The speed of technological progress likely will not slow down in the near future, at the time of writing 6DoF VR HMDs with integrated eye-tracking are already available at a consumer level in form of the HTC Vive Pro Eye (High Tech Computer Corporation, New Taipei, Taiwan). From the content side, both private and state-directed content providers for 360° video and fully immersive virtual content begin to become more commonplace all over the world. The authors hope that the present paper offers a first insight using eye-tracking into the differences and similarities between traditional educational 2D and 360° videos.

4.1.6. Acknowledgements

We want to express our gratitude to Mr. Udo Petruschkat and the Berufsbildungszentrum Märkischer Kreis (bbz), a vocational education institute in Iserlohn, Germany, for their assistance in producing of the instructional videos, willingness to share their facilities as well as providing this study with exam questions. Furthermore, we want to thank Mrs. Julia Herzhauser and Mr. Felix Assel who assisted in data collection as part of their Bachelor thesis.

4.2. Spatial Sound in a 3D Virtual Environment

4.2.1. Introduction

The feeling of being “immersed” in virtual environments (VEs) has long been an essential element in the design of user experiences for developers and content creators (Grassini, Laumann, & Skogstad, 2020). Virtual environments are immersive when they afford perception of the environment through sensorimotor relationships that mimic our natural existence (Slater & Sanchez-Vives, 2016). The degree of immersion depends on various factors of the visual experience such as the field of view, display latency, display resolution and also the number of other sensory modalities available in the virtual environment. For example, a slowly updated display is less immersive than one that can catch up to the speed of our head movements. In its simplest form, an immersive VE has both visual and auditory modalities (Slater & Sanchez-Vives, 2014; Serafin, Geronazzo, Erkut, Nilsson, & Nordahl, 2018).

Immersive sound in a VE can be achieved by incorporating environmental sounds, sounds of our own actions and simulation of the acoustics of the environment, which will affect the perceived sound (Serafin, Geronazzo, Erkut, Nilsson, & Nordahl, 2018). Sound can be incorporated into a VE as simple stereo sounds or as spatial sounds, which render real-world cues such as sound reflections and acoustic changes due to body movements, resulting in the virtual experience being perceived as more authentic. Spatial sounds not only increase the feeling of presence or ‘being there’ but also elicit more head and body movements from the user on account of being more immersive (Nordahl & Nilsson, 2014). However, spatial sounds are complex and both acquisition and reproduction are demanding in terms of the equipment, effort and expense involved. Binaural sound is sound that is perceived as being present in a specific location in space -distance, elevation and azimuth. As the name suggests, it is achieved by simulating how the sound reaches each of our ears. Finding where a sound originates - sound localization, is essential to veridical perception of an environment. High fidelity in spatial sound rendering is uncompromisable since conflicting visual information could interfere with sound localization (Jackson, 1953), as commonly seen in the capture effect or ventriloquism effect (Bertelson, 1998; Alais & Burr, 2004).

Serafin and colleagues (2018) describe ‘ear adequate’ headphones, individually administered binaural signals, head movement tracking and room acoustics as some of the requirements of a spatial soundscape to ensure high fidelity in the audio-visual environment. The position of the sound source, the

position of the receiver (user), the individual ear and head properties of the receiver, the positions of other objects in the environment and the acoustic properties of the room can all together be used to generate an individualized soundscape. The level of complexity in the type of soundscape incorporated into the VE is application dependent (Larsson, Väljamäe, Västfjäll, Tajadura-Jiménez, & Kleiner, 2010). Therefore, it is more pragmatic and economical to use spatial sounds only when they are effective and add value to the specific VE. For example, spatial sound may not be essential for a virtual lesson with an instructor speaking, whereas it will be an advantage in a table tennis training environment, where auditory feedback will improve gameplay.

In an investigation of the efficacy of different sound types in a 3D VE, Høeg and colleagues (2017) used a visual search task, where the participant was asked to search for a specific visual target randomly positioned in a scene. To assist the user in finding the target, a sound was played to indicate the target location. This auditory cue is akin to a friend calling our name from a crowd, which would help us find them more easily. In this study, the effect of different auditory cues on participant reaction times (RT), that is time to search for the target in the scene, was measured. The authors found that binaurally presented cues facilitated RTs more than stereo cues or the absence of cues by being spatially and temporally synchronous with the visual elements of the display. This finding is significant because it essentially shows that sound localization was better with binaural audio in a virtual environment. However, since the visual environment in this study was a simple 3D visual search display based on a 360° video, it is not clear whether the advantage of a binaural cue will also be present in a more dynamic virtual environment with environmental noise, which is more likely to occur in a real-world setting.

It has been observed that the introduction of noise can obscure audio cues and hinder the detection of visual stimuli (Hidaka & Ide, 2015). In recent research by Malpica, Serrano, Gutierrez and Masia (2020), the introduction of different types of noise led to a severe drop in visual detection and recognition performance in virtual reality (VR), irrespective of the type of noise introduced.

In an inverse effect, sounds went undetected under high perceptual load in the visual modality in an effect known as ‘inattentional deafness’ (Macdonald & Lavie, 2011). Moreover, visual distractors, even when irrelevant to the task, capture attention (Theeuwes, 1994; Lavie, 2010; Forster & Lavie, 2011; Lavie & Dalton, 2014).

The above findings on the effect of auditory and visual noise on perception indicate that behaviour is affected by the presence of noise—both visual and auditory. Therefore, in our study, we tested the efficacy of different types of spatial sounds, specifically, stereo and binaural, in a virtual environment with and without environmental noise. We used a visual search task with different auditory cues to test their relative effects on search performance. This task allowed us to use both visual and auditory modalities in the VE and to place our task at the intersection of the two modalities. In this manner, we studied both visual target identification and sound localization simultaneously in a noisy virtual environment. A sound localization task or a visual search task on their own would be insufficient to understand perception of auditory and visual stimuli in an ecologically valid virtual environment. Our setup enabled the study of the interaction between both visual and auditory modalities in a VE. In this manner, we mimicked a common scenario in everyday life - looking for someone in a crowd, which is made easier if they call to us. This is where the advantage of binaural audio—that it can be placed at a distance and elevation along an azimuth—comes into play. The sound source would be congruent with the visual stimulus location enhancing stimulus detection. A stereo cue, in contrast, only has slight delays between the inputs to the left and the right ear, which gives the illusion of depth, but does not enable accurate sound localization. Therefore, the binaural audio cue is expected to facilitate visual search more than the stereo cue, as already found in previous literature (Hoeg, Gerry, Thomsen, Nilsson, & Serafin, 2017; Brungart, Kruger, Kwiatkowski, Heil, & Cohen, 2019). This will not necessarily be true in the condition with environmental noise, where both auditory and visual distractors will interfere with target search and localization.

The interim results from our study have already been described elsewhere (see Ruediger et al., 2019). The descriptive results indicated lower performance variability in the presence of an auditory cue with a slight indication that the binaural cue may be more advantageous than the stereo cue. Participants did not report differences in mental load between the experimental conditions on any dimension measured using the NASA-TLX questionnaire (Hart & Staveland, 1988). Participants generally reported high spatial presence in the task as measured using the Igroup Presence Questionnaire (IPQ; Schubert, Friedmann, & Regenbrecht, 2001).

In the present investigation, we derived four measures from the eye tracking data we collected during the experiment—two measures pertaining to the

spatiotemporal characteristics of eye movements made during the task and two measures pertaining to the physiological response to the task. We chose these measures to obtain a comprehensive understanding of behaviour in a rich virtual environment. These measures would tease apart the different cognitive processes that contribute to task performance, allowing us to assess the effect of the environment and the different auditory cues.

The first measure, time to first fixation (TFF), quantifies the time for target search, which is an indicator of the speed of target localization. The second measure, gaze trajectory length (GTL), quantifies the length of the search path, which gives us insight into the search process adopted by participants. The TFF results are expected to replicate the results obtained by Høeg et al. (2017). We expect that the binaural cues will result in shorter search times (TFF) and shorter search paths (GTL) than the stereo and no cue cases in the noise-free environment. Such a result would indicate quicker target detection with binaural cues in a realistic environment, which would strengthen the case for binaural sound use in VEs.

We do not have a specific prediction about whether the same results will be obtained in the condition with environmental noise. Even if the cues are effective, the presence of distracting noise could make the search task more effortful. In our interim analysis, although there was no discernible pattern in the mental effort report of participants, there was a report of frustration in the conditions with environmental noise (Ruediger et al., 2019). Therefore, in the present study, we focused on two measures of mental effort that could be derived from the eye tracking data-blink rate and pupil size. Pupil diameter is a well-established indicator of cognitive load that increases with increase in load (Beatty, 1982; Beatty & Lucero-Wagoner, 2000; Chen & Epps, 2014; Mathôt, 2018). It has been tested as an indicator of mental effort in practical applications such as combat (de Greef, Lafeber, van Oostendorp, & Lindenberg, 2009), driving (Palinko & Kun, 2012; Benedetto, Pedrotti, Minin, Baccino, Re, & Montanari, 2011) and surgery (Zheng, Jiang, & Atkins, 2015). In contrast, blink rate is a more ambiguous measure. In some studies, blink rate has been reported to decrease with cognitive load (Veltman & Gaillard, 1998), while in others, blink rate has been reported to increase with load (Chen & Epps, 2014). A more complicated relationship of blink rate with different types of loads has been found in other studies (Recarte, Perez, Conchillo, & Nunes, 2008; Merat, Jamson, Lai, & Carsten, 2012). In our study, we expect pupil size to increase in the conditions with environmental noise, while no specific

prediction is made for blink size. We also expect pupil size to be higher in conditions without the cue than with stereo or binaural cue. An advantage for binaural cue in terms of cognitive load measures would be the ultimate benchmark for the utility of binaural sounds in VEs.

4.2.2. Materials and Methods

The dataset used in this article was obtained from a VR experiment. The stimuli used, pre-tests for stimulus validation, study setup and description of the data obtained in the experiment are detailed in Ruediger et al. (2019).

4.2.2.1. Participants

A total of 20 participants (8 female) from the University of Kaiserslautern aged between 22 and 32 years (M = 27.32; SD = 2.97) volunteered to perform the experiment with informed consent. Most participants reported relatively little previous experience with virtual reality on a 5-point scale (M = 2.73; SD = 0.86) ranging between 'First time use' and 'Already living in VR'. Data from two participants, whose data were not recorded in one or more experimental conditions due to technical errors, were removed from the analysis. One more participant with extreme values was removed from the analysis as explained in Section 2.5.2. Data from the remaining 17 participants were analysed.

4.2.2.2. Stimuli

The VR stimuli were presented in an HTC Vive with an integrated Tobii eye-tracker. Each stimulus consisted of a scene acquired using simultaneous 360° video and audio recording of a real-world handball game stadium. Both the scene and the scene acquisition method were selected in order to achieve maximum fit with reality. A sport environment integrated rich visual and auditory information in the scene. Moreover, the simultaneous video and audio recording ensured that auditory noise was spatially synchronized with the visual stimulus. This synchrony ensured that the acoustic cues were separable from the environmental noise.

There were two stadium conditions: empty and full. In the empty stadium condition, the scene included a video of the empty stadium with sparse activity from groundskeepers, etc., resulting in low visual and auditory background noise. In the full stadium condition, the scene included a live audience in the stadium and players entering the handball court, which resulted in a condition with a noisier visual and auditory background. Besides the two stadium conditions, there were three auditory cue conditions: no cue, binaural cue, stereo cue. The auditory cue

was an air horn signal, which is a typical sound at handball games, as horn signals are used by fans to cheer for the team. In addition, this horn signal was evaluated in a pre-test to ensure that participants were able to distinguish the stereo and binaural cues (Ruediger et al., 2019). The sound for the cue was rendered using the Adobe Premiere Toolbox in both stereo and binaural using the generic standard head related transfer function. The combination of the two stadium and three cue conditions resulted in six scenes (videos).

The visual target presented in the scene was a set of three Minions (fictional, yellow creatures from the popular movie franchise) stacked vertically (Figure 1). The minion was chosen as a visual search target to mimic the problem of finding a person in a crowd in a VR setting. These targets blend in with the yellow home team shirts, but at the same time, are easily recognizable because of their unique and broadly known appearance. The colour scheme of the visual stimuli ensured they could not be identified in the peripheral visual field; however, they were distinct enough to be quickly identifiable once focused. These targets were presented in one of six locations on the azimuth plane (-135°, -90°, -45° and 45°, 90°, 135°) for each stadium and cue condition, resulting in 36 trials (2 stadia x 3 cues x 6 locations). A pre-test was conducted to determine the number of recognizable directions of a sound cue for the chosen signal. The pre-test revealed that only changes in the azimuth plane were perceived correctly. Therefore, no variations in the elevation were made in the experimental setting.

4.2.2.3. Data Acquisition

Data was acquired at 100 Hz sampling frequency using a Tobii Pro VR Integration eye-tracker retrofitted to an HTC Vive (2016 version). Data from each eye was recorded via one eye-tracking sensor and ten infrared illuminators with a total trackable field of view of 110° for eye movements.

4.2.2.4. Procedure

Participants were instructed to perform a search task in a VR environment. After the participant was seated in the lab, they wore the VR headset and their eyes were calibrated using a 9-point calibration. The instructions for performing the task were displayed on the screen before the participant was presented with the six experimental trials for one stadium-cue combination. Each trial began with a large, blue cross in the centre of a stadium scene, presented for approximately five seconds. The participant was asked to fixate on the blue cross and begin searching

for the target as soon as the blue cross disappeared. In the two conditions with auditory cues, cue presentation was synchronized with the disappearance of the blue cross. Participants had been instructed to look around the stadium scene and find the visual target, which was presented for 6840ms from trial onset. On finding the visual target, they were asked to fixate on the target until it disappeared. The end of the trial was indicated by two, large red arrows directing the participant toward the central blue cross. After these 6 trials, the participant was asked to answer the NASA-TLX questionnaire (Hart & Staveland, 1988). The NASA-TLX is a well-established tool to provide a reliable assessment of perceived mental workload (Hart & Staveland, 1988; Said et al., 2020). A 20-level version of the NASA-TLX was used for this study (1 = low to 20 = high).

Figure 1: Fixation behaviour from one participant in the empty binaural (top) and another participant in the full binaural (bottom) conditions overlaid on a snapshot of a trial. Fixations are indicated by purple and red circles. The area where the central fixation cross was presented is indicated by a black square. The six locations are indicated as Tx (T1 to T6), where x is the location's position in the sequence of six target presentations in a condition. Participant fixations are concentrated over the fixation cross and also in the six target locations. Background is partially scrambled to remove advertisements and faces are blurred to hide identities.

The above procedure was done for all six stadium x cue combinations. All six target locations were presented in random sequence for each of the six stadium-cue combinations, which were also presented in random order for each participant. After the end of all trials, the participant was also asked to fill in the IPQ (Schubert, Friedmann, & Regenbrecht, 2001). The IPQ consists of 14 items with a seven-point Likert scale (0 to 6). The 14 items load on the four factors Spatial Presence (SP), General Presence (GP), Involvement (INV) and Experienced Realism (REAL). We observed (Ruediger et al., 2019) that SP was rather high, while the single item factor GP, INV and REAL were expectedly lower. The entire experiment, including the 36 trials, the six NASA-TLX questionnaires and the IPQ, took approximately 30 min to complete.

4.2.2.5. Data Analysis

Data were analysed only from those trials where the visual targets were successfully found. Target hits were assessed using predetermined areas of interest, which were superimposed on the stimulus material via Tobii Pro Lab v1.152 software (Tobii AB, Danderyd, Sweden). If the visual target was fixated on for at least 70 ms, the fixation was considered a target hit and the trial was included in the analysis. Trials with extreme values, as explained below, were also removed from the analysis, such that the same set of trials were analysed for each measure. The fixation measures obtained using Tobii Pro Studio and all the other measures were computed using custom-written Python scripts. Only the results of eye-tracking data are analysed and described below. The results from the NASA-TLX and IPQ have been described in Ruediger et al. (2019).

4.2.2.5.1. Time to First Fixation (TFF)

The TFF was calculated as the time between the appearance of the visual target and the first fixation on the target using Tobii Pro Studio. The TFF was considered an indicator of search performance. However, during the experiment, experimenters observed participant differences in the speed at which the task was performed. Since participants were not given any instructions regarding speed for performing the task, participants may have adopted a slower or faster pace at which to perform the task. This can be observed in the high TFF variances (see boxplot Figure 2a), which might render it incomparable between participants. Therefore, we computed an additional measure of the search path as described in Section 2.5.2.

For the analysis, trials with TFF with extreme values (less than 50 ms) were removed since the search interval was not reliable. This resulted in 4% data loss (3 trials from 2 different participants lost).

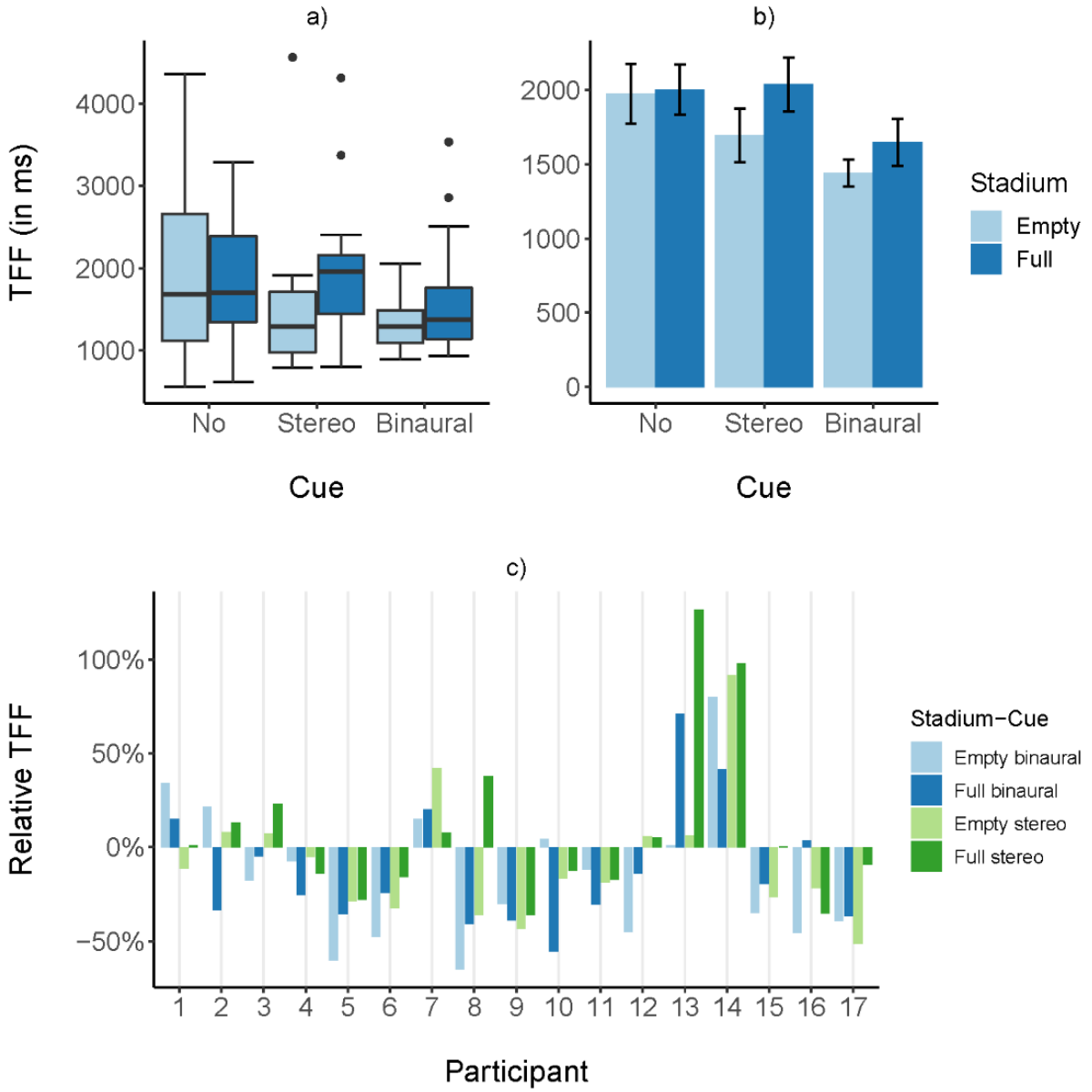

Figure 2: (a) Boxplot with median TFFs and (b) barplot with mean TFFs and error bars indicating standard errors of mean for 17 participants in the no cue, stereo and binaural cue conditions for the empty and full stadium conditions. (c) Relative changes in TFF for 17 participants in no cue, stereo and binaural cue conditions for empty and full stadium conditions. The no cue condition is interpreted as a baseline for the stereo and binaural conditions.

