Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FACULTY
OF MATHEMATICS, INFORMATICS
AND NATURAL SCIENCES

DISSERTATION

Development and Evaluation of Immersive User Interfaces for Cognitive and Physical Training

an der Universität Hamburg eingereichte Dissertation

Lucie Marcella Miriam Kruse

Human-Computer Interaction
Fachbereich Informatik
Fakultät für Mathematik, Informatik
und Naturwissenschaften
Universität Hamburg

Hamburg, 2025

Betreuer & Erstgutachter:	Prof. Dr. Frank Steinicke
Zweitgutachter:	Prof. Dr. Robert Lindeman
Leitung der Prüfungskommission:	Prof. Dr. Janick Edinger
Stellvertretende Leitung der Prüfungskommission:	Prof. Dr. Eva Bittner
Datum der Disputation:	27.02.2025

Acknowledgements

I would like to thank everybody who helped me in the process of writing this dissertation.

First, Prof. Dr. Frank Steinicke, who gave me the freedom and support in researching a topic I am passionate about and that I believe can contribute in making the world a little bit better. Thank you for welcoming me into your group, first as a student assistant and then as a PhD student, for inviting me to be a part of the research community, and for your advice on research directions and future perspectives. I never thought that I would be a PhD candidate some day, but you made this possible.

Collaborating with the HCI group of Universität Hamburg has been a pleasure. Thank you to everyone who provided advice on user studies, on technologies, and everyone who proofread (part of) my thesis, or for just listening to my thoughts. Thanks for all the collaborative celebrations of our successes, and for providing encouragement to each other when things did not go as planned. We have achieved so much together, and I hope to continue this success in the future.

Further, I could always count on the immense help of the Hospital zum Heiligen Geist (HzHG), who assisted with our user studies: They offered guidance on how to design our studies, managed participant invitations, and provided a safe space for conducting our immersive experiences. Even after our joint project ended, they continued to support me and I am very grateful for this collaboration.

Of course, this work would not be possible without the participation of our end users: older adults. For everyone who participated in our focus groups, interviews, and user studies, please know that your help and insights were incredibly valuable. You contributed to advancing research in the domain of human-computer interaction, and I learned a lot from you.

I would like to thank my family and friends, who were always there for me during this PhD journey. You listened to me when I was discouraged, provided useful ideas for improving my work, or distracted me when I was stressing out about what could go wrong. Thanks to my parents, who provided me with a place to work without distraction, who always believed in me, and who let me take pictures of them using our applications for this dissertation. You are the best (role) models I could have ever hoped for. I also want to thank my aunt, who, early on, showed me the joy of sharing knowledge with others, and how developing our own teaching strategies can help others in their learning process. Finally, a huge thank you to my partner. Thanks for all of the above, and thank you for just being there.

Abstract

Healthy aging is important to live a fulfilling, independent life. This includes, among others, striving for a good psychological, cognitive, and physical well-being. In order to remain active as long as possible, it is important to continuously promote these aspects of health. One possibility to do this is through virtual reality (VR) exergames – applications that combine exercise and gaming to create a motivating approach, fostering long-term adherence. In order for them to show the desired effects, they need to be designed in a way that is understandable, operable, and motivating.

This dissertation investigated the effect of different VR exergames and design decisions on the user experience of older adults, aiming to answer the research question how to design immersive cognitive-physical training for older adults. In five user studies and one systematic literature review, different aspects of VR exergames were investigated. The first study was conducted with older adults with mild dementia, and evaluated an exergame over the course of nine weeks for its effectiveness in improving psychological, cognitive and physical well-being. In a second study, another exergame was compared to traditional video-based training with older adults without dementia, aiming to determine whether these two approaches are comparable. The third study examined the effects of different dynamic difficulty adjustments with younger and older adults, and identified transferabilities and differences between these two user groups.

Further, one aspect of VR exergames was investigated in more depth: The integration of intelligent virtual agents (IVAs). With a systematic literature review, the usage and preferences regarding IVAs in the healthcare domain were investigated. We then conducted a user study to compare the effect of agent visibility on task performance, and social presence. In a final user study, we integrated an IVA communicating with the help of GPT-4o into a VR exergame and evaluated the user experience, communication and preferences of older adults.

The results of this dissertation show that VR exergames can improve psychological, cognitive, and physical well-being, and that they are comparable to video-based exercises. Further, it shows that in the design of VR applications for older adults, a multitude of considerations have to be taken into account, which largely depend on the user group and the context of the application. This emphasizes the need to integrate the end users in the design and evaluation process and to continue research in this important area.

Zusammenfassung

Gesundes Altern ist wichtig, um ein erfülltes, unabhängiges Leben zu führen. Um so lange wie möglich aktiv zu bleiben, ist es wichtig, gesundheitliche Aspekte wie das psychologische, kognitive und körperliche Wohlbefinden kontinuierlich zu fördern. Eine Möglichkeit, dies zu tun, sind Exergames in der virtuellen Realität (VR): Anwendungen, die Bewegungen und Spiele kombinieren, um einen motivierenden Ansatz zu schaffen, der eine langfristige Aufrechterhaltung des Trainings fördert. Damit sie die gewünschten Effekte zeigen, müssen die Exergames so gestaltet sein, dass sie verständlich, bedienbar und motivierend sind.

Diese Dissertation untersucht den Effekt verschiedener VR Exergames und Designentscheidungen auf die Nutzungserfahrung älterer Erwachsener. Die Forschungsfrage, welche in fünf Nutzungsstudien und einem systematischen Literaturreview untersucht wurde, beschäftigt sich damit, wie man immersives kognitiv-physisches Training für ältere Erwachsene gestalten kann. Die erste Studie wurde mit älteren Erwachsenen mit leichter Demenz durchgeführt und bewertete ein Exergame über einen Zeitraum von neun Wochen hinsichtlich seiner Wirksamkeit zur Verbesserung des psychologischen, kognitiven und körperlichen Wohlbefindens. In einer zweiten Studie wurde ein Exergame mit traditionellem videobasiertem Training verglichen, um herauszufinden, ob diese beiden Ansätze vergleichbar sind. Die dritte Studie untersuchte die Auswirkungen verschiedener dynamischer Schwierigkeitsanpassungen von Exergames bei jungen und älteren Erwachsenen und identifizierte Übertragungsmöglichkeiten und Unterschiede zwischen diesen beiden Nutzer:innengruppen.

Darüber hinaus wurde ein Aspekt von VR Exergames näher untersucht: die Integration intelligenter virtueller Agenten (IVAs). In einem systematischen Literaturreview wurden die Nutzung und die Präferenzen bezüglich IVAs im Gesundheitsbereich untersucht. Anschließend wurde eine Nutzungsstudie durchgeführt, um den Effekt der Sichtbarkeit von Agenten auf die Leistung bei kognitiv-physischen Aufgaben und die wahrgenommene soziale Präsenz zu vergleichen. In einer abschließenden Nutzungsstudie wurde ein IVA, der mit Hilfe von GPT-4o kommuniziert, in ein VR Exergame integriert und die Nutzungserfahrung, die Kommunikation und die Präferenzen älterer Erwachsener wurde evaluiert.

Die Ergebnisse dieser Dissertation zeigen, dass VR Exergames das psychologische, kognitive und körperliche Wohlbefinden von älteren Erwachsenen verbessern können und dass sie mit videobasierten Übungen vergleichbar sind. Darüber hinaus zeigen sie, dass bei ihrer Gestaltung eine Vielzahl von Betrachtungen zu berücksichtigen ist, die weitgehend von der Nutzer:innengruppe und dem Anwendungskontext abhängen. Dies betont die Notwendigkeit, die Endnutzer:innen in den Design- und Evaluationsprozess miteinzubeziehen um dieses wichtige Thema weiter zu erforschen.

Contents

List of Figures

List of Tables

Part I

Background

Chapter 1

Introduction

1.1 Motivation

The global population is aging, with the World Health Organization (WHO) estimating that the number of people aged 60 or above will exceed 2 billion by 2050 [3]. This may pose financial and social challenges on the healthcare system, both in primary and long-term care. Investing in strategies to promote healthy aging can help to alleviate this burden, and improve the independence of older adults.

One important aspect influencing healthy aging is the quality of life (QoL). QoL encompasses several dimensions, such as emotional and physical well-being, or personal development and activity. A good QoL allows individuals to enjoy daily activities, engage in passions, and maintain social connections [90]. For older adults, a positive QoL plays a significant role in enhancing overall well-being, fostering greater autonomy and a better ability to adapt to challenges in life [342].

While we do not have control over all aspects of life, there are areas where we can make meaningful improvements. For example, research has shown that physical and cognitive abilities [130], as well as psychological well-being [45] can be improved and maintained through regular exercise. In particular, the combination of physical and cognitive exercises has the potential to reduce neurological degeneration in conditions like dementia [324]. Therefore, working towards the improvement of these three core areas – psychological, cognitive, and physical well-being – through cognitive-physical training is the goal of this thesis.

Traditional training methods often require qualified trainers, suitable training plans, and they can quickly become repetitive. In contrast, serious games or exergames have a motivating nature that encourages users to train their abilities on a regular basis [82, 270, 341]. These exercise programs combine exercise with gaming, and can be adjusted to the needs and wishes of the users. For example, it is possible to create individualized virtual environments (VEs), referring to the past of the users and encouraging reminiscence activities [197]. Further, exergames can be tailored to the skills of the users, featuring easier and harder tasks, and adjusting themselves dynamically [139]. With the help of reward systems and competitive social activities, long-term adherence can be encouraged [341].

In this context, virtual reality (VR) shows enormous potential. In VR, realistic virtual worlds can be created and these can elicit the feeling of presence – the illusion of being in a different place [309]. Additionally, VR often features natural movement tracking in six-degrees of freedom, which makes intuitive and novel interactions possible. Compared to 2D displays, VR has the advantages of improved immersion, flow and motivation [372].

While VR technology has enormous potential in this context, it can also introduce challenges and usability barriers. For older adults, the usage of VR can be challenging due to age-related health factors and lack of experience with innovative technology [295]. Further, ergonomic problems with the head-mounted display (HMD) [127, 133, 150], complicated interaction methods [133] or abstract interface concepts [308, 337] can deteriorate the experience. These difficulties in operating or understanding the application can impact their training experience, preventing them from improving their cognitive-physical fitness, and limiting their adherence [382]. Hence, it remains an important research question how to design inclusive, intuitive VR exergames that show a high usability, a good user experience, and which can therefore lead to promising training effects for older adults.

For VR exergame, there are various design decisions that can be considered. This ranges from the choice of environmental setting, to choosing suitable, meaningful content, to picking intuitive interaction methods. Aside from the goal of creating motivating experiences, the well-being and safety of the end users should always be in focus [150], and the applications should follow the basic principles of accessibility, being perceivable, operable, understandable and robust [50]. Therefore, it is crucial to integrate the end users in the design process, e.g., through employing a User-Centered Design (UCD) approach [121]. Through this method, the end users are consulted for their opinion and experience, and usability tests on designed prototypes identify barriers and possibilities for further improvement.

In this thesis, we investigated the impact of different immersive training interfaces and different game adjustments on the user experience of older adults, aiming to develop systems that motivate the users to be cognitively and physically active. Further, we aimed to improve their independence through the use of intuitive user interfaces and interactive intelligent virtual agents (IVAs) guided by artificial intelligence (AI). The general research question investigated in this thesis is:

How should we design immersive cognitive-physical training applications, especially for older adults?

1.2 Structure

This thesis is divided into four parts. Part one explains the motivation and background for this thesis. Part two presents three works on VR exergames for older adults, while part three investigates the integration of IVAs into cognitive-physical training applications. Finally, part four discusses the findings of this dissertation and provides future areas of research. The four parts are divided into different chapters. In detail, the following chapters are presented in this dissertation:

After this introductory chapter, Chapter 2 provides an overview on the scope of (healthy) aging, as well as cognitive-physical training in immersive environments, inclusive design, and the relevant literature surrounding it. Here, barriers and design recommendations are pointed out, which are considered in the development of all immersive training applications in the following chapters.

Chapter 3 explains the research methods used in this thesis, ranging from the UCD approach to interviews and standardized questionnaires. In a mixed methods approach, quantitative and qualitative methods were used to gain a comprehensive understanding of the preferences, needs and opinions of the study participants. Furthermore, it is highlighted in which chapters these methods were used.

Chapter 4 to Chapter 8 present the development and evaluation of the research created within the EXGAVINE project, funded by the 'Federal Ministry of Education and Research' (BMBF) [40]. This project resulted in the creation and evaluation of three immersive movement games as interactive concepts for people with neurological diseases: (i) *Memory Journalist*, (ii) *Maestro Game* and (iii) the *Canoe Game*. For each exergame, a study was performed with older adults. In a reflective discussion, implications, design considerations, and future research directions concerning the design of immersive training applications for different user groups are pointed out.

Chapter 9 to Chapter 13 focus on the implications derived from the previous part in more depth - in particular the question how to provide a better assistance during immersive training to improve the independence of older adults in using the exergames. Our proposed solution for this are intelligent virtual agents (IVAs). IVAs are computer-generated digital entities that can be animated to act like humans; e.g., by using natural speech, or by exhibiting social behavior [70]. These agents can provide assistance to VR users, e.g., through providing hints, drawing the user's attention to important game elements, correcting their movements or motivating them. In the first part, a systematic literature review is provided on agents in the healthcare sector. Consequently, a study is presented that compares the effect of different visibility settings of IVAs during cognitive-physical training on performance, the feeling of social presence and eye behavior. Finally, the results of a user study with older adults on using an IVA as an assistant in the aforementioned *Canoe Game* are discussed, reflecting on how assistive IVAs are perceived and used by older adults. This last study integrates the insights on immersive interface design collected from prior studies, and evaluates the current topic of generative AI with an underrepresented user group, providing insights on older adults' usage, need and opinion on AI.

In Chapter 14 to Chapter 17, all conducted studies are discussed and put into context. Here, we reflect on the general research question of this thesis, and whether the evaluated interface concepts resulted in a good user experience. Furthermore, design recommendations and future research directions are presented, emphasizing the need for more research on this fast evolving and important topic.

1.3 Scientific Publications and Authorship Contribution

This dissertation is based on several publications that have been published in peer-reviewed national and international venues. The original publications were partly rephrased and complemented with additional information. Furthermore, some figures were edited or added. Additionally, I published papers that are outside of the scope of this dissertation, which will also be listed in this section. Finally, I co-authored several publications, which might be cited within the dissertation, but are also not part of its main scope.

1.3.1 Publications Related to This Dissertation

For these publications, I contributed significantly as the main author in the design of the studies and the development of the applications, the collection and analysis of data, the writing of the paper and the management of the publication process. The original publications included in this thesis are the following:

(1) **Kruse, L.**, Karaosmanoglu, S., Rings, S., Ellinger, B., Apken, D., Mangana, T. F., and Steinicke, F. (2021). A Long-Term User Study of an Immersive Exergame for Older Adults with Mild Dementia during the COVID-19 Pandemic. In Proceedings of the International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments (pp. 9-18). DOI: `https://doi.org/10.2312/egve.20211322` [167]

(2) **Kruse, L.**, Karaosmanoglu, S., Rings, S., Ellinger, B., and Steinicke, F. (2021). Enabling immersive exercise activities for older adults: A comparison of virtual reality exergames and traditional video exercises. Societies, 11(4), 134. DOI: `https://doi.org/10.3390/soc11040134` [168]

(3) **Kruse, L.**, Karaosmanoglu, S., Rings, S., and Steinicke, F. (2022). Evaluating difficulty adjustments in a VR exergame for younger and older adults: Transferabilities and differences. In Proceedings of the ACM Symposium on Spatial User Interaction (pp. 1-11). DOI: `https://doi.org/10.1145/3565970.3567684` [169]

(4) **Kruse, L.**, Hertel, J., Mostajeran, F., Schmidt, S., and Steinicke, F. (2023). Would You Go to a Virtual Doctor? A Systematic Literature Review on User Preferences for Embodied Virtual Agents in Healthcare. In IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (pp. 672-682). DOI: `https://doi.org/10.1109/ISMAR59233.2023.00082` [165]

(5) **Kruse, L.**, Mostajeran, F., and Steinicke, F. (2023). The Influence of Virtual Agent Visibility in Virtual Reality Cognitive Training. In Proceedings of the ACM Symposium on Spatial User Interaction (pp. 1-9). DOI: `https://doi.org/10.1145/3607822.3614526` [174]

(6) **Kruse, L.**, Mostajeran, F., and Steinicke, F. (2023). High Levels of Visibility of Virtual Agents Increase the Social Presence of Users. In IEEE Conference on Virtual

Reality and 3D User Interfaces Abstracts and Workshops (VRW) (pp. 843-844). DOI: `https://doi.org/10.1109/VRW58643.2023.00264` (Poster) [173]

(7) **Kruse, L.**, Rings, S., and Steinicke, F. (2025). My Focus was on the Game: Investigating the Use of AI Assistants in a Virtual Reality Exergame for Older Adults. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA). DOI: `https://doi.org/10.1145/3706599.3719808` [175]

1.3.2 Additional First-Author Publications

Additionally, I am the main author of the following publications, which are outside of the scope of this thesis:

Kruse, L., Langbehn, E., and Steinicke, F. (2018). I can see on my feet while walking: Sensitivity to translation gains with visible feet. In IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (pp. 305-312). [171]

Kruse, L., Jung, S., Li, R. C., and Lindeman, R. W. (2020). On the Use of Jumping Gestures for Immersive Teleportation in VR. In Proceedings of the International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments (pp. 113-120). [166]

Kruse, L., Langbehn, E., and Steinicke, F. (2021). Where are you? Influence of Redirected Walking on Audio-Visual Position Estimation of Co-Located Users. In IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) (pp. 550-551). (Poster) [172]