4.2.2.5.2. Gaze Trajectory Length (GTL)

Gaze trajectory length was calculated by adding up the pairwise differences between normalized gaze point coordinates G (x, y) of subsequent timestamps recorded until the first fixation on the target occurred (after n time steps).

$$GTL = \sum_{i=1}^{n} (G_{i-1} - G_i)$$

$$GTL = |GTL|$$

Since the gaze points were normalized with respect to the extent of the scene, GTL values are a factor of the scene width. For example, a GTL value of 2 implies that the gaze path was twice the scene width.

GTL is an indicator of the search path adopted by the participant. With easier search, the GTL would be shorter, whereas, in more prolonged search trials where the participant searches in more locations on the scene, GTL would be longer. The GTL suffers from the same individual differences as the TFF. As a consequence, reliable comparisons are only possible within each participant or by introducing a relative dimensionless measurement, e.g., the ratio of the measures between the different conditions relative to the no cue condition as depicted in Figures 2c and 3c.

Data from one participant with extreme gaze trajectory lengths was removed from the analysis. The extremely long search paths appeared to be due to a technical error.

4.2.2.5.3. Blink Rate

The blink rate during search was used as a measure of cognitive effort. To identify blinks, the pupil size data stream from the eye recording was used. Blinks are represented as missing values in the pupil data. However, the pupil value could also go missing because of other small eye movements, measurement artefacts, etc. Therefore, the pupil size data was first preprocessed by identifying small artefacts as missing values of 50 ms or less. These values were filled with the last valid pupil value. On this artefact-corrected pupil series, blinks were identified using the algorithm devised by Hershman et al. (2018). The algorithm identifies the correct start and end of the blink by identifying a decrease preceding and an increase succeeding a sequence of missing pupil values. Blinks that occur within 50 ms of each other are also merged into one larger blink. However, the pupil value could

also go missing because of other small eye movements, measurement artefacts, etc. Therefore, for the purpose of obtaining the blink rate, missing values were identified as blinks only when the blink duration was greater than 100 ms. After this step, the number of blinks was counted for each trial from the trial onset until the first fixation on target. Finally, blink rate was calculated as

$$Blink\ rate = \frac{Blink\ count}{TFF}$$

4.2.2.5.4. Pupil Size

Like blink rate, pupil size was also used as a measure of cognitive load. For this purpose, the pupil size data series marked with the blinks identified in the previous step was used. After identification of blinks, irrespective of blink duration, blink regions were interpolated using an order-3 spline 100 ms before and after the blink. Using this interpolated series, we calculated baseline-corrected average pupil size from trial onset until the first fixation on the target. Although there was a potential baseline interval of no activity when the blue cross was presented, it could not be used because participants did not always fixate the cross steadily. Therefore, the mean pupil size for each participant across all conditions was calculated as the baseline pupil size. This baseline was subtracted from the mean pupil size in each condition and target location giving us the demeaned Pupil Size.

4.2.2.5.5. Statistical Analysis

The data was analysed using repeated-measures ANOVA with stadium and cue as factors. For pupil size alone, an additional analysis was performed for the empty stadium trials with cue as a single factor. In case of violation of sphericity, the Greenhouse–Geisser corrected p-values are reported. For post hoc tests with multiple pairwise comparisons, Tukey-adjusted *p*-values are reported.

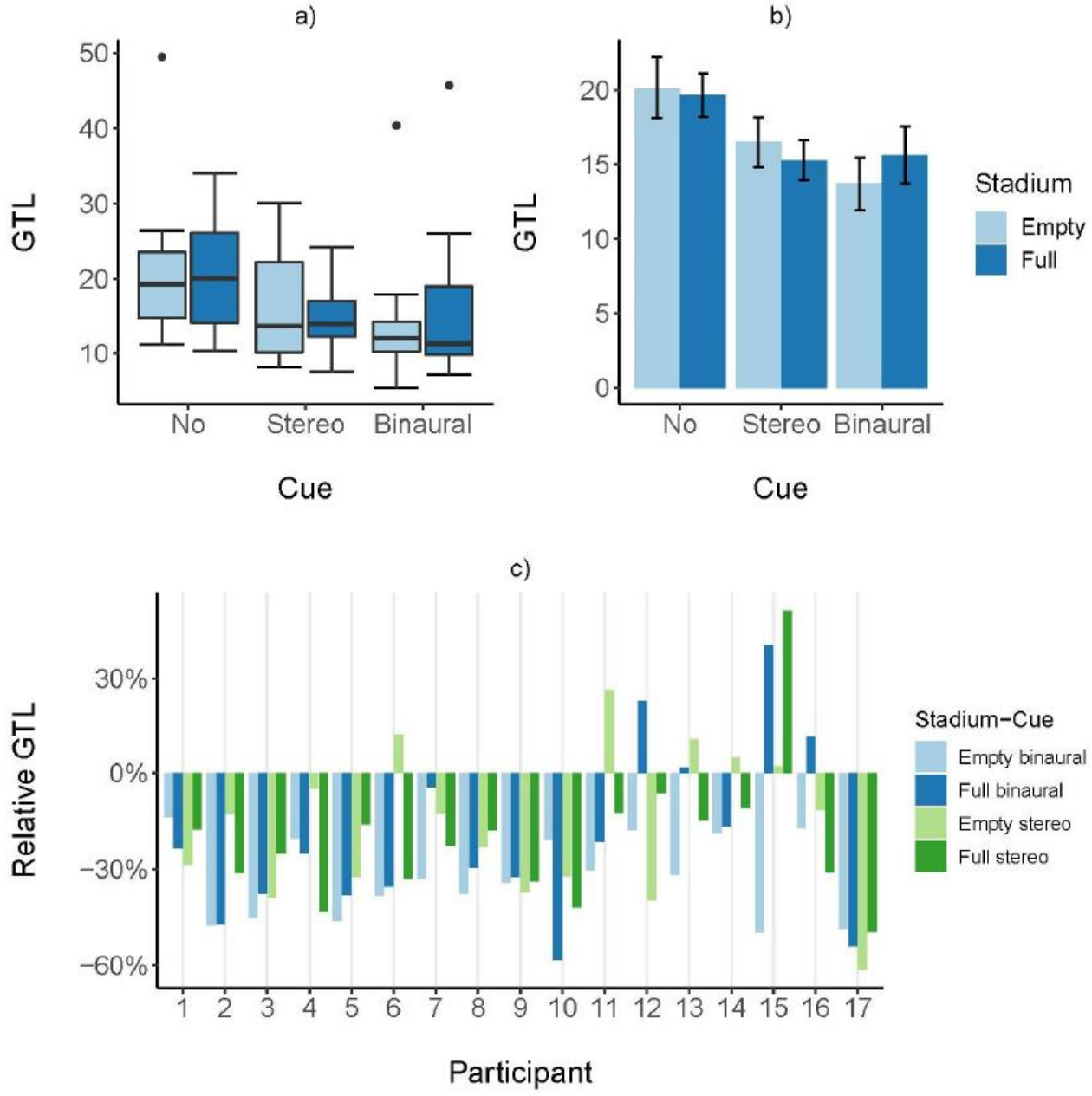

Figure 3: (a) Boxplot with median GTLs and (b) barplot with mean GTLs and error bars indicating standard errors of mean for 17 participants in the no cue, stereo and binaural cue conditions for the empty and full stadium conditions. GTL is measured in terms of the number of scene widths spanned. (c) Relative changes in GTL for 17 participants in no cue, stereo and binaural cue conditions for empty and full stadium conditions. The no cue condition is interpreted as a baseline for the stereo and binaural conditions.

4.2.3. Results

4.2.3.1. Hit Rate

We assessed search performance by performing a repeated-measures ANOVA on target hit rate with stadium and cue as factors. There was a significant effect of cue on target hit rate, $F(2, 32) = 3.3$, $p = 0.049$, $\eta^2_G = 0.032$. Post-hoc tests showed

that the binaural cue conditions had a higher hit rate than the no cue conditions (p = 0.04) averaged over the empty and full stadium conditions as shown in Table 1.

	No Cue	Stereo Cue	Binaural Cue
Empty stadium	92 (15.7)	97 (6.5)	96 (12.5)
Full stadium	91 (17.8)	94 (14.4)	98 (5.5)

Table 1: Mean target hit rate (and standard deviation) for the two stadium conditions (empty and full) and three cue conditions (binaural, stereo, no cue) for 17 participants.

4.2.3.2. Time to First Fixation (TFF)

As mentioned earlier, there was large variability in the TFF across participants, especially in the no cue condition (Figure 2a). An ANOVA performed on the mean TFF values revealed a significant effect of cue, $F(2, 32) = 7.5$, $p = 0.003$, $\eta^2_G = 0.07$. Post hoc tests showed lower TFF in the binaural cue condition (Figure 2b) than in the stereo ($p = 0.03$) and no cue conditions ($p = 0.002$) averaged over the stadium conditions. While the measure TFF is generally highly objective, it suffers from individual participant effects, such as training or experience in related tasks or orientation in virtual reality in general. As a consequence, it might be erroneous to only evaluate the mean values for 17 participants. Therefore, we additionally investigated the TFFs for each participant to understand individual differences. For each participant, we calculated the relative change ratios between each of the cued conditions with respect to the no cue condition for each stadium condition and participant as depicted in Figure 2c. We found that search times were lower in the empty than the full stadium conditions for 12 participants in the binaural cue conditions and for 11 participants in the stereo cue conditions. For 14 participants, the binaural cue showed lower TFFs than the stereo cue conditions, while one participant showed the opposite effect. For the remaining two participants, the difference did not show a clear pattern.

4.2.3.3. Gaze Trajectory Length (GTL)

There was a significant effect of cue on GTL, $F(2, 32) = 16.6$, $p < 0.001$, $\eta^2_G = 0.09$. Post hoc tests for values averaged over both stadium conditions showed a higher GTL (Figure 3) in the no cue condition than in the stereo and binaural cue conditions (both $p < 0.001$). Here, again, we computed GTL as a percentage change relative to the no cue condition. We observed that for most participants, GTL decreased compared to the no cue condition (Figure 3c), which confirmed our findings from the ANOVA.

4.2.3.4. Blink Rate

We observed large within-subject and between-subject variability in blink rates. Some trials had no blinks at all, while others had a very large number of blinks (Figure 4a). The binaural cue condition had the least number of blinks. Across cue conditions, three were fewer blinks in the empty stadium condition than the full stadium condition. A 2 x 3 repeated measures ANOVA on mean blink rates did not show a significant effect of stadium or cue (Figure 4b).

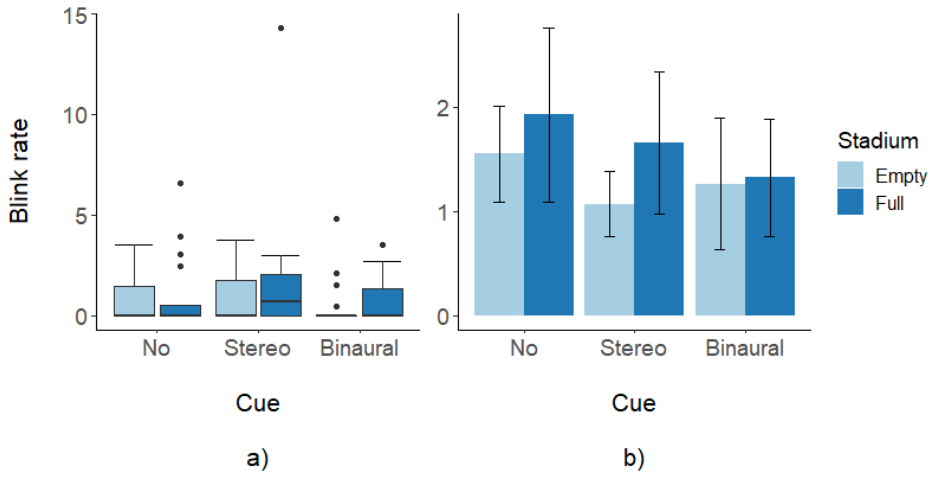

Figure 4: (a) Boxplot with median blink rates and (b) a barplot with mean blink rates and error bars indicating standard errors of mean for 17 participants in the no cue, stereo and binaural cue conditions for the empty and full stadium conditions. Blink rate is the number of blinks made per second.

4.2.3.5. Pupil Size

ANOVA on average demeaned pupil size revealed a significant effect of stadium, $F(1, 16) = 53.3$, $p < 0.001$, $\eta^2_G = 0.597$, with higher pupil size in the full stadium than in the empty stadium condition ($p < 0.001$). There was also a significant interaction between stadium and cue, $F(2, 32) = 5.5$, $p = 0.01$, $\eta^2_G = 0.067$).

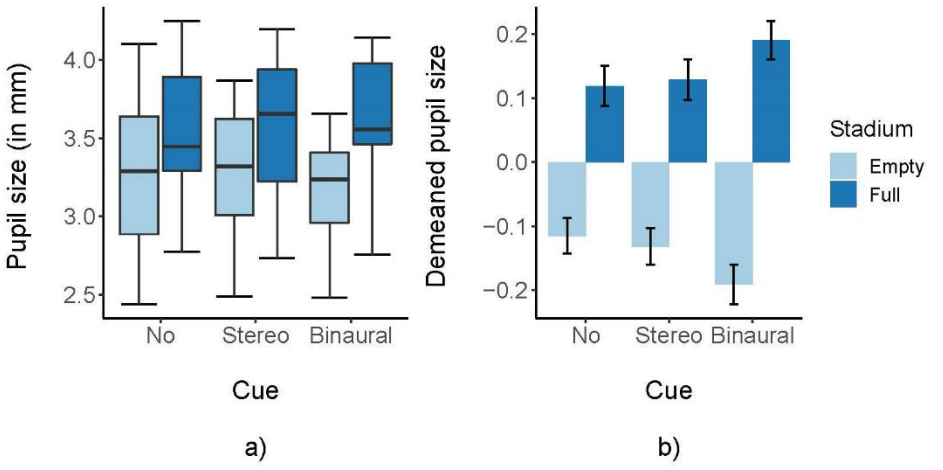

Figure 5: (a) Boxplot with median pupil sizes and (b) a barplot with demeaned pupil size and error bars indicating standard errors of mean for 17 participants in the no cue, stereo and binaural cue conditions for the empty and full stadium conditions.

Post hoc tests revealed a higher pupil size in the full stadium condition for all cue conditions ($p < 0.001$), although there was no effect of cue itself in any of the stadium conditions (Figure 5).

However, pupil size is sensitive to luminance variations. Since the empty and full stadium conditions are visually different, the full stadium condition has much larger luminance variations, which might affect pupil size differently, as seen in the pupil size deviations from mean. Therefore, we assessed pupil size again only in the empty stadium condition with only cue as a factor. This analysis did not show a significant effect of cue on pupil size.

4.2.4. Discussion

In this study, we compared the effect of three types of auditory cues (no cue, stereo cue and binaural cue) on visual search behaviour in two types of virtual environments using four measures—time to first fixation (TFF), gaze trajectory length (GTL), blink rate and pupil size.

4.2.4.1. Behavioural and Eye Position Measures

First, we found a performance advantage for binaural cue in comparison to trials where cue was absent, whereas, such an advantage was not present for stereo cues. This improved performance was also visible in the target search duration (TFF). Participants were quicker to find the target with the help of a binaural cue than with a stereo cue or in the absence of an auditory cue (Section 3.2). This result is in line with the quicker search times obtained in the studies by Hoeg and colleagues (2017) and Brungart and colleagues (2019).

The gaze trajectory length (GTL) measure, which quantified the length of the search path (Section 3.3), revealed a cue advantage as well. Trials with no auditory cue showed longer search paths than trials with binaural and stereo cues, clearly showing a benefit of the auditory cue. However, there was no difference between the search paths of the stereo and binaural cues.

Although not statistically significant, the boxplots and summary barplots (Figure 2a,b) show that, in the presence of a cue, search durations (TFF) were higher when the stadium had distractors (full stadium condition) than when it did not (empty stadium condition). This effect may be attributed to distracted search in the full stadium only in the presence of auditory cues, since such an effect is not present when there is no auditory cue. This is visible in the individual participant data (Figure 2c), where 12 of 17 participants show lower search times in empty

than full stadium conditions for the binaural cues (11 for stereo cues). While research combining task performance with fully moving VEs is scarce, related research could provide additional insight. Olk and colleagues (2018) reported slower detection of stimuli in a VE when those stimuli were harder to distinguish from surrounding objects either due to their distance or distinctiveness. This indicates that the minions, which were chosen to merge with the yellow elements of the VE, were indeed less distinctive when the players and audience with yellow uniforms appeared in the full stadium condition. Moreover, the appearance of a person - a strong social element - was recently shown to influence participants' visual attention in virtual reality (Hekele, Spilski, Bender, & Lachmann, 2021), where a person was fixated significantly more in a 360° video compared to a 2D video. In our task, the players and audience in the full stadium condition would have similarly attracted attention. This validates the minions as an appropriate visual target for our task. Additionally, the presence of people, even though task-irrelevant, also negatively impacted task performance in our study.

However, no such difference is seen between the two stadium conditions for the search paths. To understand this disparity, we additionally investigated TFF and GTL by separating the trials based on the location of the targets (Figure 6). For TFF, in the empty stadium, we found a high association between the target distance from the centre (in degree) and the time to fixation of the target. For the full stadium, this association holds true only in the presence of a binaural cue and to a lesser degree in the presence of a stereo cue. This implies that the full stadium interferes with the search process as expected.

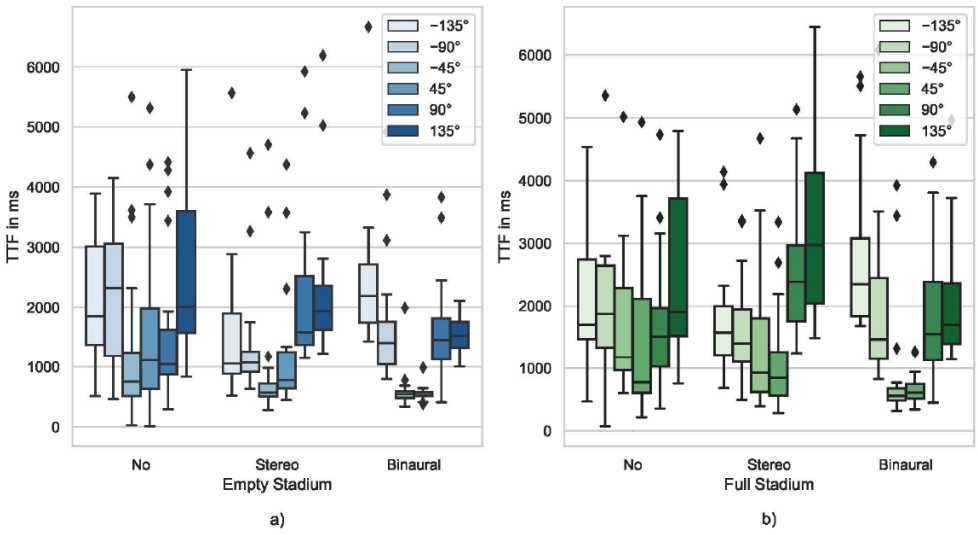

Figure 6: Boxplots of TFF for no cue, stereo and binaural cue conditions for empty (a) and full stadium (b) conditions grouped by target location (search trials). The target locations are marked by their relative angle to the starting position.

In contrast, for the gaze length trajectories, we could not find such a pattern. It remains unclear how the eye and head movements necessary for different target positions in the different stadium conditions moderate the overall results. This could have been because of a design shortcoming. As mentioned earlier, our participants did not always fixate exactly on the large, central blue cross before the start of each trial. This distracted trial beginning could mean that the search would not have always started from the middle of the display, leading to inconsistent GTLs. One solution to this problem would be to force the participant to fixate on the central cross to begin the trial. Alternatively, a more dynamic setup would have averted this problem by presenting the cue depending on the participant's current gaze location.

Another source of inconsistency in the results was individual differences. Comparing the per participant effects in Figures 3 and 4, the individual measure varies strongly for both gaze paths (GTL) and search duration (TFF). We could not spot any general pattern describing the relative degree of changes in either of the measurements. Larger sample sizes are required in future studies to mitigate the effect of this variability. Any variability stemming from differing levels of familiarity with virtual reality technology, although low as mentioned in Section 2.1, can also be explored with a larger sample size. In spite of the individual effects heavily moderating the degree between the absence and presence of an auditory cue in this visual search task, we found that the presence of any auditory cue speeds up the search performance.

Overall, the search duration (TFF) and search path (GTL) measures presented to be useful metrics of search behaviour in our task. Together, they have revealed a search advantage of auditory cues, with the binaural cue being slightly more advantageous than the stereo cue.