Kruse, L., Wittig, J., Finnern, S., Gundlach, M., Iserlohe, N., Ariza, O., and Steinicke, F. (2023). Blended collaboration: Communication and cooperation between two users across the reality-virtuality continuum. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-8). [177]

Kruse, L., Rudschies, C., Rings, S., and Steinicke, F. (2024) Considering Psychotherapy in the Metaverse. In: EMRN A Metaverse for the Good Conference (pp. 119–121). [176]

Kruse, L., Knaack, L., and Steinicke, F. (2024). Back to (Virtual) Reality: Preferences and Effects of Entry and Exit Transitions to Virtual Experiences for Older Adults. In Proceedings of the ACM Symposium on Spatial User Interaction (pp. 1-2). (Poster) [170]

1.3.3 Co-Authored Publications

Finally, I contributed in several research papers as a co-author, which are also outside of the scope of this dissertation. Here, I helped with the design of these studies, the development of the applications, the collection and analysis of data, the writing of the paper, or with supervision of students' work:

Rings, S., Karaosmanoglu, S., **Kruse, L.**, Apken, D., Picker, T., and Steinicke, F. (2020). Using exergames to train patients with dementia to accomplish daily routines. In Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY) (pp. 345-349). [270]

Hartfill, J., Gabel, J., **Kruse, L.**, Schmidt, S., Riebandt, K., Kühn, S., and Steinicke, F. (2021). Analysis of detection thresholds for hand redirection during mid-air interactions in virtual reality. In Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology (VRST) (pp. 1-10). [114]

Karaosmanoglu, S., Rings, S., **Kruse, L.**, Stein, C., and Steinicke, F. (2021). Lessons learned from a human-centered design of an immersive exergame for people with dementia. Proceedings of the ACM on Human-Computer Interaction (CHI PLAY) (pp. 1-27). [150]

Hertel, J., Gabel, J., **Kruse, L.**, Wollborn, M., and Steinicke, F. (2022). Co-Design of an Augmented Reality Maintenance Tool for Gas Pressure Regulation Stations. In IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) (pp. 720-724). [123]

Karaosmanoglu, S., **Kruse, L.**, Rings, S., and Steinicke, F. (2022). Canoe VR: An immersive exergame to support cognitive and physical exercises of older adults. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (pp. 1-7). [149]

Schmelter, T., **Kruse, L.**, Karaosmanoglu, S., Rings, S., Steinicke, F., and Hildebrand, K. (2023). Towards More Inclusive and Accessible Virtual Reality: Conducting Large-scale Studies in the Wild. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (pp. 1-5). [290]

Lehmann, N., **Kruse, L.**, and Steinicke, F. (2023). Assistance in Virtual Reality Exergames: Preference for Species of Agents in Relation to Personality of Users. In Proceedings of Mensch und Computer (pp. 422-426). (Poster) [187]

Oberfrank, L., **Kruse, L.**, and Steinicke, F. (2023). Sphere Saber: A Virtual Reality Exergame to Study Age-Related Differences in Selective Visual Attention and Information Processing. In Proceedings of the ACM Symposium on Spatial User Interaction (pp. 1-10). [239]

Rudschies, C., Rings, S., **Kruse, L.**, Schauenburg, G., Marmarshahi, H., and Zimmer, C. N. (2023). Psychotherapy with the help of ChatGPT? Current Technical and Ethical Boundaries of Intelligent Virtual Agents. In Mensch und Computer - Workshopband. [277]

Rings, S., **Kruse, L.**, Grabowski, M., and Steinicke, F. (2024). Meeting Intelligent Agents: How Virtually Anything is Possible with an Intelligent Virtual Agent at the Table. In EMRN A Metaverse for the Good Conference (pp. 32-34). [271]

Ertugrul, N., **Kruse, L.**, and Steinicke, F. (2024). Make Me Run: The Effects of an Immersive Learning Experience on Physical Running Exercise. In GI VR/AR Workshop (pp. 10-18420). Gesellschaft für Informatik e.V. [84]

Korkmaz, L., Schmidt, S., **Kruse, L.**, and Steinicke, F. (2024). I Talk - You Write. Exploring Asymmetric Text- and Voice-Based Communication Between Humans and Virtual Agents. In Proceedings of the ACM International Conference on Human-Agent Interaction. [163]

Chapter 2

Related Work

In this section, fundamental concepts related to the *Aging Process and Dementia* and *Quality of Life* are explained, followed by related work on *Immersive Training through Exergames*, and *Accessible Design of Immersive Experiences*.

2.1 Aging Process and Dementia

When aging, people collect a variety of different experiences and perspectives, which can provide them with valuable insights and knowledge. Throughout the natural aging process, the body also undergoes a variety of physiological changes. This can affect tissues, bones and organs [10], for example, declining muscle strength and endurance, lower flexibility and joint range of motion or a higher risk for diseases such as dementia [45]. Further, sensory perception and processing speed declines with age [231], resulting in slower responses to stimuli.

Cognitively, abilities can be divided into different domains, such as attention, memory, executive functions or language [231]. Executive functions decline with advancing age, including decision making, problem solving, or multitasking [231]. Through the impairment of cognitive abilities, activities of daily living can be impacted, the autonomy of older adults can be restricted and this can result in a lower quality of life [45]. Further, the working memory or procedural memory abilities decline with age. In contrast to that, some aspects of memory are relatively stable with age, for example the episodic memory which saves, among others, autobiographical memories, or procedural memories, such as remembering how to play an instrument [231].

Additionally, neurological conditions like dementia can significantly impact the life of older adults. Dementia is a degenerative illness, which causes the impairment of several abilities including memory, learning, and the ability to perform daily activities [246]. Worldwide, there are more than 55 million people living with dementia, and almost 10 million new cases each year [246]. Age is the strongest known risk factors for dementia, with most people with dementia being 65 years and older [246].

Dementia can arise from a variety of factors, all connected to damaged nerve cells and their connection to the brain [246]. It can be distinguished between primary dementia, which is usually irreversible and makes up to 90% of dementia causes, and secondary dementia, where another disease is the leading cause for the symptoms of dementia [101]. Additionally, several risk factors – including smoking, excessive alcohol consumption, physical inactivity, social isolation, and depression – can elevate the likelihood of developing dementia [246].

First, the most common form of primary dementia is Alzheimer's disease (AD), which causes plaque in the brain to damage cerebral cells. The causes of Alzheimer's disease are not fully understood yet, with genetic factors accounting for less than 2% [101]. AD is the most common form of dementia, with around 60-65% of cases [101]. Second, there is vascular dementia [101]. Impaired circulation of the blood flow can lead to the death of nerve cells in the brain. Additionally, other neurological symptoms, numbness or paralysis can occur. Risk factors can be high blood pressure, diabetes or, again, smoking. A third form of dementia is associated with Lewy bodies. Here, deposits of protein inside nerve cells, called Lewy bodies, affect the chemicals in the brain [118].

Additionally, secondary forms of dementia exist, where another disease, potentially treatable, can cause symptoms that are similar to dementia. For example, metabolic diseases, excessive alcohol consumption or certain medication as well as deficiencies in essential vitamins can contribute to this [101].

In the context of dementia, mild cognitive impairment (MCI) is also often mentioned. Here, older adults also have impaired memory abilities, but it is not as severe as dementia [4]. Nevertheless, they are at a higher risk for developing dementia than fit older adults their age: It is estimated that around 20% of people with MCI will develop dementia within one year [4], which emphasizes the importance of an early intervention.

As there is currently no cure for the primary forms of dementia, the goal of treatment is to improve the QoL of the older adults [101]. This can be done by focusing on training the remaining skills of the older adults, strengthening their self-esteem and enabling them to perform activities of daily living. Treatments often work with movements or engaging the different senses of the older adults [102]. As an example, music, art or biography work can be integrated into an individual therapy which is usually adjusted to the older adult's skills, preferences and needs. Here, it is important to keep the performance pressure as low as possible to not further overwhelm the older adults [102]. Transferred to digital interventions, the applications should be designed as intuitively, understandable and operable as possible to enable the best experience, as learning new skills can be difficult due to the memory impairments of the older adults.

2.2 Cognitive and Physical Training

While there are many factors influencing QoL, this thesis focuses on two aspects that have the potential to improve psychological, cognitive and physical well-being: (i) *Cognitive Training* and (ii) *Physical Training*, and the mix of these two methods.

2.2.1 Cognitive Training

In cognitive training, different abilities can be addressed, such as memory, visuospatial abilities, or processing speed. In a between-subject design study with older adults, participants in a study by Willis et al. [363] did a ten-session training for memory, reasoning or processing speed, as well as training boosters around one and three years after the original study. The authors found that those abilities that were originally trained improved, and an effect could be observed even after five years, showing the long-term effect of cognitive training. In a similar study setup, Ball et al. [17] found that 87% of people in the processing speed group, 74% of the people in the reasoning group, and 26% of the people in the memory-training group showed a reliable cognitive improvement after the original intervention period, and booster training enhanced these training gains. It has also been shown that cognitive training can improve the visuospatial abilities of older adults with and without MCI, their frontal lobe function, and their ability to process new information [202], indicating its effect for people with cognitive impairment. Kueider et al. conducted a systematic review on literature surrounding computer-based cognitive interventions for older adults [178]. They found positive effect sizes in the 38 articles they reviewed and concluded that the effects achieved through computer-based training are comparable or better to those of paper-based training.

2.2.2 Physical Training

Physical exercise training can increase fitness and physical function in fit older adults [116] and people with dementia [124]. Different methods for physical training exist and the selection of a suitable option depends on the individual needs and preferences. For example, physical training for older adults can be provided through physiotherapists, in individual or group workout settings, or, alternatively, in home-based settings through videotapes [346], exercise programs [192], or technological applications [211]. The Otago exercise programme [42] is one of the most renowned programs to reduce falls in older age, and it has been shown that falls can be reduced by around 35% [274]. Falls can happen for several reasons, but two of the contributing factors are lack of strength and balance [42]. The Otago exercise programme implements several strength and balance exercises, which should be performed three times a week, and complimented with walking training at least twice a week [42]. Several adjusted versions were implemented and have shown their effectiveness [204], including an augmented reality implementation [379].

Physical exercises can also improve some aspects of cognitive well-being [49, 94, 109, 124], and it is an essential component in interventions aiming to reduce cognitive frailty [325]. Most physical tasks require cognitive input and attention [36], therefore, physical activities also have the potential to train cognitive abilities. Additionally, studies have indicated a decreased risk for, and a reduced progression of dementia when doing physical activities, which also highlights its importance for cognitive abilities [38, 275]. Finally, physical exercise can improve people's mood [45]. Different training methods were evaluated for their effectiveness in improving cognitive fitness, for example, resistance training, flexibility, and strength training, or cycling. Hall et al. have shown that aerobic training can improve executive functions such as working memory or decision-making [109]. Moreover, aerobic and strength exercises, as well as stretching, can lead to a better cognitive performance, as shown through a twelve-week program in a study by Langlois et al. [183]. In their meta-analysis, Colcombe et al. [49] explored the effect of exercise on cognitive processes such as speed, visuospatial abilities, controlled processing and executive control. In these four areas, they found the largest benefits of fitness training on executive-control processes, and that there are several moderating factors, such as training duration, length and type of the intervention.

2.2.3 Dual Cognitive-Physical Training

One way to simultaneously train cognitive and physical abilities is through dual-training [324]. These interventions combine a cognitive with a physical task, for example, combining a math or memory task with a strength or balance exercise [253]. Here, both, cognitive and physical abilities, can be trained either simultaneously or after each other. It has been shown that particular executive functions benefit more from simultaneous cognitive-physical training compared to only physical training [79]. Simultaneous training also includes switching attention between the cognitive and physical task, thereby enhancing the ability to shift attention [79].

2.3 Immersive Training through Exergames

While there are several methods to implement training options, this thesis focuses on a technological approach utilizing VR. VR has several advantages in this context, for example that training sessions are repeatable and adjustable, data can be recorded and provided to healthcare providers, and virtual experiences can be designed according to the needs and wishes of the users. To reflect on the benefits and possibilities of VR, this section will explain the basics about (i) *Virtual Reality Training* and (ii) *Serious Games*.

VR uses computer-generated images that are usually displayed on a stereoscopic HMD. Through the HMD, users can be immersed into a virtual world and interact with the world naturally, e.g., through physical movements in space. In an isometric mapping, their physical movements are directly transferred to their virtual representation, enabling natural interactions with the digital content such as locomotion or grabbing virtual objects. VR has many advantages in terms of user engagement, which are largely due to two phenomena: (i) immersion and (ii) presence. Immersion in this context refers to sensory aspects of the experience [309], e.g., the graphics frame rate, the tracking capacities, latency, the field of view, or the image quality shown on the display. A high immersion can be enabled through the technology that is used, e.g., a high visual quality of the images, low latency, and the addition of other sensorimotor stimuli like audio.

The second phenomenon is presence. This is often defined as the feeling or illusion of "*being there*", and is compared to visiting a place, rather than just perceiving images of it [310, 313]. The goal of VR technology is to have both high immersion and presence, in order to engage the users and deliver a realistic experience. Further, the feeling of presence can be called the *Place Illusion*, and additionally, there is the *Plausibility Illusion*, which refers to the illusion that a scenario is actually occurring [309]. Again, if these two illusions occur at the same time, this can lead to realistic behavior in immersive VR.

Aside from immersive VR using HMDs, different systems exist that can be counted as immersive media, but use different technology. For example, this includes systems like the MS Kinect [216] that uses motion tracking, but displays the virtual environment on a 2D screen, or projection based systems like the CAVE [54]. While these systems are not evaluated in this thesis, prior research on exergames and exercise training using them, especially with older adults, offers valuable insights into design considerations, which will be included in the following summary.

Exergames or serious games combine exercise training with gaming [241]. This way, adherence, enjoyment and motivation [390] for exercise training can be increased, and daily life skills can be trained in a playful way [82, 328]. In their review, Karaosmanoglu et al. [148] found that from 51 exergames that reported a target age range, 31 focused on older adults, highlighting the potential for this user group. The majority of exergames are designed for the purpose of physical training, rehabilitation, or prevention [148], and prior work often highlights the enjoyment of participants when using immersive technology for training activities [81, 82]. Several effects of exergames have been evaluated, for example, its potential to improve physical and cognitive fitness [137, 229], to increase the mood of participants [127], or to decrease depression [370].

2.3.1 Improvements in Mood

Direct effects of using (immersive) technology have been observed in a variety of studies. As shown in a study by Garcia et al., after using VR for relaxation and stimulation, participants with cognitive decline reported decreased stress [97]. Positive effects were also observed in the development of unique experiences through the co-creation of digital photo albums [58]. After experiencing an immersive concert in VR, a participant in a study be Hodge et al. showed a more positive mood compared to before the experience, which was visible in her body movements, indicating a change of mood [127]. While these short-term effects look promising, long-term studies are required to verify the positive effect of exergames on mood, and to distinguish it from a novelty effect.

2.3.2 Feasibility

The feasibility of exergames has been positively evaluated in several studies [71]. Gerling et al. evaluated an exergame using the Nintendo Wii Balance Board [99, 100]. In their setup, players shifted their body weight to carry out different tasks like avoiding falling objects on a 2D screen. While the authors identified barriers related to ergonomic issues in long-term playing, the older adults quickly engaged with the game content. Van Diest et al. evaluated an unsupervised exergame at home [69]. In order to control an ice-skater on a frozen canal, players had to perform leaning motions in front of a TV using an MS Kinect. Aside from confirming the feasibility of their exergame, the authors also found that the balance of some participants improved during the six week study.

2.3.3 Physical Training

One primary goal of exergames is to train the physical abilities of the users. In prior studies, this was often evaluated through self-assessments of the participants or through observations of the researchers. In a study by Unbehaun et al., a participant felt like the training led to an improvement in her whole body, and researchers also noticed improvements in balance and stability among their participants [339]. They evaluated four balance games with the MS Kinect where people with dementia performed lower and upper body movements based on the Otago exercise catalogue [42]. In their qualitative results, caregivers pointed out the potential of the exercises to improve gait, balance and mobility [339]. In a study by Eisapour et al., therapists also commented that their exergame led to a greater range of motion than what they expected from the participants [81]. In more formal evaluations, small-to-medium effects on physical fitness through the use of VR interventions were observed [158]. For people with cognitive frailty, positive effects on gait speed, hand grip strength and the Timed-Up-Go Test could be found [325].

2.3.4 Cognitive Training

Cognitive abilities have also been in focus of research in several papers, with a special focus on improving cognitive functions to promote greater independence and, consequently, an improved quality of life for older adults. It has been shown that long-term VR interventions can improve cognition and brain function in older adults with MCI [235, 334]. Additionally, after watching short films in VR, people with dementia showed more alertness [276]. In a 4-week study, Huang et al. [138] investigated the effect of immersive virtual environments (IVEs) on various cognitive outcomes in older adults. Their findings showed that exergame training in IVEs can improve performance in inhibition and task switching compared to a non-immersive version, and the feeling of presence can be a mediator of this outcome. Maeng et al. [202] found that cognitive VR training improved the visuospatial ability of older adults with and without MCI, their frontal lobe function and their ability to process new information. In a systematic review, four studies were identified that reported exergames to improve cognitive function of people with MCI [390]. Importantly, transfer effects in the cognitive domain of working memory could also be shown in a study by Eggenberger et al. [79]: the positive effects of a video game dancing intervention were transferred to an untrained cognitive domain, indicating the potential for exergames to have an effect even outside of their training scope.

2.3.5 Opportunities for Other Stakeholders

While there are several advantages for the primary users, exergames can also benefit secondary users such as caregivers or clinicians. One advantage is that older adults can experience exergames at home, during their preferred time, partially decreasing the need for supervision [82] or transportation organization [339]. Especially for people with dementia, the logistical challenge of transporting them to the place of exercise training adds additional workload upon caregivers. This might be reduced through home-based exergame solutions, and enable caregivers or relatives to regain leisure time [339]. Furthermore, technical solutions can aid clinicians and physiotherapists to measure the therapy outcome [284], e.g., by automatically tracking the movement or duration of the training, as well as changes in relevant measurements such as range of motion [80, 82] or cognitive performance.

2.4 Accessible Design of Immersive Experiences

In the context of VR, accessibility means to give all people equal opportunities to experience a virtual environment, independent of their abilities or circumstances [228]. From a usability point of view, it means to make applications usable by as many people as possible. While there is no ideal way to make an application accessible for every individual, developers should aim to create designs that are accessible and user-friendly for a diverse range of users [228]. In addition, by June 2025, legal regulations will be strengthened in Germany, requiring certain hardware, applications and services to be accessible [18]. In accessibility research, the Web Content Accessibility Guidelines (WCAG) have established four principles for accessibility [50]: Perceivable, Operable, Understandable and Robust, abbreviated with the acronym "*POUR*". These are the basic requirements that each application or website needs to fulfill to be accessible. First, the presented information needs to be perceivable, e.g., users need to be able to obtain the information with at least one of their senses. Second, users need to be able to operate the interface, e.g., to perform interactions to control the application. Third, the information as well as the operation of the user interface needs to be understandable to users. And lastly, the application needs to be robust, meaning that it must remain accessible as technology advances, and also support the use of assistive technology.

In the context of VR, a set of guidelines [119] has been arranged from traditional guidelines such as WCAG or accessibility requirements made by hardware manufacturers. These guidelines range from visual implications such as large interactive elements and easily readable fonts, to re-configurable input mappings and the option to experience VR without the need for sound and speech. Further, they inform researchers about the need to account for various body types and a diverse range of motion, to allow as many (gameplay) settings as possible to be adjustable, and they prompt developers to not make precise timings essential. The accessibility of VR applications for older adults has been assessed in several works. In the following section, identified barriers as well as design recommendations are summarized.

2.4.1 Barriers

Aside from traditional barriers that affect all users, like usability and cybersickness problems [47], which can lead to frustration of the participants [337], traditional exergames, and even exergames or applications designed specifically for older adults, can pose several age- and health-related barriers that hinder participants from experiencing immersive applications and its positive effects for cognitive, physical and psychological well-being in their entirety.

Physical Barriers

Some barriers are related to physical abilities of the participants. For example, people in wheelchairs are often not able to participate in immersive experiences, for example, if full-body motions are required [308]. For others, excessive physical motion when using a Kinect system can lead to fatigue [308]. However, not only large motions can be problematic. For example, people with arthrosis or arthritis can have limited mobility in their hands, and limited fine motor skills, making it harder and painful to press small buttons on a controller with the needed accuracy [356], and sometimes buttons can be pressed accidentally [133]. Fast game mechanics can also pose a challenge to some older adults [133], and some participants in a user study were worried about mobility impairments, e.g., if a game is designed to be played with two arms, but only one arm is mobile [133]. The weight of a VR HMD is also often pointed out as a barrier [127, 133, 150], in combination with the fear of looking "*silly*" when wearing the HMD [127]. Designers and developers should aim to reduce these barriers to create *operable* interfaces.

Cognitive Barriers

Cognitively, some innovative concepts like avatars and different perspectives can be unfamiliar or too abstract for older participants to understand, especially without prior explanation [308, 337]. Additionally, talking about technology and devices can be a barrier that needs to be mediated [28], and might be due to having little experience with interacting with innovative technology [295]. This also applies to VR hardware, for example, when talking about controllers or specific button names like the "*trigger*" button, especially for non-native English speakers. For people with dementia, an impaired language processing skill can lead to difficulties in understanding verbal instructions [160], which can be further amplified when using unknown technical words. Difficulties with interactions on digital devices have also been observed in people with dementia, for example, performing simple tasks on a touch screen or using certain gestures [308], which might also be due to audio-visual capabilities. Cognitive impairments might also influence the beliefs of a person. For example, one user in a VR flower planting experience asked the study coordinators where the garden she just finished planting was located, assuming that it was a real-time visual representation of a physical garden [308]. Designing immersive applications in an intuitive way, and making sure that concepts are explained well, can help to create an *understandable* application.

Social Barriers

While social activities connected to the exergame experience have shown much potential, the social dynamics during those interactions can also lead to problems or negative perceptions. For example, in a group activity, participants in a study by Unbehaun et al. [337] were often critical of each other, exclaiming that an active player did not understand the task of the game. While training in a group setting with other people can be motivating, it can also be frustrating for bystanders and the players themselves for interpersonal matters. It has been reported that the pressure of performing well, but not understanding the needed interactions, can lead to associating negative experiences with using the system in a group [337].

On the other hand, some participants were frightened to train alone, and preferred to have other people in the room with them, emphasizing the need for external supervision [339]. While most caregivers are open to try out new technology, others have concerns about the extra workload that is added to their busy schedule [308], for example, by maintaining the devices [28], or by creating virtual content for the older adults, such as an interactive photo album [58]. Additionally, it is often the relatives or caregivers who take on the role of motivators [336], which can be exhausting in the long run. While family members can serve as an entry point into an experience, for example, if they try the application themselves or encourage their relatives to do so, their absence could potentially lead to reticence to try out unfamiliar systems [127]. In other studies, participants were only motivated to use a proposed system when the research team was present, or even refused to use the system without being guided by them [337]. This amplifies the role of social dynamics for training, but also the barriers of training alone without these social motivators.

Emotional Barriers

Finally, especially for people with dementia, their performance and motivation highly depends on their mood. For example, the visit of a family member, prior activities or medication input can affect the performance during exergame sessions [82]. In other studies, when participants had a positive mood they could accept photos in an interactive photo album, while they would complain or get annoyed at inaccuracies if their mood was negative [58]. Moreover, negative memories can also be triggered from virtual experiences [308]. These considerations also need to be taken into account when designing the content of VR applications.

2.4.2 Design Recommendations

Prior research shows that a good usability is essential to design an effective training application. For example, the application needs to fit the users' physical abilities, otherwise it cannot be operated. Likewise, it needs to suit their cognitive abilities, otherwise the user is preoccupied with understanding the training program, and unable to concentrate on the actual cognitive tasks [382]. An emphasis should be put on designing meaningful, intuitive and safe applications that can be adapted to the needs and preferences of the users. While guidelines for good design practices exist, applications should always be evaluated with the end users and other stakeholders and their feedback should be valued and iteratively integrated to design an accessible application. In order to design a suitable game, several recommendations have been proposed in literature.

Integrating On-Boarding Strategies

As a suitable on-boarding strategy, it is recommended to describe the VR experiences beforehand and show the older adults pictures or videos of people wearing the HMDs, and of the experiences they are going to see [127]. A special emphasis should be on providing a sense of trust and safety to the user, e.g., by communicating with the older adults or by providing suitable training [133]. In some prior work, pre-training sessions for familiarization purposes were conducted [80] or recommended [71] to ensure the users' comfort and feasibility. This might also include suitable calibration of the experience, e.g., the range of motion or desired difficulty [80]. Additionally, it is important to inform the participants about the possibility of cybersickness, and suitable countermeasures [127], although applications designed for older adults generally show a low occurrence of cybersickness [71, 74] and the benefits often outweigh the risks [74], e.g., by the positive experience older adults have with XR.

Providing Physical Accommodations

In order to avoid the risk of falling, and to account for people sitting in wheelchairs, several XR exergames support only seated experiences [81, 127], or use or recommend both seated and standing options [148, 150, 284]. Further, it is recommended to limit the amount of required head and body movements to increase accessibility [133]. To enable all players to use the system without the need to learn complicated controls, it is recommended to design the interactions as intuitive as possible [150], which might include using custom-made devices, limiting the use of controller buttons [82] or designing simple touch interfaces.

Designing Helpful Instructions and Assistance

Goals and objectives of the experience should be clear and apparent without requiring significant cognitive or physical effort [80]. Moreover, different prompts are suggested to provide assistance, e.g., visual, auditory or haptic ones [30], and visual cues could be used to guide the users' attention to important information [71]. This assistance could further be tailored to the users' knowledge and experience, e.g., by providing more advanced tips [25]. For people with dementia, it is recommended to give detailed instructions and repeat them, if necessary [150].

Adjusting the Application to the Users

Adjusting the difficulty or content of training applications was also done in several works, e.g., in terms of exercise duration [284], distance and speed [80], or to account for mobility impairments allowing to exercise with either one arm, or with both [284]. Generally, it is recommended to adapt serious game challenges to users' abilities [25, 150] to ensure that all users can achieve a similar performance, independent of the players physical or cognitive abilities, and to ensure an enjoyable, challenging experience for all. Visually, it is recommended to design a simple and intuitive interface with easy navigable menus, large elements like text and buttons and clear, explicit instructions, as well as the option to customize these settings [133].

Providing Meaningful Application Content

In terms of content, it is often emphasized that the experiences should be relevant, familiar and meaningful to the user [80, 82, 128, 133, 150, 307]. For example, Hodge et al. [128] captured 360° videos and audio recordings on a day out with people with dementia and their families and then ran workshops with them, where these memories were shared, curated and relived. In a study by Hosseini et al. [133], caregivers emphasized the importance of designing VR games with relevant content and the personalization of these games. A rural farm environment was designed by Eisapour et al. [82] for their VR exergame for people with dementia. This aimed to enhance familiarity for the users. Everyday life sounds can also be beneficial for people with dementia to evoke memories and emotional responses [134] and to foster conversations about past experiences. Vibrant colors, immersive environments and high-speed dynamics were also mentioned as key factors influencing enjoyment [133].

Integrating Social Interactions and Rewards

Further, many researchers have recommended to integrate social interactions into cognitive-physical training opportunities [337], including a ranking list [337, 341], or a multi-player setup [25, 150]. Nevertheless, less than 10% of articles in a systematic review on reward mechanisms in exergames for older adults by Bilius et al. explored competitive or cooperative serious games [25]. As rewards, besides a ranking list or score system, different methods were used such as developeable avatars, animations and pictures, or feedback messages [25]. In general, it is recommended to use positive, adaptable and diverse rewards [25].

Chapter 3

Methods

For the research conducted within the scope of this thesis, a variety of different data acquisition and data evaluation methods were employed, ranging from quantitative methods such as usability surveys to qualitative methods such as interviews. The collected data was evaluated according to the goal of the respective user study, for example, through Frequentist or Bayesian statistics [152], or through qualitative methods like thematic analysis [32]. The following section explains the most relevant data collection and evaluation methods from HCI research that have been used for the research conducted in the scope of this thesis.

3.1 User-Centered Design

UCD directly includes the end users in the design process and will be explained according to Sharp et al. [121]. This is important to ensure that the end product has a high usability and that end users, who will perform activities with the product on a regular basis, really need and use it. UCD places the primary focus of the design process on the users and their goals, and not on the technology used to achieve this. It generally involves three principles: Early focus on the users and the task, empirical measurements, and iterative design [104]. First, the users' behavior and the context in which the application will be used, should be studied in order to determine priorities, preferences and intentions of the users [121]. Furthermore, the cognitive and physical aspects of the end users should be understood and acknowledged. Finally, users should be consulted to be part of the development process from the first idea to the final product. Empirical measurements like interviews or questionnaires can be used to assess their opinion and the usability and user experience of the application. With the knowledge gathered from such measurements, the prototype should be iteratively improved to ensure that the requirements of the end users are met.

Older adults should be involved in the whole development process, but traditional methods may need to be adapted to suit their needs [77]. Some older adults might have limited technological skills, preventing them to contribute new ideas [77]. Furthermore, the user group of 'older' adults is heterogeneous in itself, with adults in their 60's having different technological knowledge, experience or cognitive-physical fitness than older adults in their 90's [77].

UCD was used throughout the work on this thesis, with different involvement of the users at different stages, e.g., in feasibility testing and usability studies. Whenever possible, we tried to include the end users directly, while sometimes we needed to adapt the data collection process. The corresponding adaptions taken to ensure accessibility of the research methods are explained in the corresponding subsections.

3.2 Interviews

In UCD, interviews are conversations about certain topics, which are usually conducted live. According to Sharp et al. [121], there are three main kinds of interviews, and they differ according to the amount of control that the interviewer has on the conversation: unstructured, semi-structured, and structured interviews. A suitable choice of the correct method depends on the goal of the study, as well as the research questions that are addressed.

- Unstructured interviews are exploratory, without steering the interview into a certain direction. In these interviews, open questions are often used, without an expectation of the answer format. Advantages are a rich data generation and the fact that topics that the interviewer has not considered could be mentioned by the interviewee. Nevertheless, these interviews generate a lot of unstructured data which in turn needs to be analyzed in depth.
- In a structured interview, questions often have predetermined sets of answers. This type of interview should be chosen only if clear goals of the interview can be formulated, for example, to find out how often a certain technology was used before, or which of two alternatives is preferred. Here, each participant receives the same questions in the same order, similar to a survey. The advantage of this format is that it is time-saving and questions are often easier to answer if alternatives are given. Nevertheless, important in-depth information on topics that were not considered when designing the interview might be missed.
- Semi-structured interviews combine the two other methods, using both open and closed questions. The interviewer has a predetermined script of questions, so all interviews cover the same broad topics, but can ask follow-up questions to deepen their understanding of a certain topic. This way, the advantages of both prior interview types can be combined. Nevertheless, a bias can be caused due to the interviewer's phrasing of questions and subsequent probing, which can potentially guide the interviewee into agreeing to what the interviewer suggested.

When conducting interviews with older adults, cognitive and physical impairments need to be taken into account. For example, many older adults have auditory impairments, which might results in a lowered motivation to participate in interviews since conversations can take longer and it can lead to misunderstandings [361]. Therefore, Wenger et al. suggest that interviewers should speak clearly and slowly, and if possible, face the participant with their mouth being visible, since a visual representation of a mouth can help people to understand each other better [46]. Furthermore, cognitive impairments can impact the way older adults answer questions. For example, if an interview is conducted after experiencing the technology, it might be hard to remember all the details of the experience due to an affected short-term memory. Especially for first-time users, the technology itself can occupy cognitive resources and their inexperience can lead to general, non-specific feedback in interviews [67].

With people with dementia, sometimes proxies are interviewed [60, 361]. These are people who are close to the person with dementia, for example, family members or close caregivers. Conducting interviews with proxies instead of the older adults themselves can have advantages if users are struggling with fatigue or illness, but the participant should be interviewed whenever possible to not result in a bias or exclude user groups from voicing their own opinions.

For the research described in this thesis, semi-structured interviews were used. Specifically, these were employed for the research in Chapter 12. For Chapter 5, two semi-structured proxy interviews with caregivers were conducted.

3.3 Standard Questionnaires

Through standard questionnaires, a large amount of (quantitative) data can be collected and compared. These questionnaires are usually validated to ensure that they accurately measure the desired outcomes [288]. They often have an established evaluation method which helps in interpreting the results. Furthermore, benchmark or comparison datasets exist for several questionnaires, enabling interpretations of data beyond the scope of the collected data, e.g., for the User Experience Questionnaire [296].

While standard questionnaires pose many benefits, they can sometimes show unreliable results with people with dementia [103] due to short-term memory loss and limited reflection abilities. Again, special questionnaire for proxies exist, but similarly to proxy interviews, their opinion not necessarily reflects the view of the person with dementia, which should still be prioritized [60].

In the following, the standard questionnaires used for this work on this thesis in the area of *Virtual Reality and Usability Research*, *Social Presence and Anthropomorphism*, and *Dementia Research* are described.

3.3.1 User Experience Questionnaire

The user experience of an application can be measured with several questionnaires. One of the most established ones is the User Experience Questionnaire (UEQ) [184, 185]. This questionnaire assesses user experience on six subscales: Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty. There are several verified translations, and also a short version is available [297]. The questionnaire presents word pairs on a 7-point semantic differential scale, which is being counted from -3 to 3. A data analysis tool is provided on the authors' website, which automatically calculates the values for all subscales and enables the comparison of the collected data with a benchmark dataset [296]. This allows comparisons of the evaluated application to other products. The UEQ was used for Chapter 12.

3.3.2 Simulator Sickness Questionnaire

Simulator Sickness or Cybersickness can occur in VR if the user's senses are in conflict [186]. One example for this is if a user is sitting still in the physical world, while the virtual world around them is moving without them controlling it. This may lead to symptoms like dizziness, nausea or eye strain. The exact causes for cybersickness are still in debate, but there are several established theories to examine the underlying reasons why cybersickness can occur in VR. The most established theory is that cybersickness might occur if visual cues, somatosensory senses and the vestibular system are receiving contradictory information, e.g., if there is a sensory conflict [186]. For example, this can happen when artificial scene or user movement is induced in VR. Further, cybersickness can be intensified by technical limitations, such as rendering artifacts (flickering), tracking errors, lag and latency. Here, the movements of the user and the reaction of the systems to these movements are not synchronous, which again leads to a sensory conflict.

The Simulator Sickness Questionnaire (SSQ) [154] was originally developed to measure sickness in flight simulators, but has been widely adopted for VR research [26]. It lists 16 symptoms of simulator sickness, which can be categorized into three symptom clusters: (i) Nausea, (ii) Oculomotor and (iii) Disorientation [154]. These symptoms have to be rated on a scale from 0 (no perception of the symptom) to 3 (severe perception). These values are then multiplied by scaling factors, whose function it is to produce scales with similar variability. The SSQ is often administered before the experience, and immediately after [26]. Through the establishment of baseline measures, direct effects of the VR experience can be measured and potential individual symptoms or risk factors can be identified beforehand. While earlier research indicates that scores above 20 are considered as a bad simulator [321], newer research in VR is calling for more differentiated threshold values [26]. The SSQ was administered for the work described in Chapter 6 and Chapter 7.