4.2.4.2. Physiological Measures

The next two measures we tested—blink rate and pupil size—were physiological measures, both of which have been studied in virtual environments. Although not statistically significant, we found lower blink rates in the empty stadium conditions than in the full stadium conditions. However, the trials without cues did not show such an effect. On the contrary, we observed lower blink rates in the easier trials with the binaural cue. Blink rate is an inconclusive measure of cognitive effort (Marquart, Cabrall, & de Winter, 2015). Blink rate is known to decrease in cases of extreme focus and increased workload, as observed in surgeons

(Veltman & Gaillard, 1998; Zheng, Jiang, Tien, Meneghetti, Panton, & Atkins, 2012). Veltman and Gaillard (1998) indicate a distinction in the underlying factors that affect blink rate. They found that blink rate decreased when more visual information had to be processed, while it increased when the difficulty of the task increased. In an experiment systematically varying visual and cognitive demands, Recarte and colleagues (2008) found that blink rate decreased with visual load and increased with mental load. In a driving task, Merat et al. (2012) found a similar fall in the blink rate with increased visual information in the absence of a secondary task. Adding a secondary task increased blink rates, although some results did not fit this pattern, indicating a tradeoff in blink behaviour between visual information intake and mental workload. These U-shaped patterns in blink rates have been interpreted differently by others. Berguer and colleagues (2001) found that surgeons had lower blink rates when performing surgery than at rest, but blink rate increased while doing the same in a laparoscopic environment. They interpreted this result as the outcome of a conflict between task demand or stress and concentration. Zheng et al. (2012) found, in a VR laparoscopic surgery setting, that those participants who reported more frustration and mental effort in the NASA-TLX blinked less frequently. It is also worth noting that some studies have not reported an effect of mental load on blink rate, while pupil size or other measures responded to load (Benedetto, Pedrotti, Minin, Baccino, Re, & Montanari, 2011; Ahlstrom & Friedman-Berg, 2006).

In the context of the ambiguous nature of factors affecting blink rate that were discussed above, our results did not show a discernible pattern to draw parallels to any of the above literature. Although the full stadium condition had higher visual information in the display, this information was task-irrelevant, and therefore, it cannot be equated to the demand of having to process additional visual information as described above. Our task may have been too easy to elicit an effect of visual or mental load in comparison to the difficult driving and surgery scenarios that have been studied. In addition, the median blink rates and an investigation of individual participant blink rates revealed high variability in the data. Large variability in blink rate was also reported by Benedetto et al. (2011). In spite of blink rate being decreased in head-mounted VR displays in comparison to monitors or natural settings (Kim, Sunil Kumar, Yoo, & Kwon, 2018), in our data, some participants showed extremely large blink rates (up to 15), which may indicate poor data quality. Blinks are identified when the pupil is not detected by the eye tracker. The Tobii Glasses eye tracker we used was embedded in the VR headset, which

should have resulted in lesser data loss. However, loss of the pupil size data stream occurred more frequently for some participants (as high as 19% for one participant). Some participants wore glasses and/or lenses, which may have resulted in higher data loss. This is a shortcoming of video-based eye-tracking, which needs to be overcome to increase the reach of eye tracking integrated VR setups. A simple solution to this problem would be to record a video of the participant's eyes, which would allow us to manually identify blinks.

The comparability of our results with existing literature is additionally made difficult by the fact that the blink sensors (remote eye tracker, head-mounted eye tracker, EOG) and blink detection algorithms (manual, automatic, different duration thresholds, etc.) are all different. If future studies report data quality and the precise parameters used for blink detection, it will become easier to reach a consensus on this complex measure.

Our last measure, pupil size, showed only an effect of the stadium with larger pupil sizes in the full stadium condition. This effect is clearly due to luminance differences between the two scenes. Pupil size responds to both changes in luminance and cognitive effort (Mathot, 2018) and our result shows that the task-evoked pupillary response (TEPRs) was not separable from luminance effects. However, even in the empty stadium trials, where we can reasonably assume equivalent luminance between cue conditions, any increase in cognitive effort that may have been present was not seen in our results except as a small decrease in median pupil size. It should be noted that TEPRs are small changes that require a large number of trials to be averaged and reflect large changes in cognitive load (Beatty, 1982; Beatty & Lucero-Wagoner, 2000), both of which did not apply to our study.

Although our scene was chosen to be visually and auditory realistic, which was an advantage for immersion and presence in VE, the realistic stimulus was also partly the reason why the physiological measures did not perform well. We could conduct the same study with more fine-grained control over the visual and auditory noise levels in the environment. The experiment could have only auditory noise or only visual noise as additional conditions to isolate the effect of noise from the amount of visual input that needs to be processed. This would enable the use of both pupil size and blink rate.

For future use of pupil size in our paradigm, besides correcting the design constraints mentioned above, technical sources of error need to be accounted for as well. Eye movements themselves cause distortions in pupil size, which are most

evident in different camera viewing angles. Correcting these distortions requires complex mathematical models (Mathur, Gehrmann, & Atchison, 2013). Most eye trackers incorporate these perspective-distortion corrections; however, individual differences might still exist (described and modelled in Mathur et al., 2013). One solution is to use measures that are resistant to luminance changes and pupil distortions, such as the Index of Pupillary Activity (Duchowski et al., 2018), which measures changes in the oscillatory behaviour of pupil data. However, this measure requires long trial durations, which was not the case in most of our trials.

4.2.4.3. Eye Tracking and Immersive VR

Eye-tracking gives us access to many dimensions of behaviour. It extends the study of simple behavioural responses by giving us a more fine-grained insight into human interaction with the environment. In our task, we could have asked participants to simply press a button on detecting a target. However, recording eye movement data instead, allowed us to look more closely into the search strategy of each participant. It also allowed the participant to perform the task more naturally without having to remember buttons that they might usually have to press. Such eye movement paradigms make stimulus-response paradigms more seamless and ecologically valid. However, as discussed above, some of the measures obtained from eye tracking data have shortcomings that need to be overcome. Higher fidelity signals will be required in the future for effective use of systems that can provide metrics of cognitive effort for improving user experience and for providing user feedback.

In spite of the lack of results from the physiological measures, our eye position measures have revealed a definitive advantage of an auditory cue for target localization and detection in a virtual environment. We also found that visual and auditory noise interfered with target localization in the presence of facilitating cues. There is also an indication of the usefulness of the binaural cue, which was seen in spite of the large individual differences between the different degrees of environmental noise. This evidence is in support of the use of spatial sound in different virtual environments to improve responsiveness and immersion in the environment. Our study can be extended in future research with different environments and different degrees of noise to obtain a more comprehensive understanding of sound localization and perception in realistic VEs. This would enable the design of more effective virtual environments with appropriate use of binaural sounds.

4.3. Interim Summary

Study 3 compared the learning outcome after either watching an educational 360° video presented on an HMD or watching the same content presented on a classic 2D video on a tablet. Eye-Tracking data were collected in both conditions which gave novel insights about user attention in video-based learning.

Study 4 featured an HMD-based visual search design with additional auditory stimuli to investigate differences in task performance depending on noise and different kinds of audio cues. The study provided new insights into the value binaural audio can bring to virtual reality as well as introducing visual and auditive noise to visual search paradigms.

Further discussion of these results as well as embedding them into the larger picture of the present work will take place in chapter 5.

4.4. Further Related VR Research

Study 3 and 4 were two VR-based experiments which were described in detail in this chapter. Apart from those two experiments, several further studies were conducted in the last few years. One study has been submitted in a first version in early 2020. The paper was concerned with design recommendations of virtual environments for users with disabilities and discussed ways to implement a barrier-free VR framework for car mechanics (Hekele, Bender, Spilski, Homrighausen, Werth, & Lachmann, 2020). In this way, the study directly followed the original string of experiments which was planned to follow study 3. This string of experiments will be elaborated on in more detail in chapter 5.2. The data collection for this study was conducted as pre-test for later studies as part of several public events with vocational educators around Germany. Participants could experience a virtual car workshop which was carefully constructed to allow for realistic interaction even with restricted mobility of the user. Participants were instructed to explore the environment at their own speed but could request assistance in either verbal or written form at any time. After the VR part, they were asked to fill out the IPQ presence questionnaire (Schubert, Friedmann, & Regenbrecht, 2001) and were asked for general feedback to improve the environment for future use cases with people with disabilities. During the review process the reviewers recommended to conduct additional data collection and the addition of eye-tracking and movement data to gather additional user data which

would be helpful for the process. However, no follow-up studies were possible due to the Covid19 pandemic, and the project ended before research could be resumed. Therefore the manuscript remains unpublished. Had research continued as planned, study 3 would have been the entry point to investigate different aspects to improve vocational education using virtual environments especially for people with disabilities. Unfortunately, this was not feasible as testing remained nigh impossible until 2021 (see chapter 5.2. for additional information).

Outside of vocational education, a large interdisciplinary research project was started in 2022 which set out to investigate dual load paradigms embedded within a driving task in virtual reality. Wallet and colleagues (2009) identified driving in virtual reality as a use case for transfer learning. To this end, a group of researchers and students designed a virtual driving environment with obstacles which participants were instructed to drive along using a steering wheel. In the experiment, participants wore an HMD with an integrated eye-tracker (HTC Vive Pro Eye; High Tech Computer Corporation, New Taipei, Taiwan) and had to solve either cognitive or linguistic working memory tasks either while driving (high-load condition) or after stopping the car (low-load condition). Behavioural data looked promising as participants performed worse in the high-load condition, making more mistakes in the working memory tasks as well as colliding with more obstacles in VR. Due to problems with the Unity integration of the eye-tracker, a significant portion of eye-tracking data was not properly recorded or incomplete. Due to the difficulty of integrating the different data streams and problems in recruiting a comparable sample to the original study, the study ultimately had to be scrapped and remains unfinished. A follow-up study using a mobile eye-tracker within a large human-in-the-loop driving simulator was planned for late 2023 but remained in a concept stage. The driving studies, while covering a different concept, are relevant for the larger picture because they build upon earlier insights from our lab, in particular study 4 through the combination of multiple channels for stimuli presentation. Similar to the aforementioned barrier-free VR framework the driving research has and would have continued to rely on a matching multisensory experience to increase immersion (Ruotolo et al., 2013) but also enable stimuli and instruction presentation in either a visual or auditive way, depending on the individual's needs (Agrewal, Simon, Bech, Baerentsen, & Forchammer, 2020) and to minimize problems with the detection of objects in the peripheral field (Olk, Dinu, Zielinski, & Kopper, 2018).

Chapter 5: General Discussion

5.1. Key Results from Studies 1-4

The presented research aimed to provide insights into several aspects of task difficulty using a mix of lab-based (study 1, 3) and VR-based research (study 3, 4) as well as field research (study 2). The underlying goal of this body of research was to identify common factors in vocational professions and investigate how humans interact with emerging technologies and digitalised materials.

Study 1 had the goal to compare three assessment modes of the established trail making test (Oswald & Roth, 1987) in a within-subjects design to determine whether the type of presentation affected task performance. The study found significantly faster task completion speed in the trail-making test using tablet and pen compared to the same test on a tablet using one's finger and compared to the classic pen-and-paper version of the TMT. Participants made a comparable number of errors across conditions and linear regressions revealed that in the subsample of older participants task completion times was improved the fewer errors they made or if they used the tablet and pencil combination. In addition to the performance metrics, over two thirds of participants reported a preference for the tablet and pencil version compared to both other presentation modalities. Participants also reported they performed better in the tablet and pencil version, which is in line with the faster completion time. On the flip side, self-report data showed a trend that participants felt more comfortable in the paper and pencil version as well as reporting a trend towards higher feelings of control.

Study 2 used structured interviews and surveys to gather insights into the physical and cognitive demands on construction workers in their daily work environments. The main hypothesis was explorative due to the lack of previous research investigating different domains of job-specific demands in construction workers. Based on pre-tests and related research it was assumed that physical demands would be higher than cognitive demands in construction workers. In the collected data medium to high demands across a broad range of physical and cognitive domains was revealed in both interviews based on the German version of the F-JAS (Kleinmann, Manzey, Schumacher, & Fleishman, 2010) as well as the NASA TLX (Hart & Staveland, 1988) survey. We concluded that previous research as well as the public opinion might have underestimated the cognitive demands in a job which is traditionally seen as heavy physical labour. Practical implications were formulated and focussed on implementing processes to create more awareness on

cognitive demands and avoid cognitive overload which could have detrimental effects on task performance.

In study 3 investigated learning outcomes from educational videos using a single-factor between-subjects design. The key research question was whether a 360° video would prove beneficial for the learning outcome of vocational educational content compared to a 2D video presented on a tablet. The main hypothesis was undirected and investigated differences in the learner's visual attention between the two experimental groups. To this end, the total fixation duration was used as a correlate of attentional focus (see e.g. Jeelani et al., 2018) on the two focal points of the instructional video – the instructor and the engine block. On the instructor, significant differences were found with participants spending on average four times as much time fixating on the instructor in the 360° condition compared to the 2D video (98.5 and 24.9 seconds, respectively). No differences in total fixation duration could be found for either of the other two areas of interest, therefore hypothesis 1 was only partially confirmed. The second hypothesis predicted improved learning outcome in the 360° video condition compared to the 2D video group. No differences between the experimental groups were found, both performing far above average, with participants scoring a median of 59 out of 64 points in both conditions. Additional analyses investigated presence and immersion in both experimental condition as a manipulation check whether the HMD-presented video led to increased immersion. This manipulation check could be confirmed, participants reported higher levels of perceived immersion and presence in the 360° condition as measured by the IPQ questionnaire (Schubert, Friedmann, & Regenbrecht, 2001).

Study 4 also provided several interesting insights into a VR-based visual search paradigm with different audio cues. Using a handball stadium in either an empty state or with fans on the benches and players on the field before a game, visual and auditive noise was introduced to visual search while four eye-tracking measures were collected. The main research question expected shorter times to first fixation (TFF) and shorter gaze trajectory lengths (GTL) in the empty stadium with binaural cues compared to stereo or no cue conditions, in line with previous research. These predictions were partially confirmed, with binaural cues leading to shorter TFF compared to stereo and no cue conditions, and both binaural and stereo cues leading to shorter GTLs compared to the no cue condition. No specific predictions in regards to TFF and GTL were held towards the full stadium condition, we did expect environmental noise to lead to increased mental effort, which was measured with blink rate and pupil diameter measures. Analyses of

blink rate revealed no significant differences between any combination of cue and stadium condition, while pupil size was significantly higher in the full stadium condition with post-hoc tests indicating that pupil size was higher for all cue conditions in the full stadium. This effect was likely due to differences in luminance as discussed in study 4. Both pupil data and blink rate had high variance between subjects which might be due to technical difficulties in blink detection or participants who are not used to virtual reality.

5.2. Critical Review and Expansion of Experimental Paradigms

While the previous chapters contained their respective discussions of shortcomings and limitations of each individual study, the intention in this chapter is to take a more holistic perspective on the line of research. Insights from more recent research will be discussed and conclusions between the papers drawn.

Both study 1 and study 2 were conducted as pre-tests and early investigations, as were the additional studies described in chapter 3.4. Both studies have been published and contained limitations, which will be expanded upon here to level the path for further research. Study 1 suffered from some conceptual drawbacks, first and foremost it highlighted the difficulties in comparing input methods on tablets and pen and paper in semi-controlled lab conditions, since especially the tablet and finger combination had a negative impact on overall task performance. Participants also reported that using the tablet with the finger felt less natural and could hinder visibility of the task compared to a pencil as discussed in study 1 – these assumptions however were post-hoc and not investigated in the study itself. Some of these drawbacks could be reduced or at least investigated more closely by using a head-mounted eye-tracker. to allow participants unrestrained movements while also capturing their visual field irrespective of the participant's body position (Franchak, Kretch, Soska, & Adolph, 2011). Study 1 contained no eye-tracking, therefore conclusions regarding specific movement behaviour such as gestures with the pen or the finger relied on indirect estimations such as the self-reported preference and behavioural correlates in the form of task-related performance ratings. Fears and colleagues (2019) described a method to use a mobile eye-tracker to investigate children's handwriting. Since the eye-hand coordination is as important on a tablet as it is in handwriting (Lee, Junghans, Ryan, Khuu, & Suttle, 2014; Lin, 2019), future research investigating digitalized versions of assessments such as the TMT should consider using an eye-tracker. As

study 1 was mostly investigating behavioural measures, this was not considered at the time the study was conducted. Another possibility would be the addition of a form of tracking the pressure of the participant's pen or finger on the tablet (Lee, Junghans, Ryan, Khuu, & Suttle, 2014; Matic & Gomez-Marin, 2019) to investigate differences in handling the same applications using different input methods.

Study 2 provided insights into the multiple demands on workers in the construction industry, which also led to some shortcomings on the acquired data. Research in the field often poses a unique set of challenges, especially if it utilizes new technology (Brown, Reeves, & Sherwood, 2011) and field research has not always been able to replicate previous theoretical results (Campion, Morgeson, & Mayfield, 1999) and often suffers to a certain degree with patterns of missing data (Graham, 2009) – a trend which was also apparent in study 2. The methods utilized were previously validated and provided a good fit for the target group, however the time to complete these questionnaires or sit down for a relatively long structured interview was still an unusual situation for participants and might have influenced the results. Other research from our lab, including the field study with car mechanics in chapter 3.4., usually involved some questions how participants felt about the study design and the deployed methods. Whenever an eye-tracker was used, participants voiced concerns that their performance would be recorded and evaluated by their employer. While we eased the participants concerns to the best of our ability, their suspicion is not unfounded since eye-tracking can be used to detect fatigue or work performance in real-time (Buettner, Maier, Sauer, & Eckhardt, 2018; Li et al., 2020). Future work, be it from a research or educational perspective, should take steps to ensure that participants feel save in order to avoid a negative impact from participants not behaving naturally.

For studies 3 and 4, some shortcomings and backstory should be elaborated on as well, incorporating insights from research since their original publication which allows for a more in-depth literature review. In study 4 two virtual environments based on a handball stadium were created in which participants had to solve a visual search task. The study provided insights into the effect of noise on task performance. However, the study is not without limitations. First, the 360° video material used as foundation for the VE was, in hindsight, not perfectly suited to the task. While the stadium environment did serve to capture the participants' attention, it came with its own set of issues. This was especially apparent in the full stadium condition where participants were very close to the players, including during the visual search task. Participant performance was worse in the noisy full

stadium condition, but this effect could have been due to the presence of other people. Although these others were just part of the prerecorded video material, their presence might have had an impact due to the perceived proximity as shown in robotics (Hostettler, Mayer, & Hildebrand, 2023) and other virtual reality research such as recent work by Tao and Lopes (2022) where the authors present a system to redirect distractions from inside and outside virtual reality into the virtual environment to improve sense of presence. In the discussion of study 4 visual search in a fully immersive VR environment was listed as a future avenue for research, and in these future studies the presence of other persons who are not relevant to the task should be carefully considered as other entities has been shown to influence user behaviour in positive and negative ways alike in virtual reality (Oh, Herrera, & Bailenson, 2019).

Study 3 found no difference in learning outcome based on the presentation modality, which, while contrary to the expected outcome, was in line with the literature (see Moro et al., 2021; Ulrich, Helms, Frandsen, & Rafn, 2021). The limitations section discussed a potential ceiling effect from the sample of university students who were trained on vocational learning content via videos, however this result wasn't followed up by testing with the target group of vocational students. Another paper by Legault and colleagues (2019) reported that the use of virtual environments improved learning outcomes for learners compared to L2 word-pair learning, but only if those learners were classified as "less successful" learners with a mean accuracy below 80%. In the higher performing group of learners, there was no difference between the experimental groups. Given the high performance of those learners, the authors assume that the method of learning might be less relevant to high-performance learners. Irrespective of the individual performance a majority of the participants in the study reported a preference for the iVR condition as well as notions that the ability to move around improved their learning experience (Legault, Zhao, Chi, Chen, Klippel, & Li, 2019). In study 3, we observed a potential ceiling effect in both experimental groups: The vocational education learning materials might have been too easy for university students to learn, making the presentation modality potentially less relevant than previously assumed. While the students were pre-tested to avoid pre-existing knowledge contained in the material and the material itself was also pre-tested and deemed difficult enough to prevent guessing and other luck-based strategies, the students were performing far above expectations in both conditions and scored on average 59 out of 64 points in a standardized knowledge test, with no difference between nVR and 2D video. Future

research should incorporate additional procedures to ensure that the material is of suitable difficulty for the learners. This goes for both experimental studies where different learning media are compared but also for material designed for educational use.