3.3.3 NASA-Task Load Index

To measure the perceived workload of an activity, the NASA-Task Load Index (NASA-TLX) has been established [113]. It was developed to assess factors that are relevant to the subjective experience of workload. The questionnaire has six subscales, namely (i) Mental Demand, (ii) Physical Demand, (iii) Temporal Demand, (iv) Performance, (v) Effort, and (vi) Frustration. With these subscales, an overall score can be calculated. First, all questions are answered on a scale from 0 to 100, in steps of 5. This calculates the RAW scores of NASA-TLX, which are used in many research publications. In a second step, the subscales can be weighted by the participants according to their individual importance. This is done with pairwise comparisons – participants pick the dimension that contributed more to their perceived workload in the given task [112]. Then, the ratings are multiplied with the subscales and the result is divided by the number of comparisons, in this case, 15. Perceived workload using NASA-TLX was assessed in the studies presented in Chapter 6 and Chapter 7.

3.3.4 MEC Spatial Presence Questionnaire

One way to measure presence in virtual environments is through the MEC Spatial Presence Questionnaire (MEC-SPQ). It has seven subscales [352]: (i) Attention Allocation, (ii) Spatial Situation Model, (iii) Spatial Presence, (iv) Higher Cognitive Involvement, (v) Suspension of Disbelief), (vi) Domain Specific Interest and (vii) Visual Spatial Imagery. It is available in three versions: with 4 items per subscale, 6 items or 8 items. The questions should be answered on a 5-point Likert scale from 1 (I do not agree at all) to 5 (I fully agree). The questionnaire should be administered directly after a media exposure. The subscale Attention Allocation was used in a study that is presented in Chapter 11.

3.3.5 Intrinsic Motivation Inventory

Intrinsic motivation is a behavior that comes from intrinsic desires and is independent from external incentives [282]. Users can be motivated to perform activities because they find them inherently interesting or enjoyable, which can in turn lead to long-term persistence. A participant's subjective experience of activities can be measured using the Intrinsic Motivation Inventory (IMI) [280], or the short German version [362]. The questionnaire has six subscales: (i) Interest/Enjoyment, (ii) Perceived Competence, (iii) Effort, (iv) Usefulness, (v) Pressure/Tension, and (vi) Perceived Choice. The Interest/Enjoyment subscale is of special interest here, because is is considered the self-reported measurement of intrinsic motivation. The questionnaire uses a 5-point-Likert scale from 1 to 5. It is used in studies presented in Chapter 6, Chapter 7 and Chapter 12.

3.3.6 Social Presence Survey

Intelligent virtual agents can elicit similar social responses in users as real humans, for example, regarding perception and behavior in personal space [16]. In order to measure how much virtual agents are perceived as real humans, several scales can be used. Bailenson et al. [16] developed the Social Presence Survey (SPS), which consists of five questions, each rated on a Likert-scale from -3 to 3. The results are summarized in order to generate a general presence score. A positive score can be interpreted as the participants perceiving the virtual agent as conscious and aware, whereas a negative score indicates that the virtual agent is perceived as unconscious and unaware. The SPS was used for a study shown in Chapter 11.

3.3.7 Temple Presence Inventory

Similarly, the Temple Presence Inventory (TPI) was developed to measure different dimensions of (Tele-)presence, meaning in how far users feel like perceiving the remote presence of another actor [195]. It consists of 42 questions which are divided into the following subscales: (i) Spatial Presence, (ii) Actor Within Medium (Parasocial Interaction), (iii) Passive Interpersonal, (iv) Active Interpersonal, (v) Mental Immersion, (vi) Social Richness, (vii) Social Realism, and (viii) Perceptual Realism. Except for the Social Richness subscale, all scales use a 7-point Likert scale to rate the user's answer. The Social richness subscale uses word pairs with a 7-point semantic differential scale. The Actor Within Medium (Parasocial Interaction) subscale of the TPI was used in a study detailed in Chapter 11.

3.3.8 Godspeed Questionnaire

Originally developed to evaluate the perception of robots, the Godspeed Questionnaire (GQS) [19] can also be used to get a better picture on the users' perception of a virtual agent, e.g., the agent's behavior or social attributes. This questionnaire has 24 item pairs and is answered on a 5-point semantic differential scale from 1 to 5. It can be divided into five subscales, namely (i) Anthropomorphism, (ii) Animacy, (iii) Likeability, (iv) Perceived Intelligence and (v) Safety. The Godspeed questionnaire was used for Chapter 12.

3.3.9 Addenbrooke's Cognitive Examination III

Addenbrooke's Cognitive Examination III (ACE III) is a screening test for cognitive impairment, especially for the diagnosis of dementia [135, 206]. It has 26 questions and tasks on five subscales: (i) Attention, (ii) Memory, (iii) Language, (iv) Visual Perception, and (v) Visuospatial Skills. A total of 100 points can be reached, with cut-off values of 83 or 88 depending on the desired sensitivity [206]. The interpretation of the questionnaire depends on the points reached, with more points generally indicating less cognitive impairment, but sociodemographic information about the participant has to be taken into account as well [39]. For example, the years of education, age or IQ of a person can have an influence on the score. The ACE III test was employed in our user study on the long-term effects of a VR exergame, which is presented in Chapter 5.

3.3.10 Dementia Quality of Life Questionnaire

The Dementia Quality of Life Questionnaire (DEMQoL) is used to measure the health-related QoL of people with dementia [316]. Two versions exist: The DEMQoL with 28 items and one final question that is directly filled out by the participant, and the DEMQoL-Proxy with 31 items and one final question, which is filled out by a caregiver.

It is used with people with mild to moderate dementia. Both questionnaire have three subscales: (i) Feelings, (ii) Memory and (iii) Everyday Life, as well as one question about the overall QoL. These are answered on a 4-point Likert scale from 1 (a lot) to 4 (not at all) for the regular questions, and a 4-point Likert scale from 1 (very good) to 4 (poor) for the final question. The DEMQoL Proxy was used for the study presented in Chapter 5.

3.3.11 Cornell Scale for Depression in Dementia

The Cornell Scale for Depression in Dementia (CSDD) was designed to assess signs and symptoms of major depression in people with dementia [8]. It is meant to be used with both, a person with dementia, and an "*informant*" or caregiver. First, a caregiver interview should be conducted, followed by an interview with the actual participant. In the case of discrepancies between the two interviewed persons, the interview should be repeated to dissolve these. Nineteen symptoms are answered on a scale from 0 (absent) to 2 (severe). The questionnaire asks for symptoms of the participant within the last week. If symptoms occur due to other illnesses, or physical disability, these should not be rated. The questionnaire has five subscales: (i) Mood-Related Signs, (ii) Behavioral Disturbance, (iii) Physical Signs, (iv) Cyclic Functions, and (v) Ideational Disturbance. Scores are then added and interpreted. Scores below 6 indicate the absence of significant symptoms of depression, while scores above 10 show probably major depression and scores above 18 indicate definite major depression. The Proxy version of CSDD was used with caregivers of older adults in our long-term study. The results are described in Chapter 5.

3.4 Physiological Measurements

Recording physiological measurements such as heart rate, eye tracking data or movement data in VR research has several advantages. First, it is objective, quantifiable data that can help to complement self-reported data from questionnaires or interviews. Further, it can provide insights into emotional responses, for example, stress, excitement, or experienced workload [20]. Finally, it can help to measure the effectiveness of a training application, for example, by the amount or kind of movement that is performed during a VR exergame [82].

Eye tracking can be used to identify areas of interest (AoI) or to measure the fixation count (FC) or fixation duration (FD) of a participant on a certain AoI [131]. This way, indications about the user's attention can be made. Furthermore, cognitive load can be assessed through psychophysical measurements like blinks, saccades, or pupil dilation [164, 305]. Eye tracking has also been explored as a screening method for the detection of cognitive impairment, MCI, and dementia [265, 323, 374], which could also be integrated in state-of-the-art VR HMDs.

Different physiological measurements were collected and are presented in Chapter 5 (movement data), Chapter 6 (heart rate, movement data), Chapter 7 (movement data), and in Chapter 11 (eye tracking data).

3.5 Thematic Analysis

Thematic analysis (TA) is a qualitative method to analyze text, such as interview data or research papers. In this section, its traditional approach is described after Braun and Clarke [32], as well as the more recent reflexive TA process [33, 35]. It provides a flexible approach that is able to provide a complex, rich and detailed analysis of the given data. Thematic analysis identifies, analyzes and reports patterns or themes in data. There are two main approaches to generating codes: (i) inductive and (ii) deductive coding. In inductive coding, the codes are data-driven, meaning that they are generated directly from the data and not from prior knowledge. In contrast, in deductive coding, codes are driven by theoretical concepts or specific interests, such as the predefined research questions. Usually, a deductive approach is used for specific research questions, and an inductive approach is used when the research questions are evolving through the coding process, or in exploratory research. Similarly, themes can be determined inductively from the data, or deductively from prior research.

Thematic analysis has several phases. First, researchers should familiarize themselves with the data. This involves actively reading the transcript or paper to identify possible codes. In the second phase, initial codes should be generated, with as many potential codes as possible. Individual parts of the data here can be coded several times if they fit into multiple topics. Thirdly, the different codes are sorted into potential initial themes. Here, codes are analyzed and grouped to find out how different codes can be combined into an overarching theme. At the end of this phase, there is a collection of themes, sub-themes and the coded data for them. In phase four, the themes are developed and reviewed, possibly merging or separating themes. First, the data extracts are reviewed in order to determine whether they form a consistent pattern. If single data points do not fit, they can be moved to another theme, or the theme itself can be reworked. Afterwards, the themes' validity is reviewed in relation to the whole data set, and in order to verify that the themes reflect the overall meanings in the data set. In the fifth phase, the themes are defined and named, determining the specific aspect of the data within the theme. Finally, in the last phase, a written report is produced. Here, it is important to provide a logical, interesting account of the data and its meaning. Extracts from the data can be presented in order to support the themes.

Reflexive TA was used for our study described in Chapter 5 with inductive coding under deductive categories. A fully inductive approach was used in our works presented in Chapter 10 and Chapter 12.

Part II

Immersive Cognitive-Physical Training for Older Adults

Chapter 4

Introducing Immersive Training for Older Adults

Figure 4.1: A woman is enjoying a virtual environment. The environment was created with Bing Image creator through the following prompt: "a 3d cartoon flower environment with sunshine". The user was inserted manually afterwards.

4.1 Motivation

As shown in prior work, the demand for well-designed and effective training interventions is rising, and the positive effects in regard to cognitive, physical and mental well-being look promising. Nevertheless, user preferences and requirements might differ based on factors like age or fitness, or external circumstances including the living and care arrangements, or official regulations during crises like the COVID-19 pandemic.

In order to benefit from the enormous potential of exergames for the well-being of older adults, these exergames need to be evaluated for feasibility, user experience and effectiveness. Understanding the preferences and barriers is crucial for refining the training applications, ensuring that they are effective, user-friendly and accessible.

4.2 Research Questions

This chapter presents the results of three studies that were conducted within the scope of the project *EXGAVINE*. Here, three exergames for older adults with and without dementia were iteratively developed. The first exergame, *Memory Journalist*, places the player in the role of a newspaper reporter. In 360° video recordings of famous places in Germany, pictures of well-known landmarks have to be taken, encouraging reminiscence activities [197] and social interactions with caregivers. This exergame was evaluated with older adults with dementia over the course of nine weeks. The methods and results are shown in Chapter 5. The leading research question for this user study was:

- $\mathbf{RQ}_1$: Can a VR exergame improve cognitive, physical as well as psychological well-being of older adults with mild dementia?

The second exergame, *Maestro Game*, immerses the users into the role of a conductor in an opera hall. Their task is to conduct three musicians to unfold their whole potential and make them play together. In order to do this, the player has to follow a dynamically generated path with their hands, encouraging upper body movements. At that time, in-person training was largely restricted or not possible due to the COVID-19 pandemic, leading to the need for home-based training options. Therefore, we compared it to traditional video-based training and evaluated factors such as intrinsic motivation, workload and spatial presence. For this study, which can be found in Chapter 6, the leading research question was:

- $\mathbf{RQ}_2$: Can a VR exergame be an alternative training option to a traditional video-based training for older adults?

The final exergame is the *Canoe Game*. Here, users sit in a virtual canoe in a tropical island scenario. While it runs down a river, the user has to collect soap bubbles with their hands. This is combined with cognitive tasks, e.g., only collecting certain bubbles, and therefore features both physical and cognitive exercises. Again, the evaluation was shaped by the COVID-19 situation at that time, restricting access to large numbers of special user groups. Therefore, the game was evaluated with both, younger and older adults. Further, we noticed the need for adjusting the exergame difficulty for differences in the fitness of different users. Therefore, this exergame can be adjusted in terms of physical and cognitive difficulty. The goal of the third user study, described in Chapter 7, was (i) to evaluate, which difficulty adjustment leads to which effects, and (ii) to compare the perception and preferences of younger and older adults in terms of difficulty. The global research question we formulated for this study was:

- $\mathbf{RQ}_3$: How does the difficulty of an exergame impact user experience and performance, and which insights can be derived from evaluating this with different user groups (i.e., younger and older adults)?

Finally, in Chapter 8, the results from all three user studies are discussed and design considerations are derived.

Chapter 5

Improving the Well-Being of Older Adults with Dementia Through VR Exergames

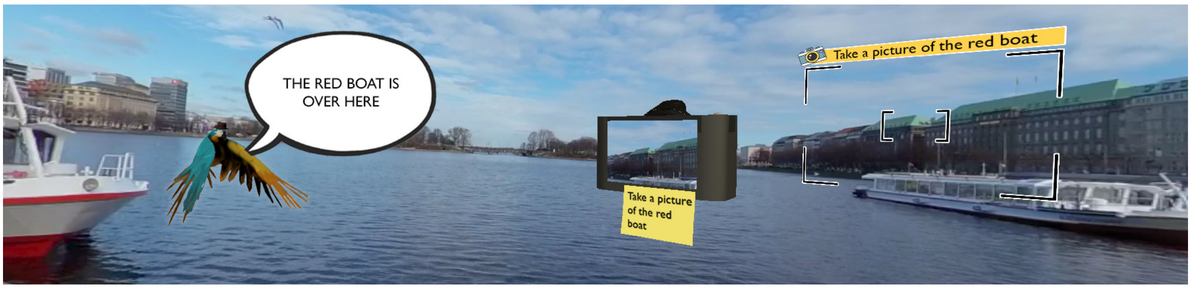


Figure 5.1: A parrot guide helps the player locate the target.

In our first user study, the goal was to investigate whether a VR exergame for older adults with dementia can improve their psychological, physical and cognitive well-being in a long-term study (**RQ**$_1$). For this, we designed an exergame where the players experienced 360° videos of famous places and had to take pictures of certain landmarks. We then evaluated the game with five older adults with dementia, while five other older adults participated in a control group. This chapter is based on the following publication:

Kruse, L., Karaosmanoglu, S., Rings, S., Ellinger, B., Apken, D., Mangana, T. F., and Steinicke, F. (2021). A Long-Term User Study of an Immersive Exergame for Older Adults with Mild Dementia during the COVID-19 Pandemic. In Proceedings of the International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments (pp. 9-18). DOI: `https://doi.org/10.2312/egve.20211322`

5.1 Introduction

With the demographics change and an ageing population, dementia has become one of the most common causes for disability or neurological diseases in later life [246].

Such diseases cause a decline in people's cognitive and physical abilities, restricting their autonomy. Different stages of dementia require individual care, and daily life tasks often require the help of healthcare professionals, imposing social and financial challenges for families and health systems [246]. Therefore, research should address the question of how we can prevent, reduce or overcome such more frequently occurring conditions. In particular during the early stages of dementia, the WHO pointed out the importance of remaining active as long as possible and training both cognitive and physical abilities [246].

VR has great potential to train cognitive and physical abilities since it offers possibilities of creating individualized immersive training programs, which can easily be adapted [81, 270]. VR can also help to provide activities that would otherwise not be possible, e.g., visiting other countries or places [276]. Especially for older adults with dementia living in senior living homes, the social and economic resources are limited, and the caregivers' workload is enormous. Hence, enabling activities for people with dementia such as taking a trip to the city requires a lot of preparation time, transportation and personnel. With the help of VR, people with dementia can virtually visit outside locations without increasing the workloads' of their caregivers, and especially during times of a global pandemic, without increasing their risk for infection.

To fully leverage the benefits of immersive VR technologies, we have developed *Memory Journalist*, an immersive VR exergame in which people with dementia can virtually visit famous places in their hometown (Hamburg), and the capital of Germany, Berlin. Furthermore, we provided a 3D printed replica of a camera device, with which the participants have to take pictures of specific objects or locations, stimulating their visual recognition, short-term as well as long-term memory. The game includes simple functional training activities generated by the task to take pictures, which requires users to perform head and torso rotations combined with arms-raising exercises. During a nine-week user study, which was conducted in cooperation with a senior living home, Hospital zum Heiligen Geist, we evaluated the effects of the immersive VR exergame on physical and cognitive abilities, and psychological well-being. The evaluation was based on interviews, questionnaires and data collection during the gameplay. The fact that the study was conducted during the COVID-19 pandemic clearly emphasized the importance and potential of using VR as a means of physical and cognitive training for isolated vulnerable groups since it allowed older adults with dementia to stay active. Overall, the long-term user study was driven by the question whether a VR exergame can improve psychological, cognitive and physical well-being of older adults with dementia ($\mathbf{RQ}_1$), which was divided into three sub-research questions:

1. $\mathbf{RQ}_{1.1}$: Is *Memory Journalist* suitable for people with mild dementia, providing enjoyment, engagement and a positive user experience?