Researchers in the training domain proposed the usage of adaptive difficulty to VR-based training programs (Aymerich-Franch & Bailenson, 2014), and similar technology has been successfully used in computer-based working memory training (Flegal, Ragland, & Ranganath, 2019) and application-based learning for language training (Shohieb, Doenyas, & Elhady, 2022). In order get to adaptive educational content in VR, the overarching shortage of VR content has to be solved (Jensen & Konradsen, 2018). The development and maintenance of VR material is expensive and while progress is being made, there are no standardized virtual learning platforms where content can be embedded easily (see e.g. Kavanagh, Luxton, Reilly, Wuensche, & Plimmer, 2017). On the other hand, the question emerges whether simple embedding of content through e.g. text or video would necessitate the use of virtual reality technology. While studies, including study 3, have shown that virtual reality not only is suitable for learning, but also increases learner motivation and engagement substantially (Parong & Mayer, 2018), the biggest advantage of virtual reality is the possibility to immerse participants fully in a virtual environment (Slater, 2018). Recent research in form of a system review by Hepperle and Wölfel (2023) gauged similarities and differences between virtual reality, screen-based and real-world experiments over the last few decades. The authors presented data which showcases that the majority of VR-based research is comparable to real-world, but not screen-based research. Hepperle and Wölfel (2023) conclude that one of the main advantages for VR is cost efficiency, since participants can easily participate remotely in experiments but also social activities if they have access to the equipment. Social activities in virtual reality are a big area of social participation, as many have experiences in the years of 2020 and 2021. Scientific evidence is provided in a review article by Bravou and colleagues (2022) for social inclusion of children with autism spectrum disorders as well as Sarupuri and colleagues (2023) in a large survey-based study investigating dancing in virtual reality during the Covid19 pandemic. The authors found that participants who reported to not partake much in social activities in the “real” world found social VR more easily accessible (Sarupuri, Kulpa, Aristidou, & Multon, 2023) This provides further support for the previously stated notion that VR can be used as a vehicle for

inclusive participation for students who might otherwise be left out due to their physical or cognitive abilities.

After study 3 was completed in late 2019, we were working on a set of follow-up studies, with one study intended to investigate how learners with cognitive or physical disabilities would interact with an educational virtual environment. After this planned study, two further studies were supposed to investigate whether vocational students with and without cognitive or physical disabilities could transfer vocational training from a virtual workshop environment to a "real" craftsman workshop. Due to the Covid19 outbreak in late 2019 and the following lockdowns in Germany throughout 2020 and 2021, none but one of these follow-up studies came to fruition. Under intense preparations, high security procedures and hygiene protocols we managed to conduct a field visit to a vocational school for people with disabilities in August 2020 to give participants with different physical and mental disabilities access to a virtual environment to gather feedback and investigate their navigation patterns to improve later versions of the virtual environment. However, at the time the deployed virtual environment was not suited to accommodate people with disabilities and further testing had to be suspended. It is to my great regret that this promising and highly important strand of research was left incomplete, but it is my sincere hope that future researchers see the same potential for inclusive learning in virtual reality as I do and use this technology to enable those learners to engage with education who are often overlooked in educational projects. First evidence for promising VR research with participants suffering from learning disabilities comes from Drigas and colleagues (2022) who created an intervention to train metacognitive skills such as mindfulness and breathing techniques. The authors reported that in particular the combination of VR-based intervention and therapeutic interventions has great potential for future scenarios. Future educational VR research, especially with disadvantaged populations, should not only try to further explore which content can and indeed should be ported into virtual reality, but also how the unique advantages of VR can be utilized in a learning environment. The next section will try to provide an overview over one major advantage of iVR environments by discussing the use of avatars and their potential for inclusive representation.

5.3. Avatar Representation in Educational Virtual Reality

The research presented in the present work has not relied on avatar-based virtual reality, future research could expand the presented paradigms to incorporate and potentially compare virtual avatars of learners to better understand the impact of virtual environments with and without avatars on learning outcomes. Study 3 has shown that presentation of video-based content on an HMD can prove beneficial in increasing learners' engagement with educational content (see Hekele et al., 2022), which is in line with previous research (Slater & Sanchez-Vives, 2016; Meyer, Omdahl, & Makransky, 2019). However, the explicit impact of avatars on learning has only recently become technologically viable to implement in scientific settings. First insights come from the research of Chang and colleagues (2019) who investigated the impact of avatar gender on learning outcomes in women and found that learning outcomes decreased when the virtual instructor used stereotype threats in the form of sexist remarks towards the female students. Interestingly, in this study no difference in the learning outcome or perception of the stereotype remarks was reported as a result of the avatar gender despite an all-female sample (Chang, Luo, Walton, Aguilar, & Bailenson, 2019). The last result seems to conflict with the Proteus effect proposed by Yee and Bailenson (2007) which would predict that the visual characteristics of an avatar affect the user's behaviour. However, this could be due to the study design which was more focussed on stereotyping and instructor behaviour than the participants' interactions with and through the avatar. In Study 3, the added depth by the 360° video was sufficient to increase the engagement with the instructor, as indicated by higher total fixation duration and number of fixations compared to the same content presented on a screen instead of an HMD.

Further insights on the usefulness of avatars for education come from Li and colleagues (2022) who compared vicarious learning outcomes from virtual sport exercises with and without an avatar as well as with and without a virtual coach and found that the inclusion of a user avatar would lead not only to identification with said avatar but also increased the participants' perceived competence as well as their motivation to maintain the training exercises in the future. Adding a virtual coach led to feelings of increased social presence which in turn further increased the intention to exercise in the future (Li, Ratan, & Lwin, 2022).

Both studies underline the impact of an instructor in virtual environments on the learning outcome of participants. While early educational virtual reality

research often found conflicting results in terms of the effectiveness of virtual reality for education (see study 3 for a detailed overview), the advances in fidelity of these virtual environments but also the avatars in those environments could have a major impact on learning outcomes in the future. Garau and colleagues (2003) noted that a realistic looking avatar with responsive gaze-tracking is already sufficient to increase the perceived quality of social interactions with such a social agent. Similar evidence comes from the field of social robotics, where participants report feelings of familiarity and show more cooperative behaviour towards a robot which is responding to their actions in a human-like way (see Hortenius, Hekele, & Cross; 2018 for an overview).

Outside of educational research a substantial body of research has concerned itself with cognition which occurs as a result of the interaction between a person and their respective environment. This construct named enactive cognition proposes that humans dynamically re-assess their courses of action depending on the capabilities of their bodies and the tools they have available (Steed, Pan, Zisch, & Steptoe; 2016). Steed and colleagues (2016) concluded that it should theoretically be possible to create avatars which could be used without requiring additional resources such as mental effort. This could, in conjunction with adequately designed environments, reduce extraneous load in learning environments (see Kirschner et al., 2018). Study 4 provided insights into the impact of visual and auditive noise on task performance in a visual search paradigm. While the study did not feature avatars and was not a fully immersive virtual environment, even pre-recorded footage from a handball stadium with players were sufficient to negatively impact task performance in virtual reality.

The importance of one's visual representation in virtual reality has been explored and discussed since it became technologically feasible in the 1990s (see e.g. Turkle, 1995; Yee & Bailenson, 2007). Research has shown the impact of avatar features on perception of external features such as object size (Banakou, Groten, & Slater, 2013), and in human-robot interaction the degree of body language adds a large amount of perceived humanness and positive emotional reaction in interactions between humans and robots (Hortensius, Hekele, & Cross, 2018). Recent advances in the ability to automatically generate personalized avatars from 3D body scans have shown to greatly increase feelings of body ownership and emotional response to one's avatar (Waltemate, Gall, Roth, Botsch, & Latoschik, 2018). Even on a manual level, the mere ability to change the looks of the avatar improves immersion as well as changing participant's behaviour to be in line with

their perception of “themselves” in form of their avatars (Yee & Bailenson, 2007; Freeman & Maloney, 2021).

Continuing this train of thought, it seems reasonable to assume that humans embodied in an avatar within a virtual environment would also prefer to interact with another “real” person who appears embodied in an avatar compared to a disembodied voice or a static image or video (see Parmar et al, 2023). Given the novelty of face- and full-body tracking, collaborative studies incorporating these methods are sparse, and earlier studies featuring expressive avatars without those tracking systems found positive effects of expressive avatars but had to utilize very experimental non-human avatars due to technological constrains (Bernal & Maes, 2017). A recent paper by Radiah and colleagues (2023) presents evidence that participants feel most comfortable with a personalized same-gender avatar as well as reporting more embodiment compared with other avatars. The authors compared personalized and non-personalized avatars which either matched the participant’s gender or were opposite to it and while participants preferred the matching personalized avatar. The study revealed several interesting insights, for one that participants prefer any avatar over no avatar at all. If participants are given a choice of avatar they prefer and feel more ownership with a personalized avatar over a non-personalized avatar - even if the personalized avatar was modelled as the opposite gender of the participant (Radiah, Roth, Alt, & Abdelrahman, 2023). The preference of an avatar is in line with other research such as a study on social VR by Freeman and Maloney (2021) who investigated the way social VR users choose to represent themselves in virtual environments. They found evidence that participants either tried to keep an avatar who was consistent with their physical self or chose avatar characteristics based on platform-specific rules, but in either case participants strongly preferred to stay with a consistent type of self-representation (Freeman & Maloney, 2021). This has implications for the generation and use of avatars in educational environments since learners could more easily get used to the avatar if the ruleset of the platform was made salient to them early on. However, this is speculative at this point and warrants further investigation.

Another form of “avatar” from a more abstract perspective could be an external object as part of the simulation. In the context of a driving simulator this could be a vehicle frame in which the VR user is placed. In Chapter 4.4. a planned human-in-the-loop driving study was described – in this case the participant would be placed within a real car or utility vehicle frame mounted on a robot arm and “drive” in a 360° dome into which a virtual environment is projected (see Bernhard,

Reinhard, Kleer, & Hecht, 2023 for an exemplary setup of the so-called RODOS simulator). While this setup is very sophisticated and costly, it provides a mixed-reality experience which is next to impossible to replicate otherwise – the level of immersion through multisensory feedback as well as the interaction with the environment through a "real" car allow for a higher fidelity experience, while maintaining the level of safety driving simulators have in comparison to real world driving research. Goedicke and colleagues (2022) presented a highly interesting mixed-reality variation of the setup described above which involved a real car in which participants wore a XR headset and had to navigate through a real course with either no obstacles, physical traffic cones or virtual traffic cones. The authors reported that participants drove similarly through the obstacle course, irrespective whether the traffic cones were "real" or virtual. Participants were also asked to rate the different conditions and rated the course with virtual obstacles as the most difficult, which was also reflected in the task completion time and error rate (Goedicke, Bremers, Lee, Bu, Yasuda, & Ju, 2022).

From an educational perspective these studies are, at the time of writing, too reliant on sophisticated software and hardware, but nevertheless can give a glimpse for some future use cases. The RODOS human-in-the-loop simulator has been fitted with an excavator chassis before, which could be used in the vocational education of construction workers, as dangerous situations can be simulated without putting either the learner or others into danger (Pause et al., 2022). For the driving research which was planned in our lab, obstacles would have been placed along a virtual road while another condition would have included other virtual drivers as traffic from the opposite road. Goedicke et al. (2022) noted that their research did not include any other traffic due to safety reasons in XR, which would have not been a concern in a fully virtual environment. To the author's knowledge, no driving research utilizing sophisticated XR or human-in-the-loop simulation have used eye-tracking to gain insights on the participant's behaviour in the simulator. In a feasibility test in early 2023 an exemplary simulation was ran in the RODOS simulator while I was wearing the mobile Tobii Glasses 2 eye-tracker. This test was intended to investigate whether any electronic interference or other technical interactions between the devices, which we planned to use, occur. However, no such interference was detected, exemplary data was collected and deemed suitable for analysis and the study would have been conceptually feasible. Future research involving large simulators such as the ones described in this paragraph should consider using a head-mounted eye-tracker to gain additional

insights into participant behaviour. I believe this could greatly assist the design of future virtual environments with educational use cases.

5.4. Future studies and outlook

While VR holds high promises to improve behavioural research and education, there exist several shortcomings in the current state of the field. A recent review from Lanier and colleagues (2019) analysed 61 articles and found that a significant portion of the published papers contain statistical errors as well as insufficient information to allow replication. The lack of replicability is problematic, since one of the expectations for VR-based research has been the increased economic validity which would come with unified methods and research designs (Blascovich, Loomis, Beall, Swinth, Hoyt, & Bailenson, 2002; Lanier et al., 2019). One research branch which is increasingly more common is the use of eye-tracking in VR-based research (Clay, König, & König, 2019). Due to the availability in customer-grade HMDs with integrated eye-tracking and pupillometry, these data can now be accessed more easily by researchers around the globe and hopefully bring a degree of comparability and replicability to this growing field. In the last few years, HMD manufacturers have also started to add additional sensors to their VR solutions, such as additional cameras to track facial expressions and heart rate (HP Reverb G2 Omnicept; Hewlett-Packard, CA, USA), inside-out tracking cameras to measure spatial parameters without external cameras (Meta Quest Pro; Meta Platforms, Inc., CA, USA) as well as controller-less hand-tracking (Pico 4 Enterprise; ByteDance, Beijing, China).

Therefore, for the outlook, researchers should consider two distinct perspectives: the technical perspective and the scientific one. This distinction is important to note since companies are eager to develop their products to gain an advantage over their competitors and thus have competitive pressure to release improved products regularly. With increasing availability of costumer-grade HMDs and tools to create virtual environments, more research groups are beginning to investigate scientific questions in virtual reality. However, as several reviews made clear, the quality of the research is often subpar and warrants improvement (Freina & Ott, 2015; Jensen & Konradsen, 2018; Lanier et al., 2019). Study 3 and study 4 were conducted with a clear research design, and data was collected and analysed in a structured and replicable way. Future research should incorporate more robust frameworks and include multiple measurements to add to our knowledge about

human behaviour in virtual and augmented reality to improve the conditions for the learners and workers of today and tomorrow.

One of the most exciting themes in working with emerging technologies is the rapid development of features. In 2019 we conducted a study with the aim to compare three ways of cognitive load: Behavioural data through error rates and other performance metrics, pupil and fixation data through an eye-tracker as well as physiological data using a dedicated device to collect electrodermal data. This mix of methods was intended to provide a base for further research and has since been developed into a full paradigm by my colleague Omar, but at the time the study ran into several issues: Missing data from the skin conductance recorder, improper lighting conditions which affected the eye-tracking data and synchronisation problems being the most prominent. If the study would be replicated nowadays, a state-of-the-art VR HMD could both present the task and simultaneously record physiological and eye-tracking data, synced automatically. Six years ago, Bernal and Maes (2017) had to come up with sophisticated "emotional beasts" to enable participants to share emotions non-verbally in VR since hand- and face-tracking weren't common – a few years later integrated face-tracking has become a key feature in commercially available HMDs (e.g. HP Reverb G2 Omnicept, Hewlett-Packard, CA, USA). Instead of relying on cables and large computers, some VR headsets became fully wireless or run completely without a computer (Meta Quest 2; Meta Platforms, Inc., CA, USA). This independence from cables or computers allows for more inclusive VR experiences for people who could in the past not afford an expensive computer or for labs where experimental space is limited. Another trend which holds potential for vocational and other forms of practical education is the push towards "room-aware" mixed-reality HMDs. These headsets utilize cameras to not only track hand and body movements but can also overlay the virtual environment on the real world (e.g. Meta Quest 3, Meta Platforms Inc., CA, USA). Future research can focus on designing and testing learning materials with the focus of blended learning experiences instead of dedicating resources to focus on either traditional learning, VR-based learning or another form of computer-assisted learning.

Chapter 6. Conclusion

6.1. Implications for the Field of Research

The present work understands itself as a showcase of experimental research which can be used as foundation for future practical and research use cases. As the reviewed literature indicate and as several studies conducted by our lab have showcased, I support the statement by Rosedale (2017) that virtual reality can and will transform the way humans learn. It is now in the responsibility of researchers and educators alike to build on this foundation and provide material to bring individually tailored learning experiences to as many people as possible, especially those who are often left behind by such programmes (see e.g. Bravou, Oikonomidou, & Drigas, 2022). Each of the four studies contained in the present work yielded some conclusions both for educational content creation, as well as virtual reality research. Study 1 provided in-depth insights on the impact of input modalities on task performance. This research could be expanded to include a broader range of participants, as it was tested with a university student sample. We stated in study 1 that a "trivial" change such as changing from finger to a pencil to interact with a tablet was sufficient to impact performance significantly. Looking back at the technological advantages in VR technology, a whole line of research could be drawn from these insights – by example the comparison of controllers versus hands-free interaction with virtual environments.

Vocational education is a good starting point for scientific investigation since the professions have a broad range of physical demands but also certain cognitive demands. As was apparent during study 2 and study 3, the vocational sector is often underrepresented in educational research which can also be seen as a chance for scientists in the future to help in the generation and evaluation of modern educational materials. Study 2 also highlighted the need for standardisation in one sector of vocational work, which could be further investigated and expanded upon by future research.

Study 3 has provided a good foundation for further research if the learning material is chosen to be more in line with the experimental sample. An immersive virtual environment would also create an interesting third condition, to investigate potential differences between a 2D video, a 360° nVR video and an immersive learning experience. Study 3 and 4 have both also provided early evidence for the viability to use eye-tracking within VR to investigate a variety of behaviours which has since been expanded by other workgroups. Future research can expand this

work by investigating blink rates, continuous pupil data or regions of interest. Research could also investigate whether there are differences in educational content by letting participants apply the newly learned content in either a virtual or potentially even a real-world setting, which at the time of writing was not yet done. Investigations into transfer learning between virtual and real-world tasks could significantly contribute to our understanding of learning in the digital age.

By adding four distinct studies and as well as additional research and an overview over some of the recent developments in the field to the growing field of research, I hope I was able to provide a small contribution to overcome some of the challenges the field is currently facing.

6.2. Personal Insight & Growth

Over the course of the last six years, I have conducted a multitude of studies, supervised several Bachelor theses and lab rotations. I've presented data in posters and talks, participated in summer schools and collaborated with partners from science and industry on research projects on vocational education, virtual reality, robotics. At the very end, fitting with the zeitgeist in late 2023, we've even started to investigate interaction with AI-powered chatbots. Out of all this work, I tried to condense the most relevant insights into the thesis you are now reading. Some of the work has never been published, is currently being analysed or simply did not fit the scope of the thesis, while several other studies and their backstory have been covered in various sections throughout the thesis. In Chapter 3.2. and 4.2. several ongoing or otherwise unpublished research was discussed briefly. This was important for two reasons, the first of which is somewhat selfish because the research presented by the papers themselves appears disjointed and incomplete without added context. The feeling of incompleteness was further fostered due to the complete halt of research activities in most of 2020 and 2021. Therefore, a key point I wanted to make with the thesis is to tell the greater, overarching story. In its final form I feel it does represent all the work my colleagues and I did better than judging the papers in a vacuum. Each study brought new insights with it, knowledge and skills which could be applied in subsequent experiments.

When I set out on this journey six years ago, I had only surface level knowledge of eye-trackers and no experience with virtual reality apart from a private interest. Thinking back to the very first study we've conducted when I started in Kaiserslautern – the field study with car mechanics in Iserlohn back in

March 2018 to investigate task performance with eye-tracking, it becomes apparent how much we, and I, have accomplished since then. Over the years I not only developed the skillset to design, run and analyse experiments with mobile and HMD-bound eye-trackers - which function fundamentally different from one another - but also contributed to a large body of research where we strived to standardize the analysis of pupil data. I cannot take full credit for the deeper analyses and especially the pupillometry algorithm because both Radha and Omar arguably had far greater contributions to the latter. Nevertheless, a small team of young scientists ran experiments, experienced the unforeseen problems so common in any research environment but continued to collect data, analyse it, and over time turned into a group of experts with a diverse skillset. The first fruits of this labour were presented in a talk on the TEAP 2023 in Trier and will soon be continued either by myself or my colleagues. As for my own skillset, it did not stop with the eye-tracking and virtual reality knowledge, I also gained invaluable experience in project management, learned to code and model in Unity and R, worked with a variety of industry stakeholders, attended conferences as participant and speaker and as a somewhat crowning achievement at the end, was personally requested by Nature as a reviewer.

All of this might not seem much to some but it did mean everything to me. The path of a PhD student is littered with insecurities, constant doubts, incessant feelings of impostor syndrome: One could always do more; your long-planned study doesn’t run half as well as you expected. Others always seem to do better while you continuously struggle, the vicious cycle seems endless. Now, standing here after six years, I can finally say I feel confident to call myself a scientist. Thank you.

References

Agrewal, S., Simon, A. M. D., Bech, S., Bærentsen, K. B., & Forchammer, S. (2020). Defining Immersion: Literature Review and Implications for Research on Audiovisual Experiences. *Journal of the Audio Engineering Society*, *68*(6), 404-417. https://doi.org 10.17743/jaes.2020.0039

Aymerich-Franch, L. & Bailenson, J. (2014) The use of doppelgangers in virtual reality to treat public speaking anxiety: A gender comparison. In *Proceedings of the International Society for Presence Research,* 173-186. USA.

Bacca, J., Baldiris, S., Fabregat, R., Kinshuk & Graf, S. (2015). Mobile augmented reality in vocational education and training. *Procedia Computer Science*, *75*, 49-58. https://doi.org/10.1016/j.procs.2015.12.203

Baceviciute, S., Terkildsen, T., & Makransky, G. (2021). Remediating learning from non-immersive to immersive media: Using EEG to investigate the effects of environmental embeddedness on reading in Virtual Reality. *Computers & Education*, *164*, 104122. https://doi.org/10.1016/j.compedu.2020.104122

Bernal, G., & Maes, P. (2017). Emotional beasts: visually expressing emotions through avatars in VR. In *Proceedings of the 2017 CHI EA conference* (pp. 2395-2402). USA. https://doi.org/10.1145/3027063.3053207

Bernhard, C., Reinhard, R., Kleer, M., & Hecht, H. (2023). A case for raising the camera: a driving simulator test of camera-monitor systems. *Human Factors*, *65*(2), 321-336. https://doi.org/10.1177/00187208211010941

Blana, D., Kyriacou, T., Lambrecht, J. M., & Chadwick, E. K. (2016). Feasibility of using combined EMG and kinematic signals for prosthesis control: A simulation study using a virtual reality environment. *Journal of Electromyography and Kinesiology*, *29*, 21-27. https://doi.org/10.1016/j.jelekin.2015.06.010

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, *13*(2), 103-124. https://doi.org/10.1207/S15327965PLI1302_01

Bonsu, N. O., Bervell, B., Armah, J. K., Aheto, S. P. K., & Arkorful, V. (2021).

WhatsApp use in teaching and learning during COVID-19 pandemic period: Investigating the initial attitudes and acceptance of students. *Library Philosophy and Practice*, 2021.

Bower, M., Lee, M. J., & Dalgarno, B. (2017). Collaborative learning across physical and virtual worlds: Factors supporting and constraining learners in a blended reality environment. *British Journal of Educational Technology*, *48*(2), 407-430. https://doi.org/10.1111/bjet.12435

Bravou, V., Oikonomidou, D., & Drigas, A. S. (2022). Applications of virtual reality for autism inclusion. A review. *Retos: Nuevas Tendencias en Educación Física, Deporte y Recreación*, 45, 779-785.

Brown, B., Reeves, S., & Sherwood, S. (2011). Into the wild: challenges and opportunities for field trial methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1657-1666), BC, Canada. https://doi.org/10.1145/1978942.1979185

Buchner, J., Buntins, K., & Kerres, M. (2022). The impact of augmented reality on cognitive load and performance: A systematic review. *Journal of Computer Assisted Learning*, *38*(1), 285-303. https://doi.org/10.1111/jcal.12617

Buettner, R., Sauer, S., Maier, C., & Eckhardt, A. (2018). Real-time prediction of user performance based on pupillary assessment via eye tracking. *AIS Transactions on Human-Computer Interaction*, *10*(1), 26-56. https://doi.org/10.17705/1thci.00103

Bujdosó, G., Novac, O. C., & Szimkovics, T. (2017). Developing cognitive processes for improving inventive thinking in system development using a collaborative virtual reality system. In *Proceedings of the 8th IEEE International Conference on Cognitive Infocommunications* (pp. 79-84), Hungary. https://doi.org/10.1109/CogInfoCom.2017.8268220

Burbules, N. C. (2016). How we use and are used by social media in education. *Educational Theory*, *66*(4), 551-565. https://doi.org/10.1111/edth.12188

Chang, F., Luo, M., Walton, G., Aguilar, L., & Bailenson, J. (2019). Stereotype threat in virtual learning environments: Effects of avatar gender and sexist behavior on women's math learning outcomes. *Cyberpsychology, Behavior,*

and Social Networking, *22*(10), 634-640. https://doi.org/10.1089/cyber.2019.0106

Choi, H. H., Van Merriënboer, J. J., & Paas, F. (2014). Effects of the physical environment on cognitive load and learning: Towards a new model of cognitive load. *Educational Psychology Review*, *26*, 225-244. https://doi.org/10.1007/s10648-014-9262-6

Clay, V., König, P., & König, S. (2019). Eye tracking in virtual reality. *Journal of Eye Movement Research*, *12*(1). https://doi.org/10.16910/jemr.12.1.3

Cognolato, M., Atzori, M., & Müller, H. (2018). Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances. *Journal of Rehabilitation and Assistive Technologies Engineering*, *5*. https://doi.org/10.1177/2055668318773991

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science, 323*(5910), 66-69. https://doi.org/10.1126/science.1167311

Dhimolea, T. K., Kaplan-Rakowski, R., & Lin, L. (2022). A systematic review of research on high-immersion virtual reality for language learning. *TechTrends*, *66*(5), 810-824. https://doi.org/10.1007/s11528-022-00717-w

Drigas, A., Mitsea, E., & Skianis, C. (2022). Virtual reality and metacognition training techniques for learning disabilities. *Sustainability*, *14*(16), 10170. https://doi.org/10.3390/su141610170

Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., Raubal, M., & 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* (pp. 1-13). https://doi.org/10.1145/3173574.3173856

Duchowski, A. T., Krejtz, K., Gehrer, N. A., Bafna, T., & Bækgaard, P. (2020). The low/high index of pupillary activity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-12). https://doi.org/10.1145/3313831.3376394

Dziuban, C., Graham, C. R., Moskal, P. D., Norberg, A., & Sicilia, N. (2018).

Blended learning: the new normal and emerging technologies. *International journal of Educational Technology in Higher Education*, *15*, 1-16. https://doi.org/10.1186/s41239-017-0087-5

Ekstrand, C., Jamal, A., Nguyen, R., Kudryk, A., Mann, J., & Mendez, I. (2018). Immersive and interactive virtual reality to improve learning and retention of neuroanatomy in medical students: a randomized controlled study. *Canadian Medical Association Open Access Journal*, *6*(1), 103-109. https://doi.org/10.9778/cmajo.20170110

Esteves, J. R., Cardoso, J. C., & Gonçalves, B. S. (2023). Design recommendations for immersive virtual reality application for English learning: A systematic review. *Computers*, *12*(11), 236-255. https://doi.org/10.3390/computers12110236

Fears, N. E., Bailey, B. C., Youmans, B., & Lockman, J. J. (2019). An eye-tracking method for directly assessing children's visual-motor integration. *Physical Therapy*, *99*(6), 797-806. https://doi.org/10.1093/ptj/pzz027

Federal Statistical Office (2022, October 27) Daten aus den Laufenden Wirtschaftsrechnungen (LWR) zur Ausstattung privater Haushalte mit Informationstechnik [Data from Ongoing Economic Inquiries on the equipment of private households with information rechnology] *https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Ausstattung-Gebrauchsgueter/Tabellen/a-infotechnik-d-lwr.html*

Ferdous, H. S., Hoang, T., Joukhadar, Z., Reinoso, M. N., Vetere, F., Kelly, D., & Remedios, L. (2019). "What's happening at that hip?" Evaluating an on-body projection based augmented reality system for physiotherapy classroom. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-12). https://doi.org/10.1145/3290605.3300464

Flegal, K. E., Ragland, J. D., & Ranganath, C. (2019). Adaptive task difficulty influences neural plasticity and transfer of training. *NeuroImage*, *188*, 111-121. https://doi.org/10.1016/j.neuroimage.2018.12.003

Fleishman, E. A., & Reilly, M. E. (1995). *Fleishman Job Analysis Survey (F-JAS)*. Management Research Institute.

Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted

eye tracking: A new method to describe infant looking. *Child Development*, *82*(6), 1738-1750. https://doi.org/10.1111/j.1467-8624.2011.01670.x

Freeman, G., & Maloney, D. (2021). Body, avatar, and me: The presentation and perception of self in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 1-27. https://doi.org/10.1145/3432938

Freina, L., & Ott, M. (2015). A literature review on immersive virtual reality in education: state of the art and perspectives. In *Proceedings of the International Scientific Conference eLearning and Software for Education* (pp. 133-141), Romania, 1. https://doi.org/10.12753/2066-026X-15-020

Goedicke, D., Bremers, A. W., Lee, S., Bu, F., Yasuda, H., & Ju, W. (2022). XR-OOM: MiXed Reality driving simulation with real cars for research and design. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 107, 1-13. https://doi.org/10.1145/3491102.3517704

Grabarczyk, P., & Pokropski, M. (2016). Perception of affordances and experience of presence in virtual reality. *Avant. The Journal of the Philosophical-Interdisciplinary Vanguard*, *7*(2), 25-44. https://doi.org/10.26913/70202016.0112.0002

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of Psychology*, *60*, 549-576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Grigorenko, E. L., Compton, D. L., Fuchs, L. S., Wagner, R. K., Willcutt, E. G., & Fletcher, J. M. (2020). Understanding, educating, and supporting children with specific learning disabilities: 50 years of science and practice. *American Psychologist*, *75*(1), 37–51. https://doi.org/10.1037/amp0000452

Göransson, K., & Nilholm, C. (2014). Conceptual diversities and empirical shortcomings–a critical analysis of research on inclusive education. *European Journal of Special Needs Education*, *29*(3), 265-280. https://doi.org/10.1080/08856257.2014.933545

Haar, S., Sundar, G., & Faisal, A. A. (2021). Embodied virtual reality for the study of real-world motor learning. *PloS one*, *16*(1), e0245717. https://doi.org/10.1371/journal.pone.0245717

Hasson, C. J., Zhang, Z., Abe, M. O., & Sternad, D. (2016). Neuromotor noise is malleable by amplifying perceived errors. *PLoS Computational Biology*, *12*(8), e1005044. https://doi.org/10.1371/journal.pcbi.1005044

Hekele, F., Bender, S., Spilski, J., Homrighausen, T., Werth, D., & Lachmann, T. (2020). *Barrier-free design of a Virtual Reality 3D environment* [Unpublished Manuscript]. Center for Cognitive Science, Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau.

Hekele, F., Spilski, J., Bender, S., & Lachmann, T. (2022). Remote vocational learning opportunities—A comparative eye-tracking investigation of educational 2D videos versus 360° videos for car mechanics. *British Journal of Educational Technology*, *53*(2), 248-268. https://doi.org/10.1111/bjet.13162

Hepperle, D., & Wölfel, M. (2023). Similarities and Differences between Immersive Virtual Reality, Real World, and Computer Screens: A Systematic Scoping Review in Human Behavior Studies. *Multimodal Technologies and Interaction*, *7*(6), 56. https://doi.org/10.3390/mti7060056

Hortensius, R., Hekele, F., & Cross, E. S. (2018). The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(4), 852-864. https://doi.org/10.1109/TCDS.2018.2826921

Hostettler, D., Mayer, S., & Hildebrand, C. (2023). Human-Like Movements of Industrial Robots Positively Impact Observer Perception. *International Journal of Social Robotics*, *15*(8), 1399-1417. https://doi.org/10.1007/s12369-022-00954-2

Hofbauer, L. M., & Rodriguez, F. S. (2023). Emotional valence perception in music and subjective arousal: Experimental validation of stimuli. *International Journal of Psychology*, 58(5), 465-475. https://doi.org/10.1002/ijop.12922

Hoogeboom, M. A., Saeed, A., Noordzij, M. L., & Wilderom, C. P. (2021). Physiological arousal variability accompanying relations-oriented behaviors of effective leaders: Triangulating skin conductance, video-based behavior coding and perceived effectiveness. *The Leadership Quarterly*, *32*(6), 101493. https://doi.org/10.1016/j.leaqua.2020.101493

Huang, R., Spector, J.M., Yang, J. (2019). Social Learning Perspective of

Educational Technology. In: *Educational Technology. Lecture Notes in Educational Technology* (pp. 107-122). Springer, Singapore. https://doi.org/10.1007/978-981-13-6643-7_7

Ichou, R.P. (2018). Can MOOCs reduce global inequality in education?. *Australasian Marketing Journal, 26*(2), 116-120. https://doi.org/10.1016/j.ausmj.2018.05.00

Kahlert, T., van de Camp, F., & Stiefelhagen, R. (2015). Learning to juggle in an interactive virtual reality environment. In *Proceedings of the International Conference on Human-Computer Interaction, USA, 528,* 196-201. https://doi.org/10.1007/978-3-319-21380-4_35

Kalyuga, S. (2005). Prior knowledge principle in multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 325–337). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.022

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need?. *Educational Psychology Review*, *23*, 1-19. https://doi.org/10.1007/s10648-010-9150-7

Kavanagh, S., Luxton-Reilly, A., Wuensche, B., & Plimmer, B. (2017). A systematic review of Virtual Reality in education. *Themes in Science and Technology Education, 10*(2), 85-119.

Kim, K. G., Oertel, C., Dobricki, M., Olsen, J. K., Coppi, A. E., Cattaneo, A., & Dillenbourg, P. (2020). Using immersive virtual reality to support designing skills in vocational education. *British Journal of Educational Technology*, *51*(6), 2199-2213. https://doi.org/10.1111/bjet.13026

Kleinmann, M., Manzey, D., Schumacher, S., & Fleishman, E.A. (2010). *Fleishman Job Analyse System für eigenschaftsbezogene Anforderungsanalysen : F-JAS; deutschsprachige Bearbeitung des Fleishman Job Analysis Survey by Edwin A. Fleishman.* Göttingen: Hogrefe.

Korbach, A., Brünken, R., & Park, B. (2017). Measurement of cognitive load in multimedia learning: a comparison of different objective measures. *Instructional Science*, *45*, 515-536. https://doi.org/10.1007/s11251-017-9413-5

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. https://doi.org/10.1371/journal.pone.0203629

Kromydas, T. (2017). Rethinking higher education and its relationship with social inequalities: past knowledge, present state and future potential. *Palgrave communications*, *3*(1), 1-12. https://doi.org/10.1057/s41599-017-0001-8

Kyriakou, P., & Hermon, S. (2019). Can I touch this? Using natural interaction in a museum augmented reality system. *Digital Applications in Archaeology and Cultural Heritage*, *12*, e00088. https://doi.org/10.1016/j.daach.2018.e00088

Lanier, M., Waddell, T. F., Elson, M., Tamul, D. J., Ivory, J. D., & Przybylski, A. (2019). Virtual reality check: Statistical power, reported results, and the validity of research on the psychology of virtual reality and immersive environments. *Computers in Human Behavior*, *100*, 70-78. https://doi.org/10.1016/j.chb.2019.06.015

Lee, K., Junghans, B. M., Ryan, M., Khuu, S., & Suttle, C. M. (2014). Development of a novel approach to the assessment of eye–hand coordination. *Journal of Neuroscience methods*, *228*, 50-56. https://doi.org/10.1016/j.jneumeth.2014.02.012

Legault, J., Zhao, J., Chi, Y. A., Chen, W., Klippel, A., & Li, P. (2019). Immersive virtual reality as an effective tool for second language vocabulary learning. *Languages*, *4*(1), 13-44. https://doi.org/10.3390/languages4010013

Li, J., Li, H., Umer, W., Wang, H., Xing, X., Zhao, S., & Hou, J. (2020). Identification and classification of construction equipment operators' mental fatigue using wearable eye-tracking technology. *Automation in Construction*, *109*, 103000. https://doi.org/10.1016/j.autcon.2019.103000

Li, B. J., Ratan, R., & Lwin, M. O. (2022). Virtual game changers: How avatars and virtual coaches influence exergame outcomes through enactive and vicarious learning. *Behaviour & Information Technology*, *41*(7), 1529-1543. https://doi.org/10.1080/0144929X.2021.1884290

Lin, L. Y. (2019). Differences between preschool children using tablets and non-tablets in visual perception and fine motor skills. *Hong Kong Journal of Occupational Therapy*, *32*(2), 118-126. https://doi.org/10.1177/1569186119888698

Ling, Y., Nefs, H. T., Morina, N., Heynderickx, I., & Brinkman, W. P. (2014). A meta-analysis on the relationship between self-reported presence and anxiety in virtual reality exposure therapy for anxiety disorders. *PloS One*, *9*(5), e96144. https://doi.org/10.1371/journal.pone.0096144

Lüdtke, O., Roberts, B. W., Trautwein, U., & Nagy, G. (2011). A random walk down university avenue: Life paths, life events, and personality trait change at the transition to university life. *Journal of Personality and Social Psychology, 101*(3), 620–637. https://doi.org/10.1037/a0023743

Makransky, G., & Petersen, G. B. (2021). The cognitive affective model of immersive learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review*, 1-22. https://doi.org/.1007/s10648-020-09586-2

Mandler, G. (2011). *A history of modern experimental psychology: From James and Wundt to cognitive science*. MIT Press. https://doi.org/10.7551/mitpress/3542.001.0001

Martín-Gutiérrez, J., Mora, C. E., Añorbe-Díaz, B., & González-Marrero, A. (2017). Virtual technologies trends in education. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(2), 469-486. https://doi.org/10.12973/eurasia.2017.00626a

Matic, A., & Gomez-Marin, A. (2019). A customizable tablet app for hand movement research outside the lab. *Journal of Neuroscience Methods*, *328*, 108398. https://doi.org/10.1016/j.jneumeth.2019.108398

Meghanathan, R. N., Ruediger-Flore, P., Hekele, F., Spilski, J., Ebert, A., & Lachmann, T. (2021). Spatial Sound in a 3D Virtual Environment: All Bark and No Bite?. *Big Data and Cognitive Computing*, *5*(4), 79-84. https://doi.org/10.3390/bdcc5040079

Milgram, P., Takemura, H., Utsumi, A., & Kishino, F. (1995). Augmented reality: A

class of displays on the reality-virtuality continuum. In *Proceedings of the Telemanipulator and Telepresence Technologies, USA,* 2351, 282-292. https://doi.org/10.1117/12.197321

Moore, G. E. (2006). Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp. 82-85. *IEEE solid-state circuits society newsletter*, *11*(3), 33-35. https://doi.org/10.1109/JPROC.1998.658762

Oh, C., Herrera, F., & Bailenson, J. (2019). The Effects of Immersion and Real-World Distractions on Virtual Social Interactions. *Cyberpsychology, Behavior, and Social Networking, 22*(6), 365–372. https://doi.org/10.1089/cyber.2018.0404

Pause, V., Emmerich, S., Steidel, S., Reinhard, R., Kleer, M., Kleeberg, V., Weber, J. & Zenner, T. (2022). Simulator-based development of a stability assistant for wheeled excavators. In *International Commercial Vehicle Technology Symposium* (pp. 3-14). Germany. https://doi.org/10.1007/978-3-658-40783-4_1

Parmar, D., Lin, L., Dsouza, N., Joerg, S., Leonard, A. E., Daily, S. B., & Babu, S. (2023). How immersion and self-avatars in VR affect learning programming and computational thinking in middle school education. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2022.3169426

Parong, J., & Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology, 110*(6), 785-797. https://doi.org/10.1037/edu0000241

Patney, A., Salvi, M., Kim, J., Kaplanyan, A., Wyman, C., Benty, N., Luebke, D.A., & Lefohn, A. (2016). Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics*, *35*(6), 1-12. https://doi.org/10.1145/2980179.2980246

Phuthong, T. (2021). Antecedents influencing the adoption of collaborative learning social-media platforms among Thai university students during the covid-19 'new normal' era. *International Journal of Emerging Technologies in Learning*, *16*(13), 108-127. https://doi.org/10.3991/ijet.v16i13.18083

Piumsomboon, T., Lee, G., Lindeman, R. W., & Billinghurst, M. (2017). Exploring

natural eye-gaze-based interaction for immersive virtual reality. In *2017 IEEE symposium on 3D user interfaces* (pp. 36-39). CA, USA. https://doi.org/10.1109/3DUI.2017.7893315

Plass, J., Kalyuga, S., & Leutner, D. (2010). Individual Differences and Cognitive Load Theory. In J. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive Load Theory* (pp. 65-88). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511844744.006

Radiah, R., Roth, D., Alt, F., & Abdelrahman, Y. (2023). The Influence of Avatar Personalization on Emotions in VR. *Multimodal Technologies and Interaction*, *7*(4), 38. https://doi.org/10.3390/mti7040038

Rickinson, M. (2001). Learners and Learning in Environmental Education: A critical review of the evidence. *Environmental Education Research, 7*(3), 207–320. https://doi.org/10.1080/13504620120065230

Rigby, J. M., Brumby, D. P., Gould, S. J. J., & Cox, A. L. (2019). Development of a questionnaire to measure immersion in video media: The film IEQ. *Proceedings of the 2019 ACM international conference on interactive experiences for TV and online video*, 35-46. https://doi.org/10.1145/3317697.3323361

Rodriguez, F. S., Spilski, J., Schneider, A., Hekele, F., Lachmann, T., Ebert, A., & Rupprecht, F. A. (2019). Relevance of the assessment mode in the digital assessment of processing speed. *Journal of Clinical and Experimental Neuropsychology*, *41*(7), 730-739. https://doi.org/10.1080/13803395.2019.1616079

Rodriguez, F. S., Spilski, J., Hekele, F., Beese, N. O., & Lachmann, T. (2020). Physical and cognitive demands of work in building construction. *Engineering, Construction and Architectural Management*, *27*(3), 745-764. https://doi.org/10.1108/ECAM-04-2019-0211

Ruotolo, F., Maffei, L., Di Gabriele, M., Iachini, T., Masullo, M., Ruggiero, G., & Senese, V. P. (2013). Immersive virtual reality and environmental noise assessment: An innovative audio–visual approach. *Environmental Impact Assessment Review*, *41*, 10-20. https://doi.org/10.1016/j.eiar.2013.01.007

Sarupuri, B., Kulpa, R., Aristidou, A., & Multon, F. (2023). Dancing in virtual reality

as an inclusive platform for social and physical fitness activities: a survey. *The Visual Computer*, 1-16. https://doi.org/10.1007/s00371-023-03068-6

Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments, 10*(3), 266-281. https://doi.org/10.1162/105474601300343603

Shohieb, S. M., Doenyas, C., & Elhady, A. M. (2022). Dynamic difficulty adjustment technique-based mobile vocabulary learning game for children with autism spectrum disorder. *Entertainment Computing*, *42*, 100495. https://doi.org/10.1016/j.entcom.2022.100495

Slater, M., Usoh, M., & Steed, A. (1994). Depth of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *3*(2), 130-144. https://doi.org/10.1162/pres.1994.3.2.130

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *6*(6), 603-616. https://doi.org/10.1162/pres.1997.6.6.603

Slater, M. (2018). Immersion and the illusion of presence in virtual reality. *British Journal of Psychology*, *109*(3), 431-433. https://doi.org/10.1111/bjop.12305

Statistisches Bundesamt. (2023a, December). *Bildung, Forschung und Kultur: Bildungsindikatoren*. [Education, Science and Culture: Indicators of Education] German Federal Statistical Office. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Bildungsindikatoren/_inhalt.html

Statistisches Bundesamt. (2023b, December). *Studienanfänger: Deutschland, Semester, Nationalität, Geschlecht.* [First semester students: Germany, Semester, Nationality, Gender]. German Federal Statistical Office, GENESIS Data Base. https://www-genesis.destatis.de/genesis/online

Stawski, R. S., Sliwinski, M. J., & Smyth, J. M. (2006). Stress-related cognitive interference predicts cognitive function in old age. *Psychology and Aging, 21*(3), 535–544. https://doi.org/10.1037/0882-7974.21.3.535

Steed, A., Pan, Y., Zisch, F., & Steptoe, W. (2016, March). The impact of a self-avatar on cognitive load in immersive virtual reality. In: *Proceedings of the 2016 IEEE virtual reality (VR)* (pp. 67-76). https://doi.org/10.1109/VR.2016.7504689

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257-285. https://doi.org/10.1016/0364-0213(88)90023-7

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Sweller, J., van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296. https://doi.org/10.1023/A:1022193728205

Sweller, J., Ayres, P., & Kalyuga, S.(2011). Intrinsic and extraneous cognitive load. In *Cognitive Load Theory*, 57-69. https://doi.org/10.1007/978-1-4419-8126-4_5

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, *68*(1), 1-16. https://doi.org/10.1007/s11423-019-09701-3

Tao, Y., & Lopes, P. (2022). Integrating real-world distractions into virtual reality. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (pp. 1-16). PA, USA. https://doi.org/10.1145/3526113.3545682

Taylor, F. W. (1911). *The Principles of Scientific Management* (1st ed.). Harper.