2. $\mathbf{RQ}_{1.2}$: Can VR exergames improve the psychological well-being of the players during times of a global pandemic?

3. $\mathbf{RQ}_{1.3}$: Can playing a VR exergame on a regular basis be a suitable addition for maintaining or even improving cognitive and physical abilities of people with

dementia?

This chapter contributes to the knowledge of the effects of VR exergames on psychological, cognitive and physical well-being of people with dementia. We extend existing research by conducting a long-term study with people with dementia and their caregivers during the COVID-19 pandemic. In comparison with a control group, who did not receive any additional training, the impact of the game will be discussed. Finally, we reflect on the implications for future research, highlighting the need for immersive exergames for people with dementia, especially in times where social contact is limited.

5.2 Methods

To develop *Memory Journalist VR*, we followed an UCD approach, which is described in more detail in [150]. First, we conducted semi-structured interviews with stakeholders (i.e., family members of people with dementia, clinical health experts, a physiotherapist) and employed contextual inquiry sessions with people with and without dementia. Then, based on the user requirement analysis, *Memory Journalist* was designed and developed. We then conducted five focus group sessions to demonstrate validity of the designed exergame. Each focus group session served as a stimuli to iterate and improve the user requirements, design, and implementation of *Memory Journalist*. The short-term evaluation of the exergame highlighted positive player experience for people with dementia, yielding the investigation on the long-term use.

5.2.1 Application

In *Memory Journalist*, the player experiences 360° 3D videos of their city of residence (Hamburg) and a second famous city in their country (Berlin). While the participant plays the VR exergame, a caregiver accompanies them through the entire experience. The players take the role of a reporter, with the task to take photographs of famous landmarks for a newspaper. Our aim was to create a realistic environment to evoke the long-term spatial memories of people with dementia by providing a meaningful experience linked to their past [80, 82, 128, 133, 150, 307]. If the participants had not visited the places before, the help of their caregiver permits them to play the game successfully, encouraging a social interaction between them [150]. A custom 3D-printed camera is used as an input device to increase the realism of the experience. A virtual version of the camera is also shown in VR. This posed an intuitive user interface for the older adults, whose affordances most of them knew from the real life [150].

Six different locations can be shown, with three of them in each city. Each location has six different missions, e.g., taking a picture of well-known churches, towers, museums or rivers. Once the picture is taken correctly, a short newspaper article with a fictional, yet fitting headline appears (see Figure 5.3 middle). Some missions provide an additional difficulty by requiring an upright or zoomed-in picture. For an upright picture, the camera has to be rotated. Zoom can be achieved by moving the camera further away from the player's head. While these movements appear effortless for younger and fitter players, the movements provide important physical training for everyday life of people with physical difficulties, e.g., for lifting or placing objects [270].

A virtual agent in the shape of a parrot is used as a narrator that presents the missions to the players. The representation as a parrot was chosen because of the parrot's ability to communicate naturally, and their ability to fly around in space, enabling natural movement in the 360° scene. If they need help, the parrot can provide two hints. The first hint is always a textual explanation of the landmark that the players are looking for, e.g., "*We are looking for a big building with a turquoise roof*". If more help is needed, the second hint is provided optically. The parrot is flying towards the landmark saying "*[The landmark] is over here*" (see Figure 5.1). Overall, the use of the parrot allows both, players and caregivers, to solve tasks if they do not know the famous landmark. A caregiver can control the VR experience using a *Remote App* on a smartphone (see Figure 5.3), which was implemented as an Express `(v4)` [245] web server using NodeJS `(v8.9.4)` [59]. With this app, missions can be started or aborted, hints can be given and the game can be stopped. Further, this eliminates the need for participants to learn a new interaction mechanic, but the control of the game flow remains in the hands of the player and their respective caregiver.

We developed the game using the Unity3D [327] game engine. The game ran on a computer with a GeForce RTX-2080 Ti graphics card and an i9-9900K CPU. A Valve Index HMD was used, with a resolution of 1440px × 1600px per eye at 120 Hz, and a field of view of 130°. Additionally, a 3D-printed camera with an attached Vive tracker was used. The shutter-button of the camera could be pressed to take a picture in-game. For the video recordings of the environment, we used an *Insta360 Pro* camera [141]. The Insta360 Pro consists of six horizontally-aligned cameras in a ring. Through stitching the images of the cameras together, 360° stereoscopic videos can be recorded. To prevent cybersickness [43] and for safety reasons, the camera was placed on a tripod at 120cm height from the ground, in order to replicate eye height for a seated VR experience. The game view was mirrored on a TV screen, allowing spectators to follow the game flow. The players were either sitting on a stationary, rotatable chair, which made it easier for them to look around in the virtual scene, or in their own wheelchair, which had to be rotated by the healthcare professional. This allowed the players to enjoy a seated, save experience while addressing their accessibility needs [81, 127].

5.2.2 User Study

The aim of the study was to investigate whether the developed VR exergame was able to provide a positive experience for people with dementia in long-term use. Several studies have shown the negative impact of social isolation and the cancellation of training options on psychological, cognitive and physical well-being [1, 201, 213, 261], which is why we and the senior living home agreed that it is important to continue our work, of course, under all hygiene standards. Furthermore, the end of the pandemic was still

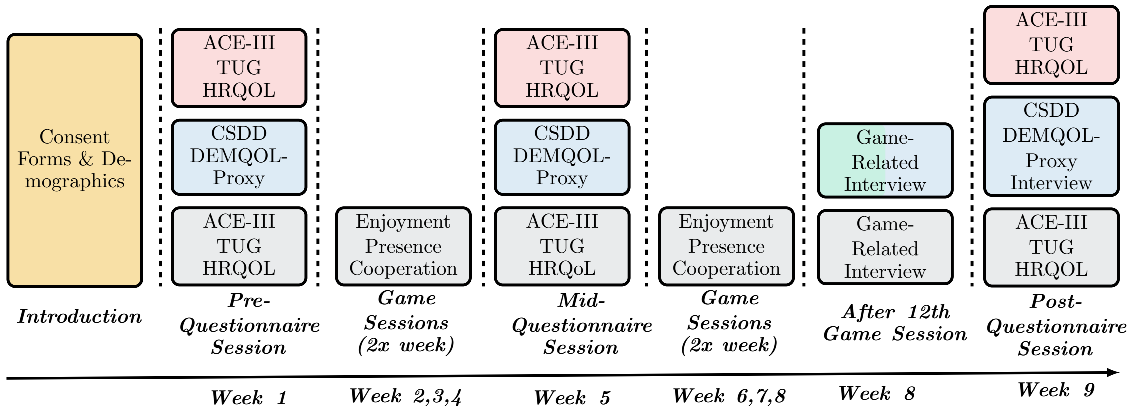


Figure 5.2: The schematic procedure of the study. The colors represent the group of people performing the task: orange (all participants), grey (test group), red (control group), blue (caregiver), and green (healthcare professional).

unknown, and positive experiences, especially at a later stage in life, are crucial for the well-being of older adults [279]. To determine whether *Memory Journalist* can be this positive experience, or if it is due to external circumstances, e.g., less social distancing, less infections, or even the season of the year, we found it important to compare one group playing the game, and one group not playing the game in a between-subject design. We decided against a control activity for the control group to investigate whether providing the people with dementia with a VR exergame as an addition to their normal life activities has an influence on their well-being.

The COVID-19 pandemic posed several challenges for research [72]. We took special care of all hygiene measurements, in accordance with the senior living facility. The study was carried out during the late summer months with relatively low COVID-19 infection rates in the VITALab.One [125], which is located in a separate building of the senior living facility. An outside door was open at all times, and medical masks were worn. Between each participant, the HMD, controllers and surfaces they touched were cleaned and disinfected. A healthcare professional was responsible for putting the HMD on the participant's head, so the researchers did not get close to them to ensure social distancing.

The study was part of a project with multidisciplinary researchers, which also included an ethics expert. Additionally, we held a so-called ELSI workshop to discuss and reflect on ethical, legal, societal, privacy and security issues prior to the study. The knowledge gained during this workshop was integrated into our work procedure. Furthermore, the ethics expert visited us during one of our earlier prototyping sessions, where they experienced the game and the interaction of the older adults with the game, and discussed the ethical problems and indications of the study. We then incorporated their feedback once again. In cooperation with the senior living home, we determined safety protocols for both, mental stress, as well as physical stress or dangers due to the COVID pandemic. For example, stop markers were developed with caregivers to peacefully end the experience in a case of emotional distress or disorientation. The legal representatives of the participants were informed about the study and signed a consent form, and of course the participants themselves participated out of their own will, and were free to stop whenever they wanted.

5.2.3 Participants

10 participants (7 female, 3 male), diagnosed with mild dementia, were recruited from the senior living facility. At the beginning of the study, the participants were randomly assigned to the test group (TG) (4 female, 1 male) or the control group (CG) (3 female, 2 male). The average age of the participants was 82.8 ± 5.49 years (TG: 83.6 ± 7.42 years, CG: 82 ± 2 years). The study was conducted with a healthcare professional (female, 52 years) participating in the game. Also, the responsible caregivers of all users (N=16, 14 female, 2 male, 48.06 ± 12.84 years, 11.75 ± 10.48 years of experience) participated in the questionnaire sessions as proxies and observers.

5.2.4 Study Procedure

The study took place over the course of 9 weeks, with 6 weeks playing the game twice per week, and 3 weeks of questionnaire sessions. The procedure can be seen in Figure 5.2. We chose to have the questionnaire sessions in separate weeks to rule out any short-term effects of the game on cognitive, physical and psychological states of the participants. After the last questionnaire session, each of the players received a printed newspaper with the pictures that they took during the game as a thank you.

Figure 5.3: Left: An older adult playing the game, with the healthcare professional assisting him. The experimenter guides the game with a remote control from their phone. Middle: A newspaper article that appears when a correct picture is taken. Right: An older adult playing the game by herself, with the healthcare professional using the remote control.

Questionnaire Sessions

Several standardized questionnaires were used to assess cognitive, physical and psychological changes. Because some people with dementia might have difficulties in answering questionnaires [58, 160, 339], we also used proxy measures that were answered by the responsible caregivers. A senior home employee filled out the ACE-III, A-B-C subversions for each session with all participants [135]. For a self-reported HRQoL rating, a customized single-item from the Dementia Quality of Life (DEMQoL) questionnaire was employed ("*Thinking about all of the things: your feelings, memory and everyday life, how would you rate your quality of life overall?*") [315]. The participants' mobility was measured with the Timed Up and Go Test (TUG) [255]. At the same time, an experimenter interviewed the caregivers of all participants. Caregivers answered DEMQoL-Proxy and the Cornell Scale for Depression in Dementia (CSDD) for the participants [8, 315, 365]. Because of changing work schedules, the caregivers were not the same throughout the sessions (See limitations).

Additionally, two types of semi-structured interviews were conducted: a game-related interview and a follow-up interview. The questions can be found in Appendix A. The game-related interview took place directly after the 12th game session. This interview aimed to obtain information regarding the game experience of the players, and the impact of the game on their well-being. A follow-up interview was conducted in the post-game questionnaire session with all caregivers. The purpose of this interview was to observe the possible long-term effects of the VR exergame on the well-being of the participants, and to compare it to that of the control group.

Game Sessions

Each game session consisted of 15-25 minutes of game play and three questions afterwards. In a tutorial, the players received their camera, and were asked to press the shutter button to take their first picture. Afterwards, a scene of one of the six different locations started. The participants had some time to look around and recognize where they were. Then the healthcare professional started the first mission, which asked the participant to take a picture of a specific location or landmark in the virtual scene. During the game, we recorded the position of the 3D-printed camera, and the reaction times.

Each game session presented three different locations with four different missions each. The locations and missions were assigned randomly in a counter-balanced way. When four missions were completed, the scene changed to a different location. As soon as 12 missions were done, or the players decided that they wanted to stop, the game ended. Then, the participants were asked about their enjoyment, feeling of presence and how well the cooperation with the caregiver worked on that day. Enjoyment and cooperation scores were acquired via custom single-item measures on a 7-point Likert scale (respectively "*I enjoyed this game experience.*" and "*I enjoyed the teamwork with the experimenter.*"). For presence, a single question from the Slater-Usoh-Steed Presence Questionnaire was asked ("*I had a sense of being at [the place].*" on a 7-point Likert scale) [313].

5.3 Results

This section covers the qualitative and quantitative results. We employed mixed-method data analysis and focused on both quantitative and qualitative data. Due to the small sample size and drop-outs from the study within the given COVID-19 restrictions, we did not focus on statistically significant differences on quantitative measures, but will report trends that emerge from the data. For detailed results, please refer to Appendix A.

For the qualitative evaluation, we used steps from both reflexive and codebook approaches of thematic analysis [32, 34]. The main researcher decided on four categories focusing on our research questions before the analysis: psychological, cognitive, and physical well-being, and player experience. The interview data was then coded inductively under these deductive priori categories.

5.3.1 Dropouts of Participants

Two test group participants could not participate for the first two sessions because their house was on lock-down, and each participant missed one or two game sessions because of doctor's appointments, sickness, or quarantine regulations. During the mid-questionnaire session, there was a lock-down imposed due to COVID-19, so data for one participant from the test group and one from the control group are missing. From the control group, one participant had an accident shortly before the mid-questionnaire session and stopped participating (CG_2). We also note that one player dropped out after the mid-questionnaire session according to their own wishes (TG_1).

5.3.2 Psychological Well-Being

We observed a decreasing trend for CSDD scores, indicating lower depression for the test group (pre: 4.2 ± 3.03; mid: 2 ± 2.45; post 0 ± 0). For the control group, the CSDD values showed a decrease from pre (1.8 ± 1.64) to mid (1 ± 1.15), but indicated an increase later (post:1.6 ± 1.14). DEMQoL and HRQoL did not show any trends.

The interviews showed that the psychological well-being of the test group participants was affected in various positive ways. First, in terms of mood: "*I would say [TG_3's] mood afterwards was euphoric, in contrast to depressed before*"-Cargiver of TG_3 (CTG_3). Second, joy of life: it has improved over the course of the study, which is stated by multiple caregivers: "*[TG_5] has fun and it is a great change to [TG_5's] daily life, especially in times of Corona where the offer of care is very limited*"-CTG_5 and "*[TG_4] is always happy when [TG_4] is picked up and we say 'Today we are going to [play the VR exergame]'*"-CTG_4. A large improvement was shown for TG_3, who showed signs of delusion before the study started. CTG_3 reported that "*This has decreased noticeably, we hear this very rarely now*". A third aspect were social interactions. While some players, e.g., TG_3 normally were rather reluctant to talk to others, CTG_3 reported that "*when [TG_3] comes back, [TG_3] also talks about it. Well, [TG_3] had a smile on their face and* *said 'Wow, that was so!'*". CTG_5 also said that "*[TG_5] is looking forward to these dates, and afterwards we can greatly talk about it*".

5.3.3 Cognitive Well-Being

The cognitive task during the game was to finish all missions as quickly as possible. The mean time spent in VR for each session was 16 minutes. On average, participants needed 18.5 seconds (SD = 12.5s) to fulfill a mission after receiving new instructions. A summary of the time needed to complete one mission is shown in Figure 5.4. Since we would be expecting a learning effect due to the repetition of the exercises, we have superimposed a fitted exponential curve to support the improvements in our players ($f(x) = 23.49 \cdot e^{-0.009413x}$). More data is needed to verify this, as this curve only explains around 50% of the variability of the data ($R^2 = 0.5090$).

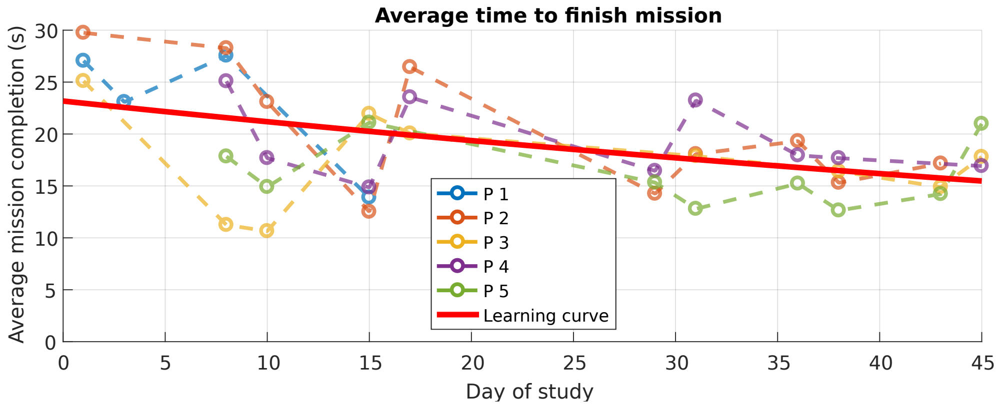


Figure 5.4: Time needed to finish a mission in *Memory Journalist*.

The total ACE-III scores of the test group rose steadily between the sessions, (pre: 60.4 ± 15.96, N=5; mid: 61.25 ± 17.29, N=4; post: 64.75 ± 13.6, N=4). For the control group, the mean total scores were 68.2 ± 17.28 in pre (N=5), 80.33 ± 1.53 in mid (N=3) and 76.5 ± 10.85 in post (N=4). It can be seen in Figure 5.5 that on average, the participants from the test group improved their score more than the control group. However, we note that given the dropout status of some participants, these scores should be evaluated at an individual level. Especially TG_3, who had a low score of 33 in the pre-session, improved to 46 in the post session, equaling an improvement of 39%.