Wissmath, B., Stricker, D., Weibel, D., Siegenthaler, E., & Mast, F. W. (2010). The illusion of being located in dynamic virtual environments. Can eye movement parameters predict spatial presence?. *Journal of Eye Movement Research*, *3*(5). https://doi.org/10.16910/jemr.3.5.2

Xiong, J., Hsiang, E. L., He, Z., Zhan, T., & Wu, S. T. (2021). Augmented reality and virtual reality displays: emerging technologies and future perspectives. *Light: Science & Applications*, *10*(1), 216. https://doi.org/10.1038/s41377-021-00658-8

Zhang, Z., & Sternad, D. (2021). Back to reality: differences in learning strategy in a simplified virtual and a real throwing task. *Journal of Neurophysiology*, *125*(1), 43-62. https://doi.org/10.1152/jn.00197.2020

Zizza, C., Starr, A., Hudson, D., Nuguri, S. S., Calyam, P., & He, Z. (2018). Towards a social virtual reality learning environment in high fidelity. In *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)* (pp. 1-4). https://doi.org/10.1109/CCNC.2018.8319187

References (Study 1)

Alfonsson, S., Maathz, P., & Hursti, T. (2014). Inter-format reliability of digital psychiatric self-report questionnaires: A systematic review. *Journal of Medical Internet Research, 16*(12), e268. https://doi.org/10.2196/jmir.3395

Bando, S., Asano, H., & Nozawa, A. (2017). Analysis of physiological effect of reading books by paper and electronic medium. *Electronics and Communications in Japan, 100*(5), 44–50. https://doi.org/10.1002/ecj.11956

Buck, K. K., Atkinson, T. M., & Ryan, J. P. (2008). Evidence of practice effects in variants of the trail making test during serial assessment. *Journal of Clinical and Experimental Neuropsychology, 30*(3), 312–318. https://doi.org/10.1080/13803390701390483

Canini, M., Battista, P., Della Rosa, P. A., Catricalà, E., Salvatore, C., Gilardi, M. C., & Castiglioni, I. (2014). Computerized neuropsychological assessment in aging: Testing efficacy and clinical ecology of different interfaces. *Computational and Mathematical Methods in Medicine, 2014*, 1–13. https://doi.org/10.1155/2014/804723

Carpenter, R., & Alloway, T. (2018). Computer versus paper-based testing: Are they equivalent when it comes to working memory? *Journal of Psychoeducational Assessment*, *37*(3), 382-394. https://doi.org/10.1177/0734282918761

Crowe, S. F. (1998). The differential contribution of mental tracking, cognitive flexibility, visual search, and motor speed to performance on parts A and B of the trail making test. *Journal of Clinical Psychology, 54*(5), 585–591.

https://doi.org/10.1002/(SICI)1097-4679(199808)54:5<585::AID-JCLP4>3.0.CO;2-K

Demaree, H. A., DeLuca, J., Gaudino, E. A., & Diamond, B. J. (1999). Speed of information processing as a key deficit in multiple sclerosis: Implications for rehabilitation. *Journal of Neurology, Neurosurgery, and Psychiatry, 67*(5), 661–663. http://doi.org/10.1136/jnnp.67.5.661

Drapeau, C. E., Bastien-Toniazzo, M., Rous, C., & Carlier, M. (2007). Nonequivalence of computerized and paper-andpencil versions of trail making test. *Perceptual and Motor Skills, 104*(3), 785–791. https://doi.org/10.2466/pms.104.3.785-791

Fellows, R. P., Dahmen, J., Cook, D., & Schmitter- Edgecombe, M. (2016). Multicomponent analysis of a digital trail making test. *The Clinical Neuropsychologist, 31*(1), 154–167. https://doi.org/10.1080/13854046.2016.1238510

Gates, N. J., & Kochan, N. A. (2015). Computerized and online neuropsychological testing for late-life cognition and neurocognitive disorders: Are we there yet? *Current Opinion in Psychiatry, 28*(2), 165–172. https://doi.org/10.1097/YCO.0000000000000141

Gerth, S., Dolk, T., Klassert, A., Fliesser, M., Fischer, M. H., Nottbusch, G., & Festman, J. (2016a). Adapting to the surface: A comparison of handwriting measures when writing on a tablet computer and on paper. *Human Movement Science, 48*, 62–73. https://doi.org/10.1016/j.humov.2016.04.006

Gerth, S., Klassert, A., Dolk, T., Fliesser, M., Fischer, M. H., Nottbusch, G., & Festman, J. (2016b). Is handwriting performance affected by the writing surface? Comparing preschoolers', second graders', and adults' writing performance on a tablet vs. paper. *Frontiers in Psychology, 7*, 1308. https://doi.org/10.3389/fpsyg.2016.01308

Gwaltney, C. J., Shields, A. L., & Shiffman, S. (2008). Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: A metanalytic review. *Value in Health, 11*(2), 322–333. https://doi.org/10.1111/j.1524-4733.2007.00231.x

Hajcak, G., McDonald, N., & Simons, R. F. (2003). To err is autonomic: Error-

related brain potentials, ANS activity, and post-error compensatory behavior. *Psychophysiology, 40*(6), 895–903. https://doi.org/10.1111/1469-8986.00107

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109*(4), 679-709. https://doi.org/10.1037/0033-295X.109.4.679

Jenkins, A., Lindsay, S., Eslambolchilar, P., Thornton, I. M., & Tales, A. (2016). Administering cognitive tests through touch screen tablet devices: Potential issues. *Journal of Alzheimers Disease, 54*(3), 1169-1182. https://doi.org/1169–1182. 10.3233/JAD-160545

Krakauer, J. W., Mazzoni, P., Ghazizadeh, A., Ravindran, R., & Shadmehr, R. (2006). Generalization of motor learning depends on the history of prior action. *PLOS Biology, 4*(10), e316. https://doi.org/10.1371/journal.pbio.0040316

Martin, T. A., Hoffman, N. M., & Donders, J. (2003). Clinical utility of the trail making test ratio score. *Applied Neuropsychology, 10*(3), 163–169. https://doi.org/10.1207/S15324826AN1003_05

Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are Internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. *Organizational Research Methods, 10*(2), 322–345. https://doi.org/10.1177/1094428106289393

Misdraji, E. L., & Gass, C. S. (2010). The trail making test and its neurobehavioral components. *Journal of Clinical and Experimental Neuropsychology, 32*(2), 159-163. https://doi.org/10.1080/13803390902881942

Montag, C., Duke, É., & Markowetz, A. (2016). Toward psychoinformatics: Computer science meets psychology. *Computational and Mathematical Methods in Medicine, 2016,* 1–10. https://doi.org/10.1155/2016/2983685

Muehlhausen, W., Doll, H., Quadri, N., Fordham, B., O'Donohoe, P., Dogar, N., & Wild, D. J. (2015). Equivalence of electronic and paper administration of patient-reported outcome measures: A systematic review and meta-analysis of studies conducted between 2007 and 2013. *Health and Quality of Life Outcomes, 13*(1), https://doi.org/10.1186/s12955-015-0362-x

Mulert, C., Gallinat, J., Dorn, H., Herrmann, W. M., & Winterer, G. (2003). The

relationship between reaction time, error rate and anterior cingulate cortex activity. *International Journal of Psychophysiology, 47*(2), 175–183. https://doi.org/10.1016/S0167-8760(02)00125-3

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics, 51*(9), 1352–1375. https://doi.org/10.1080/00140130802170387

Oswald, W., & Roth, E. (1987). *Der Zahlenverbindungstest* [The connecting numbers test]. Göttingen: Hogrefe.

Prins, N. D., van Dijk, E. J., Den Heijer, T., Vermeer, S. E., Jolles, J., Koudstaal, P. J., Hofman, A., & Breteler, M. M. B. (2005). Cerebral small-vessel disease and decline in information processing speed, executive function and memory. *Brain, 128*(9), 2034–2041. https://doi.org/10.1093/brain/awh553

Reitan, R. M. (1958). Validity of the trail making test as an indicator of organic brain damage. *Perceptual and Motor Skills, 8*(3), 271–276. https://doi.org/10.2466/pms.1958.8.3.271

Riva, G., Teruzzi, T., & Anolli, L. (2003). The use of the internet in psychological research: Comparison of online and offline questionnaires. *Cyberpsychology & Behavior, 6*(1), 73–80. https://doi.org/10.1089/109493103321167983

Salthouse, T. A., & Fristoe, N. M. (1995). Process analysis of adult age effects on a computer-administered trail making test. *Neuropsychology, 9*(4), 518-528. https://doi.org/10.1037/0894-4105.9.4.518

Salthouse, T. A., Toth, J., Daniels, K., Parks, C., Pak, R., Wolbrette, M., & Hocking, K. J. (2000). Effects of aging on efficiency of task switching in a variant of the trail making test. *Neuropsychology, 14*(1), 102-111. https://doi.org/10.1037/0894-4105.14.1.102

Sanchez-Cubillo, I., Perianez, J. A., Adrover-Roig, D., Rodriguez-Sanchez, J. M., Rios-Lago, M., Tirapu, J., & Barceló, F. (2009). Construct validity of the trail making test: Role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *Journal of the International Neuropsychological Society, 15*(3), 438–450. https://doi.org/10.1017/S1355617709090626

Shanahan,M. A., Pennington, B. F., Yerys, B. E., Scott, A., Boada, R., Willcutt, E. G., Olson, R. K., & DeFries, J. C. (2006). Processing speed deficits in attention deficit/hyperactivity disorder and reading disability. *Journal of*

Abnormal Child Psychology, 34(5), 584. https://doi.org/10.1007/s10802-006-9037-8

Tombaugh, T. N. (2004). Trail making test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology, 19*(2), 203–214. https://doi.org/10.1016/S0887-6177(03)00039-8

Van Ballegooijen, W., Riper, H., Cuijpers, P., van Oppen, P., & Smit, J. H. (2016). Validation of online psychometric instruments for common mental health disorders: A systematic review. *BMC Psychiatry, 16*(1), 45. https://doi.org/10.1186/s12888-016-0735-7

Vora, J. P., Varghese, R., Weisenbach, S. L., & Bhatt, T. (2016). Test-retest reliability and validity of a custom-designed computerized neuropsychological cognitive test battery in young healthy adults. *Journal of Psychology and Cognition, 1*(1), 11–19. http://doi.org/10.35841/psychology-cognition.1.1.11-19

Zabramski, S. (2011). Careless touch: A comparative evaluation of mouse, pen, and touch input in shape tracing task. *Proceedings of the 23rd Australian Computer-Human Interaction Conference, 11*, 329–332. https://doi.org/10.1145/2071536.2071588

Zabramski, S., & Stuerzlinger, W. (2012). The effect of shape properties on ad-hoc shape replication with mouse, pen, and touch input. *Proceeding of the 16th International Academic MindTrek Conference, 12*, 275–278. https://doi.org/10.1145/2393132.2393192

Zygouris, S., & Tsolaki, M. (2014). Computerized cognitive testing for older adults: A review. *American Journal of Alzheimers Disease and Other Dementias, 30*(1), 13–28. https://doi.org/10.1177/1533317514522852

References (Study 2)

Abdel-Wahab, M. & Vogl, B. (2011). Trends of productivity growth in the construction industry across Europe, US and Japan. *Construction Management and Economics*, *29*(6), 635-644. https://doi.org/10.1080/01446193.2011.573568

Alabdulkarim, S. & Nussbaum, M.A. (2019). Influences of different exoskeleton

designs and tool mass on physical demands and performance in a simulated overhead drilling task. *Applied Ergonomics*, *74*, 55-66. https://doi.org/10.1016/j.apergo.2018.08.004

Arndt, V., Rothenbacher, D., Brenner, H., Fraisse, E., Zschenderlein, B., Daniel, U., Schuberth, S., & Fliedner, T.M. (1996). Older workers in the construction industry: results of a routine health examination and a five year follow up. *Occupational and Environmental Medicine*, *53*(10), 686-691. http://dx.doi.org/10.1136/oem.53.10.686

Bakker, A.B., van Veldhoven, M., & Xanthopoulou, D. (2010). Beyond the demand-control model thriving on high job demands and resources. *Journal of Personnel Psychology, 9*(1), 3-16. https://doi.org/10.1027/1866-5888/a000006

Boschman, J.S., van der Molen, H.F., Sluiter, J.K., & Frings-Dresen, M.H.W. (2013). Psychosocial work environment and mental health among construction workers. *Applied Ergonomics*, *44*(1), 748-755. https://doi.org/10.1016/j.apergo.2013.01.004

Boschman, J.S., van der Molen, Henk, F., Sluiter, J.K., & Frings-Dresen, M.H.W. (2011). Occupational demands and health effects for bricklayers and construction supervisors: a systematic review. *American Journal of Industrial Medicine*, *54*(1), 55-77. https://doi.org/10.1002/ajim.20899

Bowen, P., Govender, R., & Edwards, P. (2014). Structural equation modeling of occupational stress in the construction industry. *Journal of Construction Engineering and Management*, *140*(9). https://doi.org/10.1061/(ASCE)CO.1943-7862.0000877

Campion, M.A., Morgeson, F.P., & Mayfield, M.S. (1999). O* NET's theoretical contributions to job analysis research. In Peterson, N.G., Mumford, M.D., Borman, W.C., Jeanneret, P.R. & Fleishman, E.A. (Eds), *An Occupational Information System for the 21st Century: The development of O* NET* , USA, 297-304. American Psychological Association, https://doi.org/10.1037/10313-020

Chen, J., Song, X., & Lin, Z. (2016). Revealing the 'invisible gorilla' in construction: estimating construction safety through mental workload assessment. *Automation in Construction*, *63*, 173-183. https://doi.org/10.1016/j.autcon.2015.12.018

Cunnigham, J.W., Powell,T.E., Wimpee,W.E., Wilson, M.A., & Ballentine, R.D.

(1996). Ability-requirement factors for general job elements. *Military Psychology*, *8*(3), 219-234. https://doi.org/10.1207/s15327876mp0803_6

Dainty, A.R.J., Ison, S.G., & Briscoe, G.H. (2005). The construction labour market skills crisis: the perspective of small-medium-sized firms. *Construction Management and Economics*, *23*(4), 387-398. https://doi.org/10.1080/0144619042000326738

Eaves, S., Gyi, D.E., & Gibb, A.G.F. (2016). Building healthy construction workers: their views on health, well-being and better workplace design. *Applied Ergonomics*, *54*, 10-18. https://doi.org/10.1016/j.apergo.2015.11.004

Federal Health Monitoring Information System (2019). *Durchschnittliches Zugangsalter bei Renten wegen verminderter Erwerbsfähigkeit in der Gesetzlichen Rentenversicherung. Gliederungsmerkmale: Jahre, Deutschland, Geschlecht, 1. Diagnose (ICD-10), Rentenversicherungszweig*, Gesundheitsberichterstattung des Bundes. Retrieved April 10, 2018, from www.gbe-bund.de/oowa921-install/servlet/oowa/aw92/WS0100/_XWD_FORMPROC?TARGET=&PAGE=_XWD_2&OPINDEX=3&HANDLER=_XWD_CUBE.SETPGS&DATACUBE=_XWD_30&D.003=43&D.531=1000047

Federal Reserve Bank of St Louis (2018). *Gross domestic product: private industries: construction for United States metropolitan portion.* Federal Reserve Bank of St Louis. Retrieved February 11, 2019, from https://fred.stlouisfed.org/series/NGMPCONSTUSMP

Federation of the German Construction Industry (Hauptverband der Deutschen Bauindustrie) (2019). *Verteilung der Betriebe im Bauhauptgewerbe in Deutschland nach Beschäftigtengrößenklassen im Jahr 2017.* Statistisches Bundesamt. Retrieved February 11, 2019, from de.statista.com/statistik/daten/studie/152113/umfrage/struktur-der-unternehmen-im-bauhauptgewerbe-in-deutschland-2007

Fleishman, E.A. (1979). Evaluating physical abilities required by jobs. *Personnel Administrator*, *24*(6), 82-90.

Fleishman, E.A., & Mumford, M.D. (1991). Evaluating classifications of job behavior: A construct validation of the ability requirement scales. *Personnel Psychology, 44*(3), 523-575. https://doi.org/10.1111/j.1744-6570.1991.tb02403.x

Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analysis. *Journal of General Internal Medicine, 22*(11), 1596-1602. https://doi.org/10.1007/s11606-007-0333-y

Graham, J.W., Hofer, S.M., & Piccinin, A.M. (1994). Analysis with missing data in drug prevention research. In *NIDA Research Monograph, 142*, 13-64.

Gravseth, H.M., Lund, J., & Wergeland, E. (2006). Risk factors for accidental injuries in the construction industry. *Tidsskrift for den Norske laegeforening: tidsskrift for praktisk medicin, ny raekke, 126*(4), 453-456.

Hakanen, J.J., Schaufeli, W.B., & Ahola, K. (2008). The job demands-resources model: A three-year cross-lagged study of burnout, depression, commitment, and work engagement. *Work and Stress, 22*(3), 224-241. https://doi.org/10.1080/02678370802379432

Handel, M.J. (2016). The O*NET content model: Strengths and limitations. *Journal for Labour Market Research, 49*(2), 157-176. https://doi.org/10.1007/s12651-016-0199-8

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the human factors and ergonomics society annual meeting,* USA, *50*(9), 904-908. https://doi.org/10.1177/154193120605000909

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology, 52*, 139-183. https://doi.org/10.1016/S0166-4115(08)62386-9

Herbig, B., Seibt, R., Lang, J., Böckelmann, I., Darius, S., Gauggel, B., Meifort, J., Müller, A., Oldenburg, M., Stahlkopf, H., Wegner, R., & Angerer, P. (2012). Messung psychischer Belastungen. Ausgewählte Methoden und Anwendungsfelder. *Arbeitsmedizin, Sozialmedizin, Umweltmedizin, 47*(4), 252-268.

Hogan, J.C., Ogden, G.D., Gebhardt, D.L., & Fleishman, E.A. (1980). Reliability and validity of methods for evaluating perceived physical effort. *Journal of Applied Psychology, 65*(6), 672-679. https://doi.org/10.1037/0021-9010.65.6.672

Huang, G.D., Feuerstein, M., Kop, W.J., Schor, K., & Arroyo, F. (2003). Individual and combined impacts of biomechanical and work organization factors in work-related musculoskeletal symptoms. *American Journal of Industrial Medicine, 43*(5), 495-506. https://doi.org/10.1002/ajim.10212

Ilies, R., Dimotakis, N., & De Pater, I.E. (2010). Psychological and physiological reactions to high workloads: Implications for well-being. *Personnel Psychology, 63*(2), 407-436. https://doi.org/10.1111/j.1744-6570.2010.01175.x

Interian, A., Allen, L.A., Gara, M.A., Escobar, J.I., & Díaz-Martínez, A.M. (2006). Somatic complaints in primary care. Further examining the validity of the Patient Health Questionnaire (PHQ-15). *Psychosomatics, 47*(5), 392-398. https://doi.org/10.1176/appi.psy.47.5.392

Jacobi, F. (2009). Nehmen psychische Störungen zu?. *Report Psychologie, 34*(1), 16-28. https://10.1055/s-2008-1067526

Jahedi, S., & Méndez, F. (2014). On the advantages and disadvantages of subjective measures. *Journal of Economic Behavior & Organization, 98*, 97-114. https://doi.org/10.1016/j.jebo.2013.12.016

Kaplinski, O. (2018). Innovative solutions in construction industry. Review of 2016–2018 events and trends. *Engineering Structures and Technologies, 10*(1), 27-33. https://doi.org/10.3846/est.2018.1469

Karasek, R., Baker, D., Marxer, F., Ahlbom, A., & Theorell, T. (1981). Job decision latitude, job demands, and cardiovascular disease: A prospective study of Swedish men. *American Journal of Public Health, 71*(7), 694-705. https://doi.org/10.2105/AJPH.71.7.694

Kleinmann, M., Manzey, D., Schumacher, S., & Fleishman, E.A. (2010). *Fleishman Job Analyse System für eigenschaftsbezogene Anforderungsanalysen: Deutschsprachige Bearbeitung des Fleishman Job Analysis Survey by Edwin A. Fleishman*. Göttingen: Hogrefe.

Knapp, B. G., Seven, S. A., Muckler, F. A., & Akman, A. (1991). Abilities Evaluation Method for Military Intelligence Personnel. In *Proceedings of the Human Factors Society Annual Meeting,* (pp. 1321-1325). Sage Publications. https://doi.org/10.1177/154193129103501815

Kocalevent, R.D., Hinz, A., & Brahler, E. (2013). Standardization of a screening instrument (PHQ-15) for somatization syndromes in the general population. *BMC Psychiatry, 13*, 91. https://doi.org/10.1186/1471-244X-13-91

Kristensen, T.S., Borritz, M., Villadsen, E., & Christensen, K.B. (2005). The Copenhagen Burnout Inventory: A new tool for the assessment of burnout. *Work and Stress, 19*(3), 192-207. https://doi.org/10.1080/02678370500297720

Kristensen, T.S., Hannerz, H., Hogh, A., & Borg, V. (2005). The Copenhagen Psychosocial Questionnaire – a tool for the assessment and improvement of the psychosocial work environment. *Scandinavian Journal of Work, Environment & Health, 31*(6), 438-449. http://doi.org/10.5271/sjweh.948

Kroenke, K., Spitzer, R.L., & Williams, J.B.W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*(9), 606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kroenke, K., Spitzer, R.L., & Williams, J.B.W. (2002). The PHQ-15: Validity of a new measure for evaluating the severity of somatic symptoms. *Psychosomatic Medicine, 64*(2), 258-266. http://doi.org/10.1097/00006842-200203000-00008

Leung, M.Y., Liang, Q., & Olomolaiye, P. (2015). Impact of job stressors and stress on the safety behavior and accidents of construction workers. *Journal of Management in Engineering, 32*(1). https://doi.org/10.1061/(ASCE)ME.1943-5479.0000373

Lingard, H., & Francis, V. (2005). Does work–family conflict mediate the relationship between job schedule demands and burnout in male construction professionals and managers?. *Construction Management and Economics, 23*(7), 733-745. https://doi.org/10.1080/01446190500040836

Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis. *Canadian Medical Association Journal, 184*(3), 191-196. https://doi.org/10.1503/cmaj.110829

Martin, A., Rief, W., Klaiberg, A., & Braehler, E. (2006). Validity of the brief Patient Health Questionnaire mood scale (PHQ-9) in the general population. *General Hospital Psychiatry, 28*(1), 71-77. https://doi.org/10.1016/j.genhosppsych.2005.07.003

Mausner-Dorsch, H., & Eaton, W.W. (2000). Psychosocial work environment and depression. Epidemiologic assessment of the demand-control model. *American Journal of Public Health, 90*(11), 1765-1770. https:// doi.org/10.2105/ajph.90.11.1765

Merlino, L.A., Rosecrance, J.C., Anton, D., & Cook, T.M. (2003). Symptoms of musculoskeletal disorders among apprentice construction workers. *Applied Occupational and Environmental Hygiene, 18*(1), 57-64. https://doi.org/10.1080/10473220301391

Nahrgang, J.D., Morgeson, F.P., & Hofmann, D.A. (2011). Safety at work. A meta-analytic investigation of the link between job demands, job resources, burnout, engagement, and safety outcomes. *Journal of Applied Psychology, 96*(1), 71-94. https://doi.org/10.1037/a0021484

Nepal, M.P., Park, M., & Son, B. (2006). Effects of schedule pressure on construction performance. *Journal of Construction Engineering and Management, 132*(2), 182-188. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:2(182)

Nomura, K., Nakao, M., Sato, M., Ishikawa, H., & Yano, E. (2007). The association of the reporting of somatic symptoms with job stress and active coping among Japanese white-collar workers. *Journal of Occupational Health, 49*(5), 370-375. https://doi.org/10.1539/joh.49.370

Peterson, N.G., Mumford, M.D., Borman, W.C., Jeanneret, P.R., Fleishman, E.A., Levin, K.Y., ..., & Dye, D.M. (2001). Understanding work using the Occupational Information Network (O* NET): Implications for practice and research. *Personnel Psychology, 54*(2), 451-492. https://doi.org/10.1111/j.1744-6570.2001.tb00100.x

Pichyangkul, S., Yongvanitchit, K., Limsalakpetch, A., ..., & Saunders, D.L. (2015). Tissue distribution of memory T and B cells in rhesus monkeys following influenza A infection. *Journal of Immunology, 195*(9), 4378-4386. https://doi.org/10.4049/jimmunol.1501702

Roth, E.M. (2008). Uncovering the requirements of cognitive work. *Human Factors, 50*(3), 475-480. https://doi.org/10.1518/001872008X288556

Rüßmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P., & Harnisch, M. (2015). Industry 4.0: The future of productivity and growth in manufacturing industries. *Boston Consulting Group, 9*(1), 54-89.

Rutešić, S., Ćetković, J., Žarković, M., Knežević, M. & Vatin, N. (2015). Analysis of the situation in montenegrin civil engineering sector from the point of application of national regulations and the EU technical standards in construction. *Procedia Engineering, 117*, 900-910. https://doi.org/10.1016/j.proeng.2015.08.175

Sakano, J., Yamazaki, Y., Sekiya, E., & Uehata, T. (1995). The relation between job characteristics, job-related stress, and health related symptoms among middle-aged male workers in Japan. *Journal of Science of Labour, 71*(1), 1-12.

Schneider, S., Griffin, M., & Chowdhury, R. (1998). Ergonomic exposures of

construction workers: An analysis of the US Department of Labor Employment and Training Administration Database on job demands. *Applied Occupational and Environmental Hygiene, 13*(4), 238-241. https://doi.org/10.1080/1047322X.1998.10390074

Seidler, A., Thinschmidt, M., Deckert, S., Then, F., Hegewald, J., Nieuwenhuijsen, K., & Riedel-Heller, S.G. (2014). The role of psychosocial working conditions on burnout and its core component emotional exhaustion – a systematic review. *Journal of Occupational Medicine and Toxicology, 9*(1), https://doi.org/10.10.1186/1745-6673-9-10

Sexton, M., & Barrett, P. (2003). Appropriate innovation in small construction firms. *Construction Management and Economics, 21*(6), 623-633. https://doi.org/10.1080/0144619032000134156

Shimazu, A., & Jonge, J. de (2009). Reciprocal relations between effort-reward imbalance at work and adverse health: A three-wave panel survey. *Social Science & Medicine, 68*(1), 60-68. https://doi.org/10.1016/j.socscimed.2008.09.055

Siegrist, J., & Dragano, N. (2008). Psychosocial stress and disease risks in occupational life. Results of international studies on the demand-control and the effort-reward imbalance models. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz, 51*(3), 305-312. https://doi.org/10.1007/s00103-008-0461-5

Sobeih, T.M., Salem, O., Daraiseh, N., ... Genaidy, A. (2006). Psychosocial factors and musculoskeletal disorders in the construction industry: A systematic review. *Theoretical Issues in Ergonomics Science, 7*(3), 329-344. https://doi.org/10.1080/14639220500090760

Sonnega, A., Helppie-McFall, B., Hudomiet, P., Willis, R.J. Fisher, G.G. (2017). A comparison of subjective and objective job demands and fit with personal resources as predictors of retirement timing in a national US sample. *Work, Aging and Retirement, 4*(1), 37-51. https://doi.org/10.1093/workar/wax016

Stadler, P., Bayerisches Landesamt für Gesundheit (2012). *Arbeitsschutz: Psychische Belastung von Bauleitern.* Retrieved February 14, 2019, from www.lgl.bayern.de/downloads/arbeitsschutz/arbeitspsychologie/doc/bauleiter.pdf

Statista (2018). P*rognostizierte Umsatzentwicklung in der Branche Baugewerbe in*

Deutschland in den Jahren von 2006 bis 2022 (in Milliarden Euro). Retrieved February 11, 2019, from: https://de.statista.com/statistik/daten/studie/247959/umfrage/prognose-zum-umsatz-im-baugewerbe-in-deutschland/

Statista (2019). *Projected revenue of the US commercial building construction industry from 2012 to 2017 (in billion US dollars).* Retrieved February 11, 2019, from: www.statista.com/statistics/244689/projected-revenue-of-the-us-commercial-building-construction-industry/

Thayaparan, M., Siriwardena, M., Malalgoda, C., Amaratunga, D., Kaklauskas, A., & Lill, I. (2010). Reforming HEI to improve skills and knowledge on disaster resilience among construction professionals. In *The Proceedings of the Construction, Building and Real Estate Research Conference of the Royal Institution of Chartered Surveyors (COBRA)*, Dauphine Université, Paris.

Then, F.S., Luppa, M., Schroeter, M.L., Konig, H.H., Angermeyer, M.C., & Riedel-Heller, S.G. (2013). Enriched environment at work and the incidence of Dementia: results of the Leipzig longitudinal study of the Aged (LEILA 75+). *PLoS One*, *8*(7). https://doi.org/10.1371/journal.pone.0070906

Then, F.S., Luck, T., Luppa, M., Thinschmidt, M., Deckert, S., Nieuwenhuijsen, K., Seidler, A., & Riedel-Heller, S.G. (2014). Systematic review of the effect of the psychosocial working environment on cognition and dementia. *Occupational and Environmental Medicine*, *71*(5), 358-365. https://doi.org/10.1136/oemed-2013-101760

Then, F.S., Luck, T., Heser, K., Ernst, A., Posselt, T., Wiese, B., Mamone, S., Brettschneider, C., Konig, H.-H., Weyerer, S., Werle, J., Mosch, E., Bickel, H., Fuchs, A., Pentzek, M., Maier, W., Scherer, M., Wagner, M., & Riedel-Heller, S.G. (2017). Which types of mental work demands may be associated with reduced risk of dementia?. *Alzheimer's & Dementia*, *13*(4), 431-440. https://doi.org/10.1016/j.jalz.2016.08.008

Trading Economics (2019). *Germany GDP from construction. Summary.* Trading Economics Retrieved February 11, 2019, from https://tradingeconomics.com/germany/gdp-from-construction

Visscher, H. & Meijer, F. (2007). Dynamics of building regulations in Europe. *International Conference on Sustainable Urban Areas*, ENHR, 25-28.

Wahlström, J., Hagberg, M., Johnson, P., Svensson, J., & Rempel, D. (2002).

Influence of time pressure and verbal provocation on physiological and psychological reactions during work with a computer mouse. *European Journal of Applied Physiology*, *87*(3), 257-263. https://doi.org/10.1007/s00421-002-0611-7

Winwood, P.C. & Winefield, A.H. (2004). Comparing two measures of burnout among dentists in Australia. *International Journal of Stress Management*, *11*(3), 282-289. https://doi.org/10.1037/1072-5245.11.3.282

Wofford, J.C. (2001). Cognitive–affective stress response: effects of individual stress propensity on physiological and psychological indicators of Strain. *Psychological Reports*, *88*(3), 768-784. https://doi.org/10.2466/pr0.2001.88.3.768

Yang, L., Lu, K., Diaz-Olivares, J.A., Seoane, F., Lindecrantz, K., Forsman, M., Abtahi, F., & Eklund, J.A. (2018). Towards smart work clothing for automatic risk assessment of physical workload. *IEEE Access*, *6*, 40059-40072. https://doi.org/10.1109/ACCESS.2018.2855719

References (Study 3)

Albus, P., Vogt, A., & Seufert, T. (2021). Signaling in virtual reality influences learning outcome and cognitive load. *Computers & Education*, *166*, 104154. https://doi.org/10.1016/j.compe du.2021.104154

Armougum, A., Orriols, E., Gaston-Bellegarde, A., Joie-La Marle, C., & Piolino, P. (2019). Virtual reality: A new method to investigate cognitive load during navigation. *Journal of Environmental Psychology*, *65*, 101338. https://doi.org/10.1016/j.jenvp.2019.101338

Autorengruppe Bildungsberichterstattung. (2018). *Bildung in Deutschland 2018: Ein indikatorengestützter Bericht mit einer Analyse zu Wirkungen und Erträgen von Bildung* [Education in Germany 2018: An indication-based report analyzing the effects and benefits of education]. wbv.

Bacca, J., Baldiris, S., Fabregat, R., & Graf, S. (2015). Mobile augmented reality in vocational education and training. *Procedia Computer Science*, *75*, 49–58. https://doi.org/10.1016/j.procs.2015.12.203

Baethge, M., Solga, H., & Wieck, M. (2007). Übergänge aus allgemein bildenden

Schulen in die Berufsbildung. In *Berufsbildung im Umbruch. Signale eines überfälligen Aufbruchs* [Vocational education in Upheaval. Signals of an overdue awakening] (pp. 37–58). Friedrich-Ebert-Stiftung.

Bhoir, S. A., Hasanzadeh, S., Esmaeili, B., Dodd, M. D., & Fardhosseini, M. S. (2015, June). Measuring construction workers' attention using eye-tracking technology. *Proceedings of the Canadian Society for Civil Engineering 5th Int./11th Construction Specialty Conference*, Canada, *222*, 1–10.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008). Gaze selection in complex social scenes. *Visual Cognition*, *16*(2–3), 341–355. https://doi.org/10.1080/13506 28070 1434532

Chen, Y. F., Luo, Y. Z., Fang, X., & Shieh, C. J. (2018). Effects of the application of computer multimedia teaching to automobile vocational education on students' learning satisfaction and learning outcome. *EURASIA Journal of Mathematics, Science and Technology Education*, *14*(7), 3293–3300. https://doi.org/10.29333/ejmste/91245

Cheng, Y., & Huang, R. (2012). Using virtual reality environment to improve joint attention associated with pervasive developmental disorder. *Research in Developmental Disabilities*, *33*(6), 2141-2152. https://doi.org/10.1016/j.ridd.2012.05.023

Dyer, E., Swartzlander, B. J., & Gugliucci, M. R. (2018). Using virtual reality in medical education to teach empathy. *Journal of the Medical Library Association*, *106*(4), 498–500. https://doi.org/10.5195/jmla.2018.518

Fox, J., Arena, D., & Bailenson, J. N. (2009). Virtual reality: A survival guide for the social scientist. *Journal of Media Psychology*, *21*(3), 95–113. https://doi.org/10.1027/1864-1105.21.3.95

Frederiksen, J. G., Sørensen, S. M. D., Konge, L., Svendsen, M. B. S., Nobel-Jørgensen, M., Bjerrum, F., & Andersen, S. A. W. (2020). Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: A randomized trial. *Surgical Endoscopy*, *34*(3), 1244–1252. https://doi.org/10.1007/s0046 4-019-06887-8

Freina, L., & Ott, M. (2015, April). A literature review on immersive virtual reality in education: State of the art and perspectives. *Proceedings of the International Scientific Conference of eLearning and Software for Education (eLSE)*, Romania, *1*(133), 133–141.

Grassini, S., & Laumann, K. (2020). Questionnaire measures and physiological correlates of presence: A systematic review. *Frontiers in Psychology*, *11*(349), 1–21. https://doi.org/10.3389/fpsyg.2020.00349

Hart, S. G. (2006, October). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Los Angeles, CA, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). North Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Huang, W. (2020). *Investigating the novelty effect in virtual reality on stem learning* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global database. (UMI No. 27959999).

Huang, W., Roscoe, R. D., Johnson-Glenberg, M. C., & Craig, S. D. (2021). Motivation, engagement, and performance across multiple virtual reality sessions and levels of immersion. *Journal of Computer Assisted Learning*, *37*(3), 745–758. https://doi.org/10.1111/jcal.12520

Jeelani, I., Han, K., & Albert, A. (2018). Automating and scaling personalized safety training using eye-tracking data. *Automation in Construction*, *93*, 63–77. https://doi.org/10.1016/j.autcon.2018.05.006

Jensen, L., & Konradsen, F. (2018). A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies*, *23*(4), 1515–1529. https://doi.org/10.1007/s1063-9-017-9676-0

Jeong, J. B., Lee, S., Ryu, I. W., Le, T. T., & Ryu, E. S. (2020, October). Towards viewport-dependent 6DoF 360 video tiled streaming for virtual reality systems. In *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, WA (pp. 3687–3695). https://doi.org/10.1145/3394171.3413712

Meppelink, C. S., & Bol, N. (2015). Exploring the role of health literacy on attention to and recall of text-illustrated health information: An eye-tracking study. *Computers in Human Behavior*, *48*, 87–93. https://doi.org/10.1016/j.chb.2015.01.027

Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J.

(2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. *Computers & Education*, *70*, 29–40. https://doi.org/10.1016/j.compe du.2013.07.033

Meyer, O. A., Omdahl, M. K., & Makransky, G. (2019). Investigating the effect of pre-training when learning through immersive virtual reality and video: A media and methods experiment. *Computers & Education*, *140*, 103603. https://doi.org/10.1016/j.compe du.2019.103603

Morales, T. M., Bang, E., & Andre, T. (2013). A one-year case study: Understanding the rich potential of project-based learning in a virtual reality class for high school students. *Journal of Science Education and Technology*, *22*(5), 791–806. https://doi.org/10.1007/s1095 6-012-9431-7

Moro, C., Birt, J., Stromberga, Z., Phelps, C., Clark, J., Glasziou, P., & Scott, A. M. (2021). Virtual and augmented reality enhancements to medical and science student physiology and anatomy test performance: A systematic review and meta-analysis. *Anatomical Sciences Education*, *14*(3), 368–376. https://doi.org/10.1002/ ase.2049

Nobrega, F. A., & Rozenfeld, C. C. D. F. (2019). Virtual reality in the teaching of FLE in a Brazilian public school. *Languages*, *4*(2), 36–49. https://doi.org/10.3390/languages4020036

Olsen, A. (2012). *The Tobii I-VT fixation filter* [White Paper]. Tobii® Technology. https://www.tobii pro.com/siteassets/tobii-pro/learn-and-support/analyze/how-do-we-classify-eye-movements/tobii -pro-i-vt-fixation-filter.pdf

Parong, J., & Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology*, *110*(6), 785–797. https://doi.org/10.1037/edu00 00241

Reichenberger, J., Pfaller, M., & Mühlberger, A. (2020). Gaze behavior in social fear conditioning: An eye-tracking study in virtual reality. *Frontiers in Psychology*, *11*, 35. https://doi.org/10.3389/fpsyg.2020.00035

Reinhard, R. T., Kleer, M., & Dreßler, K. (2019). The impact of individual simulator experiences on usability and driving behavior in a moving base driving simulator. *Transportation Research Part F: Traffic Psychology and Behaviour*, *61*, 131–140. https://doi.org/10.1016/j.trf.2018.01.004

Rubo, M., & Gamer, M. (2021). Stronger reactivity to social gaze in virtual reality

compared to a classical laboratory environment. *British Journal of Psychology*, *112*(1), 301–314. https://doi.org/10.1111/bjop.12453

Schubert, T., Friedmann, F., & Regenbrecht, H. (1999). Embodied presence in virtual environments. In: R. Paton & I. Neilson (Eds.), *Visual representations and interpretations* (pp. 269–278). Springer. https://doi.org/10.1007/978-1-4471-0563-3_30

Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments*, *10*(3), 266–281. https://doi.org/10.1162/105474601300343603

Slater, M. (2003). A note on presence terminology. *Presence Connect*, *3*(3), 1–5.