In the interviews, the impact of the COVID-19 pandemic on cognitive abilities of people with dementia was emphasized several times. CTG_3 explained that "*the problem is that we could not do so many groups [for cognitive training]*", which might be a reason for declining cognitive abilities. CTG_2 also stated that "*The cognitive stimulation gets worse now, that is also because no activities take place. We notice this in our inhabitants in general, that cognitive [abilities] decrease a little and the forgetfulness gets more in focus now*". The caregivers saw the VR exergame sessions as an opportunity to provide cognitive stimulation for participants during COVID-19. Further, a caregiver

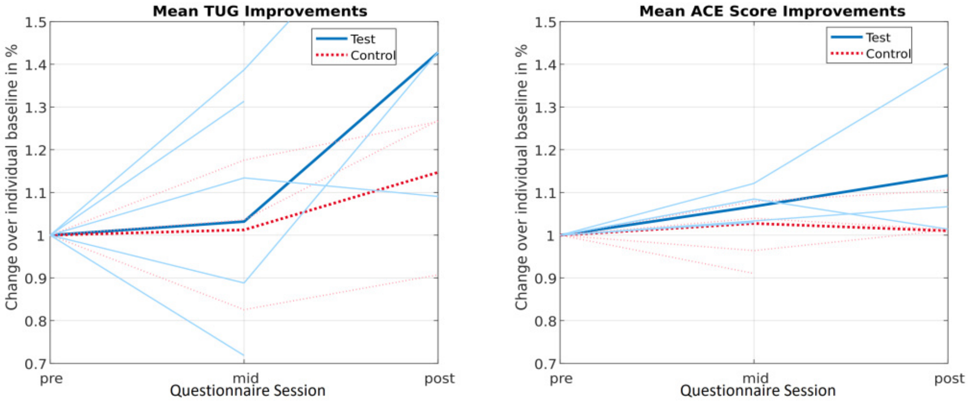


Figure 5.5: Left: TUG improvements. Right: ACE-III improvements. Bold lines show the mean value, thin lines the individual changes. The pre-session is used as an individual baseline for each participant.

explicitly indicated that one player showed signs of improved memory because of *Memory Journalist*: "*It is a small improvement. I think so, yes. [...] I even think this has improved positively*"-CTG_3. Additionally, caregivers stated that the mood and abilities of the participants depended on the day (CCG_1, CTG_2, and CTG_5), which might help to understand underlying reasons of our varying quantitative results.

5.3.4 Physical Well-Being

Movement of the 3D-printed camera was evenly distributed around the y-axis, which means that the players rotated on their chair and took pictures in every direction. The movement intensity of the camera supports the functionality of our exergame. Most movements occurred close to the head of the player, holding the camera close to their eyes. The biggest training was the act of zooming, for which the camera had to be moved away from the players' head. The mean total distance that the camera, and therefore the player's hand, traveled during one session was 26.4 meters (SD $= 7.64m$). Additionally, a training effect of the participants can be observed in less scattering of range of motion in the last three weeks of playing, compared to the first three weeks. This is a typical learning effect, as the players needed less movement to fulfill the task, meaning they get more precise.

All participants except for CG_5, who was sitting in a wheelchair, participated in TUG. Mean times for the test group decreased, starting from $29.93s \pm 29.74s$ in pre (N=5), $25.63s \pm 22.49s$ in mid (N=4) and decreasing to $21.3s \pm 16.14s$ in post (N=4). Times for the control group also decreased, with $26.22s \pm 16.96s$ in pre (N=4), $18.43s \pm 6.68s$ in mid (N=3) and decreasing to $16.24s \pm 5.92s$ in post (N=3). The improvements can be seen in Figure 5.5 (left). Looking at individual improvements, participants in the test group improved by 42.7% on average, and participants in the control group improved by 14.71%. TG_4 improved the most by 47%, almost needing half as long in the post-game session, as in the pre-game session.

In our study, we observed that players had varying physical abilities. For example, TG_4's caregiver explained that "*Sometimes TG_4 is picked up in their wheelchair, sometimes they walk by themself with their walker*". This might be a reason for their improving score. The second topic was pain, or the absence of it. The caregiver of TG_3 reported: "*before [the study started], they had extreme shoulder pain. [TG_3] had an accident, long ago, and now [TG_3] does not complain about it anymore.*"-CTG_3. While holding the physical camera was considered a good training by most older adults, TG_1 reported difficulties due to a shoulder injury that they received when using their walker. This was also a reason for them to drop out from the study. From the control group, only the caregiver of CG_2 reported that CG_2's physical well-being was worse due to their accident, while all other control group caregivers reported no physical change during the study period.

5.3.5 Player Experience

The player experience stayed high throughout the game sessions, with a little lower scores in the very first and around the middle of the game sessions (mean values for enjoyment: 5.61 ± 0.59; presence: 5.3 ± 1.16; cooperation: 5.88 ± 0.25). Results over the course of the study are shown in Figure 5.6.

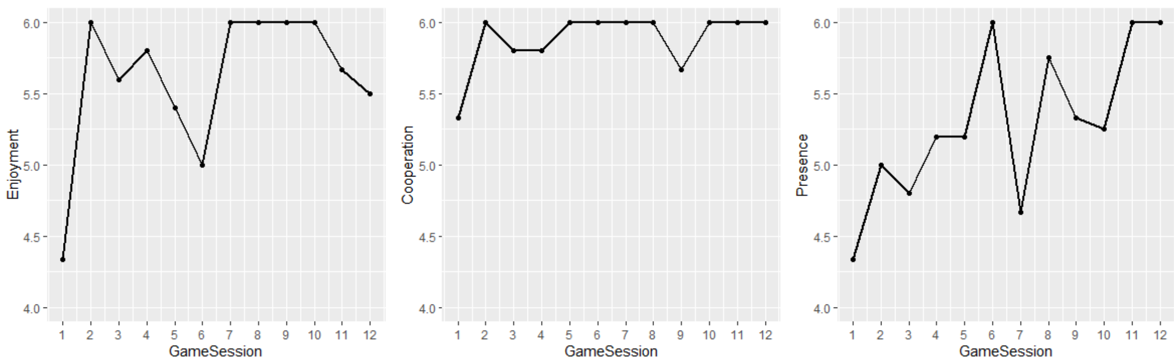


Figure 5.6: Plots showing (i) the enjoyment, (ii) the cooperation with the caregiver, and (iii) the presence in the virtual world over the course of the game sessions.

The qualitative findings of the game-related interviews also support the quantitative findings by showing high player experience for the players. For example, TG_2 reported that "*[The game] was fun. I liked to come here*", and that there was "*no boredom at all. [There] was always something to see*". The high player experience is further supported by the healthcare professional: "*the joy we brought to [TG_3]. That was great. [...]*". Further she pointed out the feeling of accomplishment: "*[TG_3] came into the room scared, then we sat [them] on the chair and [they] said "I hope I can do this". And when [TG_3] saw the pictures, total enthusiasm and totally relaxed. No fear, and no fear to fail at all. [...] It was like this until [the last session]*". More generally, she evaluated: "*I see how much you achieve and how much fun [the older adults] have. [...] We make them really happy and I think that is wonderful*". The healthcare professional also explicitly pointed out the possibilities of this VR exergame: "*I found [the game] super good, because everyone that has a person [with] dementia at home or takes care of them knows what an effort it is to travel to [the center of] Hamburg or to other cities.*", emphasizing how the exergame can achieve this is seconds and provide older adults with dementia with the opportunity to travel.

Furthermore, the realism of the virtual world was praised by many participants, yielding a feeling of presence: "*I am shaking. Birds everywhere! This is so real. I sometimes ducked my head [laughing]*"-TG_3. Moreover, the VR exergame created a feeling of curiosity: "*I felt well there. [I liked that] there were different motives. It was always interesting*"-TG_4 and "*the game was something new. And if it is new, it is interesting. It is something different. You talk to other people, you see other things, new things, new pictures*"-TG_5. The healthcare professional emphasized the repeatability of the VR exergame, indicating that some participants could play the game even longer: "*I think [TG_3] could have easily played not just 10 minutes, but half an hour or 45 minutes and she would still have been enthusiastic*", and "*[TG_5] would have liked to keep playing, she would also have been a candidate where you could say half an hour, 45 minutes*". All participants reported that they would recommend the game to others, and when we asked TG_3, they answered: "*I already did!*".

TG_1 -who only played until the 6th session– stated that they "*liked [the game], yes, yes. [It] was good, good video captures. [...] It was quite an experience if you see it like that*". However, they also reported it got boring after a while because they already knew all the places to visit. All other participants did not have any suggestions on improvements. TG_4 said: "*I think you have done this game in the optimal way.*"

5.4 Discussion

Below, we discuss the findings focusing on our research questions.

5.4.1 Psychological Well-Being

Our results in the mid-game questionnaire session show a decreased quality of life. This was probably caused by severe COVID-19 restrictions including a full lock-down of one of the houses. Previous research has found similar decreases of perceived quality of life due to such isolation measures [110, 213]. In contrast to the control group, the psychological well-being of the test group showed improvements. Especially in times of COVID-19, when all other group activities were canceled, the individual game sessions gave the players something to look forward to, and something aside from their daily routine. In addition to the positive short-term effect of improved psychological well-being while the participants played, we observed a positive mid-term effect over the course of the study. This empathizes the importance of situation adjusted activities for people with dementia in senior living homes. We argue that *Memory Journalist* can contribute to the well-being of older adults with mild dementia, especially in later times, when more caregivers are trained to use the system or other means of assisting the older adults in using the game are developed.

5.4.2 Cognitive Well-Being

The quantitative results of the test group's ACE-III scores were higher than the control group's. Due to COVID-19, cognitive training groups had to be reduced and the individual caregivers' workload was too high to provide individual training. *Memory Journalist* and the cognitive training associate with it may have led to the test group's higher scores. Furthermore, the mood and cognitive abilities of people with dementia can vary from day to day, and a positive mood can lead to a better cognitive performance [194]. As discussed above, the game improved the participants' psychological well-being and we assume that this might also have positively influenced their cognitive well-being. On the other hand, it is unclear whether these effects would also be shown under normal circumstances, without an ongoing pandemic. This needs to be tested in future studies.

Even though the short-term memory of people with dementia is often impaired, the game was remembered by all of our participants, who talked about it to their neighbors and caregivers, or made comments about looking forward to the game session. Additionally, their speed in finding the required landmarks improved, hinting for some form of remembrance. Especially those participants with a lower cognitive score in the pre-session improved by playing the exergame. This implies that the training aspect of this game is most suitable for them, following the training principle of "*Initial Values*" states that the improvement of an outcome will be greatest in those with lower initial values [79]. Therefore, we argue that playing *Memory Journalist* has the potential to maintain and improve cognitive skills.

5.4.3 Physical Well-Being

During the game play, players successfully fulfilled the required movements to take a picture, even though some of them needed assistance to rotate on their chair. This implies that even though the movements are easy for most younger or fitter people, our participants had to make an effort to perform them.

In TUG, it could be seen that almost all participants from both groups increased their speed between the sessions. While this could be a simple training effect of repeating the test itself, the test group's improvement was larger. Similar to the results for cognitive training, the players with the highest physical problems improved the most. This implies that even a low-intensity activity in VR can maintain or improve physical abilities. One player complained of shoulder pain before the study, but did not do so anymore afterwards. This might be due to the physical training they received, but of course, a more detailed examination is necessary to confirm this.

While we employed the TUG test to measure physical well-being, we argue that it is not always applicable for people with dementia due to varying physical conditions (e.g., when using wheelchairs). Future studies should either employ inclusion criteria for participation or consider these variances and thereby use different measurements for physical well-being.

5.4.4 Player Experience

All players reported that they liked the game and that it was fun to play for them, which replicated previous findings in the usage of exergames [82, 270]. Although all our participants had mild dementia, their short-term memory abilities varied. Especially the participants with worse short-term memory were enthusiastic every time they played the game, while it became boring or too slow for those who remembered it after a few sessions, similarly to older adults without dementia who tested the game [150]. We argue that playing the game regularly is the most suitable for people with increased dementia symptoms, while infrequent playing sessions will invoke feelings of joy even for users that remember the game and the videos. Therefore, future development should add more variety and videos from other places to make the game less repetitive for everyone. Furthermore, it should be noted that the participants did not indicate the need for any improvements of the game, but this might also be due to their difficulties in expressing themselves or inexperience with immersive experiences [67].

5.4.5 Limitations

One obvious limitation was the group size, with five people in each group. COVID-19 restrictions and health problems of the participants additionally made the quantitative analysis less meaningful. However, this also implies that future studies should be aware of the challenges of conducting long-term studies with this special user group, in particular under pandemic circumstances.

Another limitation is that the control group did not perform any activity while the test group played the game twice a week. It is important to note that our exergames should not be seen as a replacement for traditional therapies, but as an extension, as indicated by Zeng et al. [387]. Our current study design empathized how important it is to provide additional activities for older adults, especially in times of social distancing and if their traditional activities are not possible. Additionally, control activities like actually going outside to take pictures were not possible under the circumstances back then, but could be compared in future studies.

The caregivers (who performed CSDD and DEMOqL-proxy measures) were different in the questionnaire sessions due to changing work schedules; this decreases the comparability of these metrics across questionnaire sessions. Future studies should avoid these variances if possible, and the personal connection between caregivers and the older adults needs to be taken into account. Additionally, administering the exergame proved to be a considerable time commitment for the healthcare professional, but their involvement was essential for enabling the gameplay experience. Therefore, this has to be accounted for when exergames are employed outside of the research context, or for future studies.

5.5 Conclusion

We have developed and evaluated a VR exergame for people with dementia, which provides the possibility of participating in outside activities again, provides cognitive stimulation, and encourages them to perform functional physical movements. Even though our user study had a relatively low number of participants, the quantitative findings of the work indicate a higher cognitive, physical and psychological well-being for the test group compared to the control group. Particularly the participants with more difficulties benefited from the additional training. Over the course of the study, presence, enjoyment, and cooperation values remained high, indicating high long-term engagement with this VR exergame.

Furthermore, our qualitative findings indicate that the immersive exergame can help to improve and maintain cognitive, physical and psychological well-being and can, to some degree, counteract the negative impact that the COVID-19 pandemic had on the participants. The meaningfulness of the game, which gave the participants a relevant task, a distraction from their daily life and which left them with a feeling of accomplishment, was pointed out. The game was also a relief for caregivers and family, helping them with the provision of meaningful activities in a time where most exercise groups were not possible anymore.

We want to highlight the importance of using VR technology not only for younger users, but also for older adults and people with dementia. Especially during a pandemic, we believe that VR technology has the potential of participating in activities that are otherwise not possible, e.g., visiting different places, and it thereby has the potential to decrease the negative effects of social isolation. We argue that it is important to integrate new training methods into senior living homes as they can be seen as a good addition to traditional activities; especially when one-on-one or group training cannot take place. Most importantly, the positive impact of the experience on psychological well-being was obvious in the participants' reactions and interview findings - encouraging us to continue this meaningful research.

Chapter 6

Comparing a VR Exergame to Video-Based Exercises

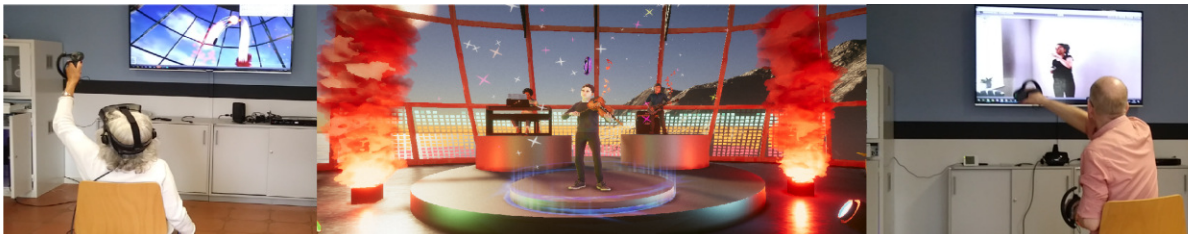


Figure 6.1: Left: An older adult playing *Maestro Game*. Middle: The virtual stage. Right: An older adult doing his workout with the 2D video.

In the last chapter, a VR exergame was evaluated with older adults with dementia and we found a positive effect on psychological, cognitive and physical well-being. Moving on, we are interested to see whether another VR exergame can also elicit similarly positive results in a short-term evaluation. Further, our first study lacked a training comparison condition, therefore it is unclear whether the results could also be found with a different approach, e.g., through video-based training. Due to the fast-evolving COVID-19 situation at the time that the study was conducted, our original target group (older adults with dementia) was no longer safe to conduct studies with. Therefore, we evaluated this exergame with older adults without dementia. In this study, our main goal was to determine whether a VR exergame can be a suitable alternative for traditional, video-based training. This chapter is based on the following publication:

> **Kruse, L.**, Karaosmanoglu, S., Rings, S., Ellinger, B., and Steinicke, F. (2021). Enabling immersive exercise activities for older adults: A comparison of virtual reality exergames and traditional video exercises. Societies, 11(4), 134. DOI: `https://doi.org/10.3390/soc11040134`

6.1 Introduction

As discussed before, for older adults, it is important to remain fit as long as possible to maintain autonomy and live a healthy, independent life. In this context, regular exercises can help to enhance and sustain both, physical and cognitive abilities [130, 183, 193]. In particular, the combination of physical and cognitive exercises can reduce neurological degeneration and improve some of the symptoms of dementia [94, 109, 116].

During the COVID-19 pandemic, physical training groups, sport teams, the access to gyms and individual physiotherapist sessions were limited, and, therefore, physical activities of older adults have decreased [289]. The unavailability of group-based physical activity, as well as several lock-downs, led to the need for alternatives such as home-based training.