Slater, M. (2018). Immersion and the illusion of presence in virtual reality. *British Journal of Psychology*, *109*(3), 431–433. https://doi.org/10.1111/bjop.12305

Tobii® Technology. (2012). *Determining the Tobii I-VT fixation filter's default values method description and results discussion* [White Paper]. Author. https://www.tobii pro.com/siteassets/tobii-pro/learn-and-support/analyze/how-do-we-classify-eye-movements/determining-the-tobii-pro-i-vt-fixation-filters-default-values.pdf

Ulrich, F., Helms, N. H., Frandsen, U. P., & Rafn, A. V. (2021). Learning effectiveness of 360 video: Experiences from a controlled experiment in healthcare education. *Interactive Learning Environments*, *29*(1), 98–111. https://doi.org/10.1080/10494 820.2019.1579234

Violante, M. G., Vezzetti, E., & Piazzolla, P. (2019). Interactive virtual technologies in engineering education: Why not 360° videos? *International Journal on Interactive Design and Manufacturing*, *13*(2), 729–742. https://doi.org/10.1007/s1200 8-019-00553-y

Wallet, G., Sauzéon, H., Rodrigues, J., & N'Kaoua, B. (2009). Transfer of spatial knowledge from a virtual environment to reality: Impact of route complexity and subject's strategy on the exploration mode. *Journal of Virtual Reality and Broadcasting*, *6*(4), urn:nbn:de:0009-6-17577. https://doi.org/10.20385/1860-2037/6.2009.4

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In A. L. Yarbus (Ed.), *Eye movements and vision* (pp. 171–211). Springer. https://doi.org/10.1007/978-1-4899-5379-7

References (Study 4)

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International journal of industrial ergonomics*, *36*(7), 623-636. https://doi.org/10.1016/j.ergon.2006.04.002

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*(3), 257-262. https://doi.org/10.1016/j.cub.2004.01.029

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276-292. https://doi.org/10.1037/0033-2909.91.2.276

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (pp. 142–162). Cambridge University Press.

Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., & Montanari, R. (2011). Driver workload and eye blink duration. *Traffic Psychology and Behaviour*, *14*(3), 199-208. https://doi.org/10.1016/j.trf.2010.12.001

Berguer, R., Smith, W. D., & Chung, Y. H. (2001). Performing laparoscopic surgery is significantly more stressful for the surgeon than open surgery. *Surgical Endoscopy*, *15*, 1204-1207. https://doi.org/10.1007/s004640080030

Bertelson, P. (1998) Starting from the ventriloquist: The perception of multimodal events. In M. Sabourin, F. Craik, & M. Robert, M., (Eds.) *Advances in Psychological Science, Vol. 2. Biological and Cognitive Aspects* (pp. 419–439). Psychology Press/Erlbaum.

Brungart, D. S., Kruger, S. E., Kwiatkowski, T., Heil, T., & Cohen, J. (2019). The effect of walking on auditory localization, visual discrimination, and aurally aided visual search. *Human Factors*, *61*(6), 976-991. https://doi.org/10.1177/0018720819831092

Chen, S., & Epps, J. (2014). Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction*, *29*(4), 390-413. https://doi.org/10.1080/07370024.2014.892428

de Greef, T., Lafeber, H., van Oostendorp, H., & Lindenberg, J. (2009) Eye

movement as indicators of mental workload to trigger adaptive automation. In D. Schmorrow, I. Estabrooke, & M. Grootjen (Eds.), *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience. Lecture Notes in Computer Science* (pp. 219–228). Germany.

Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., Raubal, M., & 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* (pp. 1-13). https://doi.org/10.1145/3173574.3173856

Forster, S., & Lavie, N. (2011). Entirely irrelevant distractors can capture and captivate attention. *Psychonomic Bulletin & Review*, *18*, 1064-1070. https://doi.org/10.3758/s13423-011-0172-z

Grassini, S., Laumann, K., & Skogstad, M. R. (2020). The use of virtual reality alone does not promote training performance (but sense of presence does). *Frontiers in Psychology*, *11*, 1743. https://doi.org/10.3389/fpsyg.2020.01743

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology, 52*, 139-183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hekele, F., Spilski, J., Bender, S., & Lachmann, T. (2022). Remote vocational learning opportunities—A comparative eye-tracking investigation of educational 2D videos versus 360° videos for car mechanics. *British Journal of Educational Technology*, *53*(2), 248-268. https://doi.org/10.1111/bjet.13162

Hershman, R., Henik, A., & Cohen, N. (2018). A novel blink detection method based on pupillometry noise. *Behavior Research Methods*, *50*, 107-114. https://doi.org/10.3758/s13428-017-1008-1

Hidaka, S., & Ide, M. (2015). Sound can suppress visual perception. *Scientific Reports*, *5*(1), 10483. https://doi.org/10.1038/srep10483

Hoeg, E. R., Gerry, L. J., Thomsen, L., Nilsson, N. C., & Serafin, S. (2017). Binaural sound reduces reaction time in a virtual reality search task. In *2017 IEEE 3rd VR workshop on sonic interactions for virtual environments (SIVE), USA,* 1-4. https://doi.org/10.1109/SIVE.2017.7901610

Jackson, C. V. (1953). Visual factors in auditory localization. *Quarterly Journal of Experimental Psychology*, *5*(2), 52-65. https://doi.org/10.1080/17470215308416626

Kim, J., Sunil Kumar, Y., Yoo, J., & Kwon, S. (2018). Change of blink rate in viewing virtual reality with HMD. *Symmetry*, *10*(9), 400. https://doi.org/10.3390/sym10090400

Larsson, P., Väljamäe, A., Västfjäll, D., Tajadura-Jiménez, A., & Kleiner, M. (2010). Auditory-Induced Presence in Mixed Reality Environments and Related Technology. In E. Dubois, P. Gray, & L. Nigay (Eds), *The Engineering of Mixed Reality Systems* (pp. 143–163). Springer. https://doi.org/10.1007/978-1-84882-733-2_8

Lavie, N. (2010). Attention, distraction, and cognitive control under load. *Current Directions in Psychological Science*, *19*(3), 143-148. https://doi.org/10.1177/0963721410370295

Lavie, N., & Dalton, P. (2014). Load theory of attention and cognitive control. In A. C. Nobre & S. Kastner (Eds.), *The Oxford Handbook of Attention* (pp. 56–75). Oxford University Press.

Macdonald, J. S., & Lavie, N. (2011). Visual perceptual load induces inattentional deafness. *Attention, Perception, & Psychophysics*, *73*, 1780-1789. https://10.3758/s13414-011-0144-4

Malpica, S., Serrano, A., Gutierrez, D., & Masia, B. (2020). Auditory stimuli degrade visual performance in virtual reality. *Scientific Reports*, *10*, 12363. https://10.1038/s41598-020-69135-3

Marquart, G., Cabrall, C., & de Winter, J. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, *3*, 2854-2861. https://doi.org/10.1016/j.promfg.2015.07.783

Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, *1*(1), 1-16. https://doi.org/10.5334/joc.18

Mathur, A., Gehrmann, J., & Atchison, D. A. (2013). Pupil shape as viewed along the horizontal visual field. *Journal of Vision*, *13*(6), 3. https://doi.org/10.1167/13.6.3

Merat, N., Jamson, A. H., Lai, F. C., & Carsten, O. (2012). Highly automated

driving, secondary task performance, and driver state. *Human Factors*, *54*(5), 762-771. https://doi.org/10.1177/0018720812442087

Nordahl, R., & Nilsson, N.C. (2014). The sound of being there: Presence and interactive audio in immersive virtual reality. In K. Collins, B. Kapralos, H. Tessler (Eds.), *The Oxford Handbook of Interactive Audio* (pp. 213-233). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199797226.013.013

Olk, B., Dinu, A., Zielinski, D. J., & Kopper, R. (2018). Measuring visual search and distraction in immersive virtual reality. *Royal Society Open Science*, *5*(5), 172331. https://doi.org/10.1098/rsos.172331

Palinko, O. & Kun, A. (2012) Exploring the influence of light and cognitive load on pupil diameter in driving simulator studies. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA'12)*, USA, 1-7. https://doi.org/10.17077/drivingassessment.1416

Recarte, M. Á., Pérez, E., Conchillo, Á., & Nunes, L. M. (2008). Mental workload and visual impairment: Differences between pupil, blink, and subjective rating. *The Spanish Journal of Psychology*, *11*(2), 374-385. https://doi.org/10.1017/S1138741600004406

Ruediger, P., Spilski, J., Kartal, N., Gsuck, S., Beese, N.O., Schlittmeier, S.J., Lachmann, T., & Ebert, A. (2019). Cognitive indicators for acoustic source localization and presence in a vivid 3D scene. *Proceedings of the 23rd International Congress on Acoustics*, 8234-8241. Germany.

Said, S., Gozdzik, M., Roche, T. R., Braun, J., Rössler, J., Kaserer, A., Spahn, D.R., Nöthinger, C.B., & Tscholl, D. W. (2020). Validation of the raw national aeronautics and space administration task load index (NASA-TLX) questionnaire to assess perceived workload in patient monitoring tasks: Pooled analysis study using mixed models. *Journal of Medical Internet Research*, *22*(9), e19472. https://doi.org/10.2196/19472

Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments*, *10*(3), 266–281. https://doi.org/10.1162/105474601300343603

Serafin, S., Geronazzo, M., Erkut, C., Nilsson, N. C., & Nordahl, R. (2018). Sonic

interactions in virtual reality: State of the art, current challenges, and future directions. *IEEE Computer Graphics and Applications*, *38*(2), 31-43. https://doi.org/10.1109/MCG.2018.193142628

Slater, M., & Sanchez-Vives, M. V. (2014). Transcending the self in immersive virtual reality. *Computer*, *47*(7), 24-30. https://doi.org/10.1109/MC.2014.198

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, *3*, 74. https://doi.org/10.3389/frobt.2016.00074

Theeuwes, J. (1994). Stimulus-driven capture and attentional set: selective search for color and visual abrupt onsets. *Journal of Experimental Psychology: Human perception and performance*, *20*(4), 799-806. https://doi.org/10.1037/0096-1523.20.4.799

Veltman, J. A., & Gaillard, A. W. K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, *41*(5), 656-669. https://doi.org/10.1080/001401398186829

Zheng, B., Jiang, X., Tien, G., Meneghetti, A., Panton, O. N. M., & Atkins, M. S. (2012). Workload assessment of surgeons: correlation between NASA TLX and blinks. *Surgical Endoscopy*, *26*, 2746-2750. https://doi.org/10.1007/s00464-012-2268-6

Zheng, B., Jiang, X., & Atkins, M. S. (2015). Detection of changes in surgical difficulty: evidence from pupil responses. *Surgical Innovation*, *22*(6), 629-635. https://doi.org/10.1177/1553350615573582

Appendix

Appendix 1: Mean level of demands based on the F-JAS

	M	CI 95%	SD
Cognitive (21 scales)	3.75	3.57–3.94	0.53
Oral comprehension	2.97	2.59–3.36	1.12
Written comprehension	3.15	2.62–3.68	1.52
Oral expression	3.06	2.62–3.49	1.28
Written expression	2.91	2.51–3.01	1.14
Fluency of ideas	4.66	4.23–5.08	1.24
Originality	2.56	1.98–3.14	1.67
Memorization	4.83	4.31–5.35	1.50
Problem sensitivity	4.60	4.13–5.07	1.35
Mathematical reasoning	3.49	3.01–3.96	1.38
Number facility	2.94	2.35–3.54	1.73
Deductive reasoning	4.47	4.04–4.90	1.25
Inductive reasoning	3.60	3.00–4.19	1.74
Sorting information	3.63	3.03–4.23	1.75
Category flexibility	2.71	2.07–3.34	1.82
Speed of closure	4.00	3.46–4.54	1.54
Flexibility of closure	4.32	3.84–4.81	1.39
Spatial orientation	3.51	2.88–4.15	1.84
Visualization	4.77	4.16–5.38	1.78
Perceptual speed	4.03	3.52–4.54	1.48
Selective attention	4.60	4.12–5.08	1.39
Simultaneous information processing	4.00	3.39–4.61	1.76
Psychomotor (10 scales)	4.63	4.43–4.84	0.60
Control precision	5.18	4.79–5.56	1.09
Multiple limb coordination	5.34	5.04–5.64	0.87
Response orientation	4.31	3.85–4.78	1.35
Rate control	4.20	3.69–4.71	1.49
Reaction time	3.80	3.22–4.38	1.69
Arm-hand steadiness	5.29	4.91–5.66	1.10
Manual dexterity	5.57	5.33–5.81	0.69
Finger dexterity	4.86	4.42–5.29	1.26
Wrist-finger speed	3.83	3.25–4.40	1.67
Speed of limb movements	4.00	3.46–4.54	1.57
Physical (9 scales)	4.84	4.59–5.09	0.73
Static strength	5.03	4.61–5.45	1.22
Dynamic strength	4.83	4.35–5.31	1.40
Strength endurance	4.56	4.04–5.08	1.48
Trunk strength	5.23	4.79–5.67	1.29
Extent flexibility	5.06	4.71–5.40	0.99
Dynamic flexibility	3.26	2.64–3.89	1.79
Gross body coordination	5.34	4.95–5.73	1.14
Gross body equilibrium	5.57	5.21–5.94	1.07
Stamina	4.64	4.29–4.99	1.03
Sensory–perceptual (12 scales)	3.67	3.40–3.95	0.79
Near vision	5.09	4.72–5.45	1.07
Far vision	4.29	3.72–4.85	1.64
Color discrimination	4.51	3.93–5.09	1.69
Night vision	1.80	1.31–2.29	1.43
Peripheral vision	3.80	3.17–4.43	1.84
Depth perception	4.00	3.42–4.58	1.69
Glare sensitivity	3.20	2.56–3.84	1.86
Hearing sensitivity	2.85	2.22–3.49	1.83
Auditory attention	2.46	1.86–3.05	1.74
Sound localization	3.40	2.72–4.08	1.97

Appendix 1 (continued):

	M	CI 95%	SD
Speech recognition	3.83	3.21–4.44	1.77
Speech clarity	4.83	4.39–5.27	1.27
Social (21 scales)	4.42	4.15–4.69	0.79
Friendliness	5.91	5.49–6.33	1.22
Behavior flexibility	4.21	3.69–4.73	1.51
Coordinating	4.40	3.85–4.95	1.61
Reliability	5.94	5.57–6.32	1.07
Expressing opinion	4.60	4.06–5.14	1.58
Negotiating	3.20	2.49–3.91	2.07
Persuasion	3.43	2.77–4.09	1.93
Communicating with others	4.51	4.09–4.94	1.25
Social compliance	4.09	3.55–4.63	1.54
Social sensitivity	3.91	3.22–4.61	2.03
Emotional control	4.17	3.51–4.84	1.93
Self-confidence	4.80	4.35–5.25	1.30
Training others	3.57	2.96–4.19	1.79
Oral fact finding	3.21	2.56–3.85	1.86
Motivation	5.61	5.36–5.86	0.73
Openness	5.06	4.59–5.53	1.37
Assertiveness	5.66	5.26–6.06	1.16
Persistence	4.12	3.52–4.72	1.72
Resistance to premature judgment	4.18	3.78–4.57	1.14
Debating	4.14	3.53–4.76	1.78
Stress tolerance	4.10	3.53–4.68	1.64

Notes: CI, confidence interval; *M*, mean; SD, standard deviation

Table 1-2: Mean level of demands based on the dimensions of the Fleishman Job Analysis Survey (F-JAS) as obtained in interviews with 35 construction workers as part of study 2..

Appendix 2: Comparison of fixation patterns in an error search (Unpublished research, see chapter 3.4)

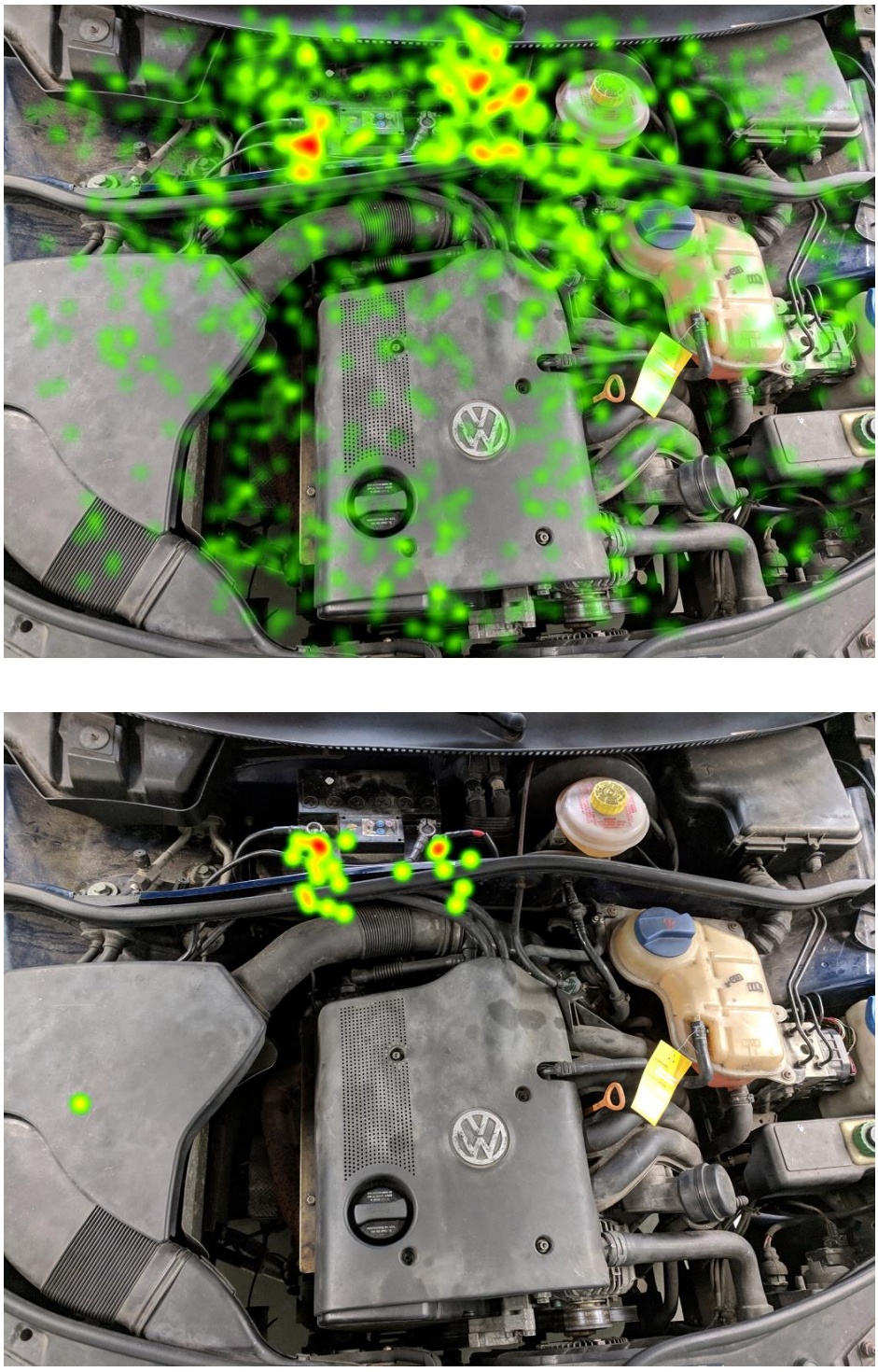

Figure 1: Heatmap visualization from visual search fixation patterns from 8 unlearned participants (top) and 1 expert with 15 years of experience (bottom). Exported from Tobii Pro Lab, using single frame snapshots. Applied Settings: Tobii I-VT Attention filter; Absolute Fixation Count; Radius 5px; Scale Max Value 5; Opacity 100%

Curriculum Vitae

Persönliche Angaben

Name: Felix Hekele

Anschrift: Erwin-Schrödinger-Straße, Gebäude 57-526, 67653 Kaiserslautern

E-Mail: felix.hekele@rptu.de

Beruflicher Werdegang

Februar 2018 – März 2024 Promotion Psychologie (Dr. rer. nat.)

Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau

Oktober 2016 – September 2017 Forschungsaufenthalt

Bangor University, Wales, Großbritannien

ERC Projekt “Social Robots” (Prof. Emily Cross)

Oktober 2013 – Januar 2016 Master-Studium Psychologie, Schwerpunkt Sozial- und Wirtschaftspsychologie

Paris-Lodron-Universität Salzburg, Österreich

Abschluss: Master of Science (mit Auszeichnung bestanden)

Thema der Abschlussarbeit: Bedrohung in sozialen Interaktionen

Abschlussnote: 1,22 (28.01.2016)

Oktober 2010 – Juli 2013 Bachelor-Studium Psychologie

Paris-Lodron-Universität Salzburg, Österreich

Abschluss: Bachelor of Science

Thema der Abschlussarbeit: Untersuchung von Reaktanz im fMRT

Abschlussnote: 1,86 (30.07.2013)