Many older adults tend to prefer face-to-face interaction, and the direct support of physiotherapists for exercises [64, 306], however, in times when this contact is not possible, alternatives have to be provided. Exergames are one possibility for home workouts. Many younger people are already familiar with online video classes or VR exergames for training, and 11% of people have increased their use of online fitness videos during the pandemic, compared to their usage before [65]. While there are a lot of exercise videos for older adults online, VR exergames are rarely used by this age group [64]. Furthermore, 2D video-based training has limitations regarding the possibilities to show complex movements from different (also egocentric) perspectives, or to track and record these motions. Moreover, VR has the potential to provide more engaging and immersive experiences, whereas it also comes with the cost of more intrusive and bulky technology. Hence, it remains an open research question whether or not immersive VR exergames can provide an acceptable alternative for 2D exercise videos for older adults.

In this chapter, we compare a traditional, recorded 2D exercise video to a VR exergame that aims to train similar muscles and body parts. The video shows familiar movements performed by a real human trainer, while the VR exergame offers an immersive virtual environment where users can practice 3D movements, with virtual agents in front of the user. The goal of our evaluation was to analyse whether VR exergames are an acceptable, comparable alternative.

We conducted a within-subject user study with 25 older adults that performed both exercise conditions, and then commented on their perceived workload, attention, enjoyment, and which program they preferred. Furthermore, we analyzed the movements and heart rate values measured during both conditions and discuss the advantages and disadvantages of both programs.

Our research was driven by the following research question:

- $\mathbf{RQ}_2$: Can a VR exergame be an alternative training option to a traditional video-based training for older adults?

Following our research question, we state the following hypotheses:

- H_1: The video and VR exergame will show no significant difference for perceived workload, attention, and enjoyment.

- H_2: The VR exergame will be preferred by the same number of people as the traditional video of well-known exercises.

- H_3: Older adults will prefer doing the exercise video for training in long-term use.

H_1 was developed due to the similar length and difficulty of the exercises in the video and the game, which was determined in focus groups and a pre-study [270]. Working with the target users in a UCD approach resulted in a VR exergame that shows a high usability for this user group, which we suppose is comparable with traditional exercise videos for older adults.

Regarding hypothesis H_2, we suppose that the novelty effects of the VR application will be balanced out by the familiarity with the traditional exercises provided in the video. The personal decision of preference will depend on the participants, their physical fitness and their openness for new technology. Since our participant group will be taken from the general population of a senior living facility, we expect these factors to be balanced out.

H_3 was established because the video exercises are more familiar and easier to access. While most older adults are familiar with (exercise) videos and the usage of computers and the internet [66], VR systems are still an unfamiliar concept and most head-mounted displays come with usability and technical issues that prevent the older generation from using them without help (e.g., setting up the hardware and starting applications).

To summarize, it is still unclear how older adults will compare a traditional exercise video to an immersive VR exergame, but the latter provides several advantages which will be discussed in the scope of this chapter.

6.2 Methods

In this section we will describe both, the VR exergame, and the exercise video, as well as the user study.

6.2.1 VR Exergame

Maestro Game was developed in a User-Centered Design approach, with several focus groups, interviews with older people and experts, and prototyping sessions of previous versions of this game, as described in our previous paper [270].

The exergame was implemented using the Unity game engine (version 2019.4) [327] with the High Definition Rendering Pipeline and Visual Effects Graph with Particle Strips. The exergame was played on a computer using an i9-9900K CPU and a GeForce RTX2080 Ti graphics card. We used the Valve Index VR headset, two controllers to act as batons, and 2 base stations. The player's in-game view was displayed on a 65 inch 4k display for the spectators.

In *Maestro Game*, the players are located in a 3D concert hall. They are playing the role of a conductor that has to conduct a band of three musicians by following a virtual

3D path in front of them with a baton. The virtual path is made up of three essential parts, as can be seen in Figure 6.2:

- A disco ball which the player is supposed to follow with the tip of the baton.
- A tunnel with staff lines extending before and after the disco ball to help the player anticipate the target's path.
- Music notes that regularly spawn along the tunnel to calculate the player's accuracy when the disco ball hits them.

In order to accommodate for different body types and mobility [119] and to calibrate a fitting path for the player, a calibration is performed at the start of the game, where players have to move their arms as far to the front, sides and up as possible. With these measurements, the shape and position of the path is determined. For each arm, an ellipsoid with the shoulder position as the center and the maximum arm reach as outer points is created (See Figure 6.3, light green). A cuboid shape that begins in front of the shoulder position and spans to the outer edge of the extended arm is then added (red cube). The length towards the other side and in the downward direction can be configured with values between 0 (ends at the shoulder position) and 1 (extends as far to the other side as possible), to enable different difficulties. All points within the intersection of the light green circle and the red cube represent potential positions the disco ball can occupy. The ball's trajectory is then dynamically calculated using a Bézier curve.

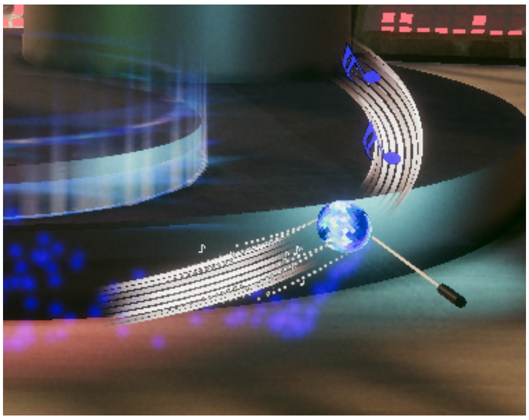


Figure 6.2: The disco ball that the player has to follow on a virtual path that resembles staff lines. The baton emits blue sparkles to highlight that the right hand has to be used. Notes are used as measurement points for accuracy.

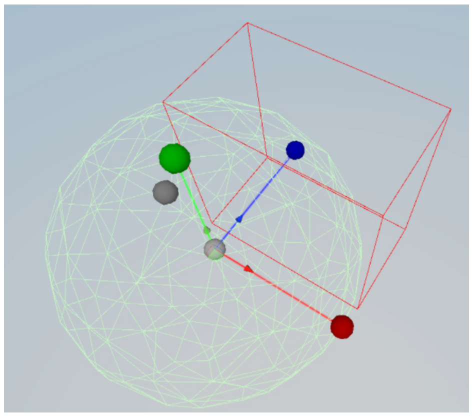


Figure 6.3: Calibration and curve creation. Light grey: shoulder position of the player. Light green lines represent the maximum reach of the player's arm, and the red cuboid shape shows where points for a Bézier curve can be chosen.

In accordance with physiotherapists, useful but comfortable movements with occasional more difficult spikes are procedurally generated. This approach does not put too much strain on the player's muscles, but also includes some challenges with the prospect of improvements.

The players are only conducting with one hand (left or right) at the same time, but change hands after 30 seconds of successfully conducting the musicians. This way, they can fully focus on following the path correctly, reducing the physical and cognitive complexity and their frustration. Oral instructions are given for a hand change, and the color of the disco ball and baton sparkles changes (blue for the right hand and red for the left hand). This dual channel output aimed to (i) compensate for impairments of one of the two senses, e.g., the visual or auditory sense, and (ii) provide a constant visual reminder of the current game rule.

The usage of music has been implemented in several other exergames, such as Beat-Saber [96] or OhShape [180], and has been shown to increase motivation and performance, and invoke old memories [190, 338]. The players can choose one out of eleven songs, which are a variety of newer and older pop, classic and rock songs. In prototyping sessions and interviews, we asked the participants for their favorite music and tried to include it in the game. The songs are split into the individual music instruments using *Spleeter* [122], so that they can be used individually for the three musicians. Depending on the instrument, the beat changes, resulting in different movement rhythms of the disco ball.

The active musician is always located at the front of the stage, with the two other musicians behind them (See Figure 6.1). When the player successfully follows the disco ball, the volume of this musician gets louder, depending on the distance of the baton to the center of the disco ball. When the maximum volume is reached, the stage starts to rotate and the next musician moves to the front. While the musicians are not actively conducted, their volume decreases again. The goal is to increase the volume of all musicians to its maximum, resulting in the "full" version of the song, which can then be heard as a reward. If the player stops following the path or if the baton deviates too far from the disco ball, the musician stops and a distorted sound of their instrument is played. When a song is completed until the end, confetti starts to fall from the ceiling, fire appears at the sides of the stage and a long applause is played. A new song can then be chosen.

The player's movements, personal settings and their calibration are tracked, enabling a later evaluation and the possibility to compare their progress.

6.2.2 Exercise Video

We chose to compare the VR exergame with a 2D gymnastics video because it creates external validity for our research as many people did their workouts with exercise videos due to the COVID-19 pandemic [65, 153]. The exercise video was recorded using an iPhone 7 Plus. It was displayed on a 65 inch 4k display and two Valve controllers were used for tracking the hand movements of the participant. The video featured exercises adapted from different sources [42, 335, 376, 378], for example side arm raises, shoulder circles or torso rotations. The details can be found in Appendix B.

To replicate a realistic as well as physically benefiting experience, the video and the movements were inspired by several YouTube videos [87, 88, 376, 378]. The senior living home employees, including physiotherapists and caregivers, reviewed and agreed to the video and exercises.

The video was integrated in a Unity application to enable tracking, data collection and individual music selection. Even though the video was prerecorded, the participants were able to freely choose three out of 12 possible songs of different genres as background music. The program automatically tunes down the volume of the songs while oral instructions are presented, so they are clearly audible.

6.2.3 Study Procedure

The goal of this user study was to compare a traditionally recorded 2D gymnastics video with a VR exergame regarding enjoyment, attention allocation, perceived workload, and preference. All participants were older adults still living independently in their own apartments, which are part of a local senior living facility that offers serviced homes. Our project works with various ethics and privacy experts, as well as the local senior living facility, who support us during the project. We contacted our local ethics department and received an acknowledgement for the study. Furthermore, an ethics workshop was held before the start of the study to discuss ethical, as well as security and privacy concerns for all of our studies, and advice from this workshop was integrated.

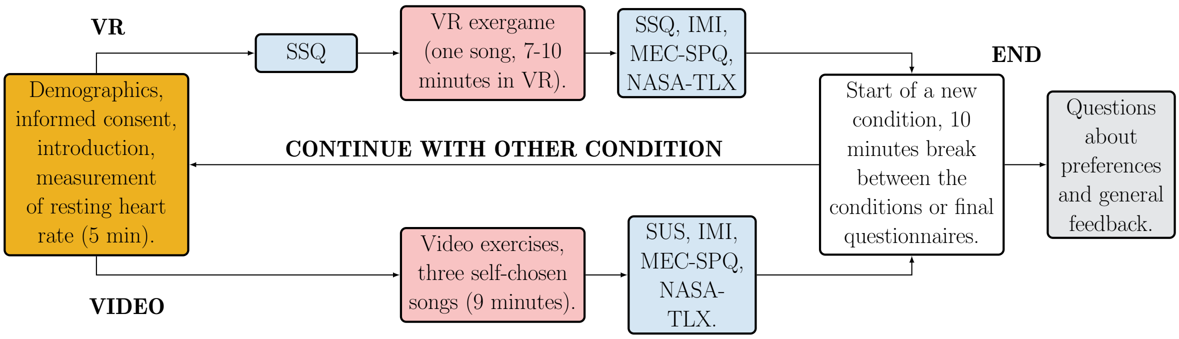


Figure 6.4: Study procedure to compare a VR exergame to video-based training.

Pre-Study

Four people (P_1 - P_4, all female, average age: 81.75 ± 4.19 years) participated in a pre-study to determine a suitable length and difficulty of the game. We planned to play the game in a standing position, but two of the (first-time VR) users asked to sit down after putting on the HMD because they felt insecure, and sitting exergames can also provide a high usefulness to players [371], so we chose this setup for all following players.

Furthermore, we originally planned to play three songs in a row, as it was suggested by our players in the prototyping sessions. However, the players in the pre-study showed signs of exhaustion and cybersickness during the second and third song (increased breathing rate, increased heart rate and dizziness), and also reported those feelings after the study. This is also consistent with literature, which indicates higher cybersickness scores for inexperienced VR users [319]. One more experienced user said that three songs were suitable for her, but since we included experienced and less experienced VR users, and older adults with varying physical fitness, we decided to choose a safe setup and only play one song. After all, this study was not meant to test efficiency of the game regarding physical training effects, but to compare the feasibility and preference of a traditional 2D gymnastics video to a VR exergame. As determined in prior prototyping sessions and in the pre-study, we also adjusted the speed and difficulty of the game to make it safe and enjoyable for all older adults, regardless of physical fitness.

Comparison Study

The procedure for the study can be seen in Figure 6.4. We evaluated the two exercise programs in a within-subject design, meaning that each participant tried out both conditions in a counterbalanced way. A consent form was given to the participants a few days in advance so they could read it, sign it and write down questions. Upon arrival, the older adults filled out a demographics questionnaire and they were allowed to choose one song for the game, and three songs for the workout video. Then we gave them the heart rate sensor and measured their resting heart rate for five minutes. Afterwards, their first condition started.

In case of the exergame, the participants filled out a before measure of the SSQ. Then they watched a short demonstration video of the experimenter playing the game, so they knew what to expect and to reduce their potential fear [127]. They then sat down on a chair in the middle of the tracking space, put on the HMD and adjusted it to fit onto their head. Afterwards, a calibration sequence was started where a virtual agent asked them to move their arms as far forward, up, and sideways as possible, followed by a tutorial. The tutorial explained how to follow the disco ball with the baton, and that the task was to move in a certain, changing rhythm. A hand change was also practised. Afterwards, the song they had chosen was started, and they conducted it until the end. When the song finished, they had some more time to look around in the virtual environment and then took the HMD off again. Then they filled out another SSQ, followed by the interest/enjoyment sub-scale of IMI, the attention allocation part of MEC-SPQ, and NASA-TLX. After a 10-minute break, the next condition was started.

In the case of the video-condition, the participants again took a seat on the chair in the middle of the room. Two Valve controllers were given to them to track their movements and the video was started. The music that the participant had chosen for the video was played in the background to make the exercises more comparable. After the video, the participants again filled out IMI, MEC-SPQ and NASA-TLX.

After the last condition, they were asked which sports program they preferred and why, which sports program they would prefer to do in long-term use and why, and lastly, were able to provide additional feedback about the game.

The study took around one hour, with 7-10 minutes spent in VR, depending on the chosen song length, and a duration of 9 minutes for the gymnastics video, as suggested in [372].

Participants

In the actual study, 25 older adults took part (P_5 - P_{29}, 3 male, 22 female, average age: 81.24 ± 4.97 years). Five older adults already had experience with VR and participated in some of our prior studies and prototyping sessions. Three participants reported never doing any sports, 15 were doing sports at least once a day, and seven exercised at least once a week. Activities mainly included yoga, gymnastics, walking, or biking. Regarding gaming, seven people said that they did not play any games (computer or board games), four people played every day, 12 people at least once a week and two people at least once a month.

6.2.4 Measurements

Several measurements were used to compare the two conditions. To evaluate the well-being of the participants after playing the VR-exergame, they filled out the SSQ [154] before and after the VR experience. In order to assess their enjoyment, the participant answered three questions on a scale from 1 to 5 ("The task was fun.", "The task was interesting", "The task was entertaining."), adapted from the interest/enjoyment sub-scale of the Intrinsic Motivation Inventory (IMI) [280, 362]. Furthermore, the sub-scale Attention Allocation from the 4-item scale of MEC Spatial Presence Questionnaire (MEC-SPQ) was presented [352]. We chose this questionnaire because of its ability to compare different media, e.g., video and VR, as opposed to well-known presence questionnaires that are used to distinguish experiences in the same type of environment [340]. Finally, the NASA-TLX questionnaire [113] was used for an evaluation of the perceived workload.

During both conditions and during a resting period, we measured the heart rate of the participants to have an additional indicator of exhaustion and physical stress, and to better compare the exertion of both programs. This was done using the Polar OH1 heart rate sensor [256], which the participants wore around their arm. Furthermore, we aimed to have a physiological measure to compare to the subjective measurements given in the NASA-TLX questionnaire.

6.3 Results

In this section, we will report the results of the study. We will report both, statistical results from Frequentist statistics ($\alpha = 0.05$) and from the Bayesian approach. Frequentist statistics is still the standard in HCI research, but falls short in evaluating in favor of the null hypothesis or depicting how much more likely one condition is over another [144, 152]. Evaluation was done using JASP [329]. The descriptive values of the questionnaires can be seen Table 6.1.

Conditions	Pre-SSQ	Post-SSQ	IMI	MEC-SPQ	NASA-TLX	Heart Rate
VR Exergame	8.98 ± 11.48	8.08 ± 11.46	4.57 ± 0.63	4.44 ± 0.72	19.77 ± 11.91	76.77 ± 9.6
Exercise Video	—	—	4.6 ± 0.68	4.63 ± 0.56	17.23 ± 10.94	81.96 ± 11.33

Table 6.1: The descriptive values of the questionnaire measures and heart rate for both conditions, mean and SD.

6.3.1 Enjoyment

Both programs, the video and the VR exergame, received similar results regarding enjoyment. No significant differences could be found using Frequentist statistics ($t(24) = -0.165$, $p = 0.871$). A Bayesian Paired Samples T-Test showed that the null hypothesis ($\text{enjoyment}_{game} = \text{enjoyment}_{video}$) is 4.68 times more likely than the alternative hypothesis. Using Jeffrey's criterion, this is positive evidence in favor of the null hypothesis [152, 262].

6.3.2 Attention Allocation

For Attention Allocation, there was no significant difference between conditions ($p = 0.122$, $t(24) = -1.605$). A Bayesian t-test shows weak evidence in favor of the null hypothesis ($\text{attention}_{game} = \text{attention}_{video}$), with a Bayes factor of 1.54.

6.3.3 Perceived Workload

The exergame was perceived as cognitively and physically more exhausting than the video. In detail, the video received a NASA-TLX score of 22 ± 22.31 out of 100 points for cognitive exhaustion (median: 10), while the exergame received 27 ± 21.21 points (median: 25). For physical exhaustion, the video was rated with 19.6 ± 20.41 points (median: 10), and the exergame with 20.8 ± 20.4 points (median: 15). Additionally, the perceived frustration was lower during the VR exergame (11.8 ± 8.65, median: 10) than during the video (15.2 ± 20.28,median: 10). While most older adults rated both conditions relatively low, there are some higher ratings of 95 out of 100 points in the cognitive or physical demand. Due to different cognitive states and physical fitness of older adults, and also different perceptions of difficulty, these ratings should also be considered for future implementations.

No significant differences between the two conditions could be found ($t(24) = 1.237$, $p = 0.228$). The Bayesian Paired Samples T-Test shows a Bayes factor of 2.398 in favor of the null hypothesis ($\text{workload}_{game} = \text{workload}_{video}$). A positive correlation could be found between physical and cognitive workload in the exergame ($r = 0.832, p < 0.001, BF_{10} = 65044$), but not for the video ($r = 0.066, p = 0.754, BF_{10} = 0.26$).

6.3.4 Simulator Sickness

For the exergame, we asked the participants to fill out the SSQ before and after the experience. A non-significant difference ($t(24) = 0.321$, $p = 0.75$) indicates that the game did not cause any cybersickness. A Bayesian t-test resulted in a Bayes factor of 4.52 in favor of the null hypothesis ($SSQ_{before} = SSQ_{after}$). We also report the descriptive values of the subscales before the study: Nausea: 6.4872 ± 10.92; Oculomotor 7.8832 ± 12.47; Disorientation 9.4656 ± 13.76, and afterwards: Nausea: 2.2896 ± 5.7, $p = 0.0442$; Oculomotor 8.4896 ± 11.64, $p = 0.7509$; Disorientation 9.4656 ± 19.69 , $p = 0.8359$.

No correlations with age, gaming experience or the number of times the older adults exercised were found.

6.3.5 Heart Rate Data

The descriptive values of heart rate measures can be seen in Table 6.1 for the two conditions. During the resting period, the average heart rate was 80.34 ± 9.22. It was higher during the video than during the exergame ($t(24) = -4.4$, $p < 0.001$, $BF_{10} = 257.99$ in favor of the alternative hypothesis $HR_{Video} \neq HR_{Game}$), with a mean difference compared to the resting heart rate of -3.75 ± 5.11 during the game and 1.62 ± 6.63 during the video. A positive correlation ($r = 0.436, p = 0.029, BF_{10} = 2.339$) was found between the heart rate during the video and the perceived performance measure from NASA-TLX, indicating that a higher heart rate was associated with a better performance. No correlation was found between the heart rate and the perceived physical exhaustion for the video ($r = -0.259, p = 0.212, BF_{10} = 0.52$) or the exergame ($r = -0.246, p = 0.236, BF_{10} = 0.482$), or between the other subscales of NASA-TLX.

6.3.6 Movement Data

During both programs, the movement of the participants was tracked with the two Valve Controllers. The mean distance traveled by the hands of the participants was $187.17 \pm 89.71m$ during the video, and $24.22 \pm 4.32m$ during the game, which is a significant difference ($t(24) = 9.219, p < 0.001, BF_{10} = 5002000$). These large differences are mainly due to the longer length of the exercise video, but also the more diverse and bigger movements performed in this condition. Additionally, some tracking inaccuracies due to fast movements occurred. The right hand, which was the dominant hand for most participants, covered more distance than the left one, with 97.78 ± 46.47 m for the right hand and 89.36 ± 46.67 m for the left hand in the video, and 14.5 ± 9.71 m for the right, and 9.71 ± 3.04 m for the left hand in the exergame condition.

The plots in Figure 6.5 show the movements of the hands in polar coordinates for both scenarios, the exergame and the video, recorded for P_{28}. Since the movement of all participants depended on the individual range of motion, a pooled plot would disguise

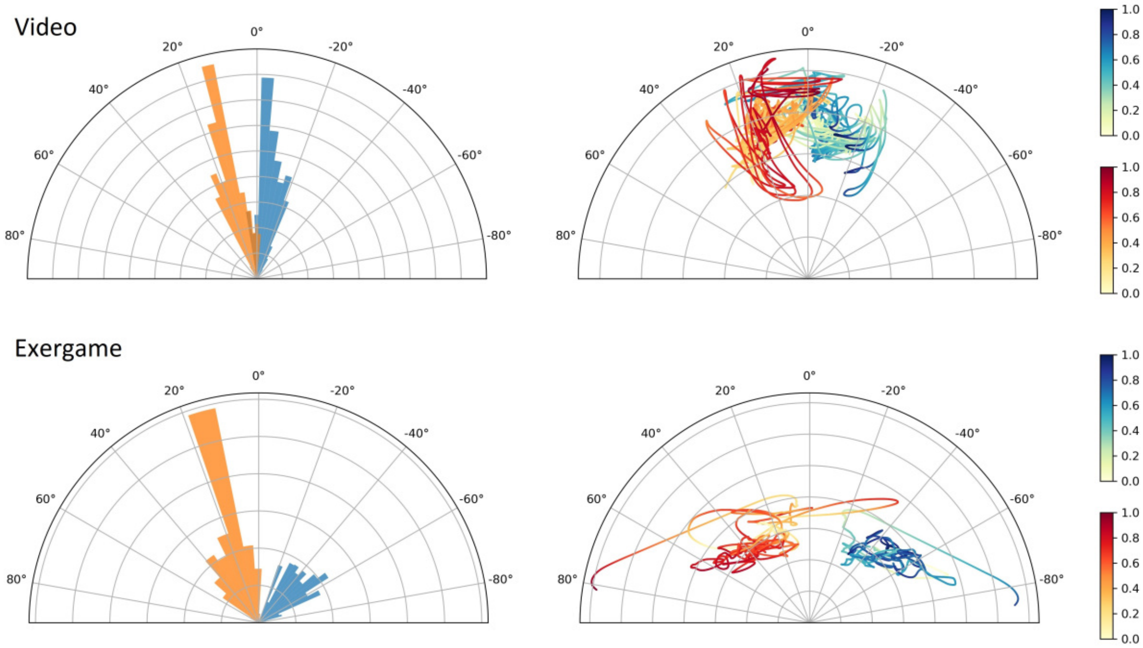


Figure 6.5: Polar plots of the movement of P_{28} for the video (upper two plots) and for the game (lower two plots). The first plot always shows the histogram of the hand movement, while the second one displays the hand movement trajectory with intensity measure.

the results, which is why we will explain the movements on this participant. The other participant's plots show similar results.

The left figure is a histogram that shows how much movement occurred in a specific area. The right figure displays the trajectories of the left and right hand. Additionally, the intensity of the movement can be seen. The darker the color, the more intense the movement. Intensity is measured as a function of movement radius (stretching of the arm) and lifting of the arm. It is considered that lifting or stretching the arm is more intense than flexing the arm. The 0° direction is considered to be the center of the body.

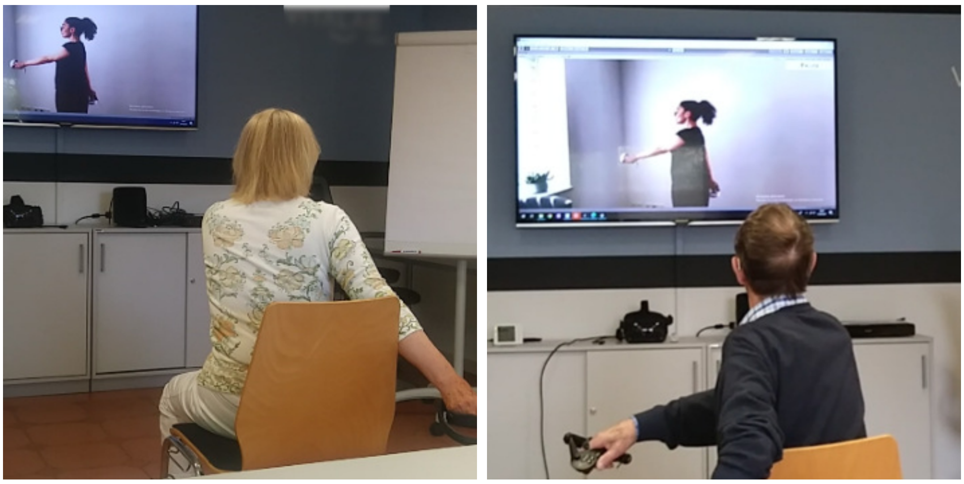
When comparing the video scenario with the game scenario, it can be seen that the video seems to be more intense than the game scenario. This can be explained by the high radius values. But when looking at the histogram one sees that the video scenario leads to a narrowed range of motion; the movement takes place in a range of 0° and 30°. In comparison, the game shows a much wider range of motion between 0° and 60°.

6.3.7 Qualitative Feedback

On the question which sports program the participants preferred, 13 people chose the video and 9 people chose the exergame. Three people ticked the boxes of both options, because they could not decide on one. Several people mentioned that they did not understand the question, since they did not perceive the exergame as sports, but just as a game. One person who preferred the *video* mentioned that they were already familiar with this type of exercise (P_{11}). Additionally, the exercises trained the whole upper body, as opposed to the exergame where mainly the arms and torso were used (P_{14}). It therefore felt more like a workout with a higher physical demand (P_7, P_{17}, P_{27}, P_{28}). Additionally, they noted that the exergame needed some getting used to (P_{15}), as opposed to the more realistic, face-to-face video (P_7, P_8). The participants that preferred the *exergame* commented that it was more fun (P_5, P_{21}) and more interesting (P_6). They also commented that compared to the exergame, the video felt more monotonous (P_9). It also required more cognitive attention to follow the path correctly (P_6, P_{13}), which was pointed out positively. Additionally they noted that they liked the virtual environment (P_9, P_{18}, P_{21}), and the music fit well to the movements (P_{22}).

For long-term use, 16 participants would prefer the video, and 7 would prefer the game. This time, two people chose both options. The main arguments for the *video* were that the exercises were known (P_{11}, P_{22}, P_{23}), more intense (P_{16}, P_{21}) and easier (P_{19}). Additionally, they showed more variety (P_{27}). For the *exergame* in long-term use, the participants again mentioned that it was more fun (P_9, P_{13}), that they liked the movements (P_{10}, P_{12}) and that it required more concentration (P_6, P_{13}). One participant wrote that it was "self-explanatory" why this was the preferred option (P_{18}). P_{22}, who favored the exergame, mentioned that while he would choose the video for long-term use, the exergame could serve as a refreshing change or as a relaxing activity.

6.4 Discussion



(a) (b)

Figure 6.6: (a) A participant doing the exercises in a reversed way compared to the video. (b) A participant looking forwards while doing torso rotations.

Our goal was to assess if a VR exergame can be an alternative for traditional exercise videos. Especially during the COVID-19 pandemic, where group and personal sport experiences were limited, many people used exercise videos to remain fit at home.

As expected in H_1, almost all results from our qualitative and quantitative evaluation remained very similar, but the results obtained in the subcategory of MEC-SPQ surprised us. Here, although not significant, the exercise video obtained a higher score for attention allocation. VR is known to provide a large feeling of presence [311], which includes turning the user's attention to the medium. The video on the other hand also proved to capture the users' attention, and even received slightly higher scores than the VR exergame. It seems like the older adults, who grew up with a TV and mostly use it everyday, maintain the ability to be captivated by it, as opposed to younger people who are already familiar with VR and value its ubiquity. This is similar to what Xu et al. [373] found in their study with young and middle-aged people, where younger players were more immersed in VR than middle-aged ones.

Prior research has shown that VR provides a higher feeling of flow and more enjoyment than a TV display [372]. This effect could not be confirmed in our study, but it has to be noted that our two training applications showed many differences in the task, environment and movements. Future studies should compare an exercise video to an exergame that also features similar movements.

Hypotheses H_2 and H_3 could also be confirmed, with around 40 % (9) of the participants preferring the exergame over the exercise video, and around two thirds (16) of participants preferring the exercise video for long-term use. Movement plots support this, with the video showing more intense movements that are more suitable for long-term training. On the other side, movement precision and range of motion in the video condition are limited due to the missing feedback and encouragement for larger movements.

Advantages of the exercise video are that there are many videos available online, which foster different topics and can be watched independently at home. A disadvantage is that there is no intuitive way to ensure that the exercises are performed correctly. This could also be seen in our movement analysis, where participants often performed the movements in a different speed, or with the reversed arm (See Figure 6.6 left). For this video, those errors had no impact on the exercise accuracy, but especially for more difficult movements or for someone with mobility impairments, the correct execution of the movements is important. Additionally, the movements in the video either required oral explanations, or the participants constantly looked at the TV to follow the movements, even if they were supposed to rotate their torso to a different direction (See Figure 6.6 right). For the oral explanations to be audible, the volume of the music had to be decreased, which could potentially break the flow of the experience.

For the VR exergame, movement could be controlled and immediate feedback could be given. If the players did not perform the correct movements, the musicians stopped playing. On the other hand, if the player did well, there was applause and confetti, and the speed of the movements increased. Therefore, the game was able to dynamically adjust itself to the abilities of the participant and encouraged them when correct movements were performed. Furthermore, because of the stereoscopic 360° 3D representation of the virtual environment, the participants' ability to replicate spatial movements in their peripersonal space was encouraged. When seeing the required movements in 2D in the video condition, spatial information such as depth of movement on the Z-axis can get lost, which is important to create cognitive representations of space [83]. Additionally, prior studies have shown that distance perception in peripersonal space in VR shows smaller errors than in extrapersonal space [13], which, due to the distance of the TV, the exercise video falls into.

The movements in this game were synchronized with the music, amplifying the feeling of flow and even encouraging some players to tap their feet or to sing along. On the other hand, the movements could become monotonous due to the limited variety of body parts engaged. The initially chosen difficulty was too easy for some participants, but our pre-study showed that the originally planned setup was too difficult for some others. Therefore, the correct starting difficulty and the number of songs that can be played should be balanced out and decided by the older adults themselves, if the game is played outside of the study context. This was also reflected in the heart rate measurements, where the heart rate even decreased compared to the resting heart rate. This might be because many participants were excited to try VR for the first time, and did not know what to expect. But they started to relax when they were wearing the HMD and when they followed the relatively slow motions of the exergame, similar to one participant in our previous study with people with dementia (See Chapter 5). Thus, our findings are in line with prior work [151] that recommends testing VR games with participants before taking baseline physiological measurements and conducting the actual experiment.

Interestingly, our findings point out that while older adults rated the VR exergame and video based exercise similar in terms of perceived physical workload, the heart rate and movement values for the video-based exercise were significantly higher than for the VR exergame. We speculate that the similar rating of perceived physical workload for the VR game might be due to the unfamiliarity with the given VR exergame mechanics and VR technology, although long-term studies are needed to verify this claim. Contrarily, prior research [373] was not able to find an influence of display type on perceived workload. Furthermore, the cognitive workload during the game might have influenced their perception of the physical workload, even though this is also not supported by literature [68]. Both cited studies were tested with younger adults, yielding the need for studies with an older user group. We conclude that novelty effects of VR with older adults should be considered in future work, as it can be overwhelming or exciting and can thus affect the overall experience.

To address the limitations of the study, the exercises performed in the video are not exactly comparable with those in the VR exergame. While the exergame mainly focused on arm movements and some torso rotations, the video provided more diverse exercises with significantly more movement. Yet, we note that the exercise movements in the video represent a real-world based scenario, and the game could be adjusted to feature more and larger movements. Additionally, the question, which *sports program* the participants preferred resulted in surprise, because the VR exergame was not perceived as sports, but rather as a game, as mentioned by several participants. This might have distorted the results of this question towards the participants choosing the video. Another point which was apparent in the questionnaire results was a ceiling effect. Both programs received very high scores, with numerous questions receiving the maximum number of points. So even if the second program the participants experienced was in their eyes "better", there was no way to correct their prior evaluation of the first program. Lastly, we only had three male participants in our study, so gender differences cannot be accounted for. On the other side, we would like to empathize that because of the longer life expectancy of females, they are the main population of senior living homes and therefore more likely to use the systems.

To discuss safety issues, it is important to follow the guidelines suggested by the HMD manufacturers. Additionally, the frailty of this user group has to be considered. Choosing a standing setup should therefore only be proposed to players that feel secure and have used VR before. This should always be their own decision, with options to play the exergames in sitting and standing positions [148], and to switch between the two whenever necessary. Future exergames should adapt to this dynamically. When employed at a senior living home, the tracking space should be secured, e.g., with the help and involvement of technicians [150].

Our findings suggest that the VR exergame is not able to fully replace the traditional workout. But with the demographic change, caregivers need to be relieved, and additional offers should be available for independent training of the older adults. Our study showed that virtual humans or virtual content was largely accepted by the older adults, which is also supported by literature [150, 217]. With new generations, this trend of accepting new technologies is likely to continue. In combination with the dynamically changing difficulty, and the ability to receive immediate feedback about movements and track the older adults' progress, VR exergames show to be a suitable alternative.

6.5 Conclusion

We have developed and evaluated a VR exergame that features rhythmic movements in 3D space, and compared this to a traditional 2D gymnastics video. Both programs received similar scores regarding enjoyment, workload and attention, and neither a significant difference using Frequentist statistics, nor a high likelihood of differences using Bayesian statistics could be found. While the video was preferred by slightly more people (13), the game also received 9 votes for preference (3 people remained undecided). For long-term use, more people (16 vs. 7) prefer the exercise video.

In a qualitative evaluation, the participants commented positively on the fun and novelty of the VR exergame, while some also found the VR exercises too repetitive. Regarding the exercise video, the older adults appreciated the known exercises and their variety, as well as the higher physical demand.

The different nature and exertion of the exercises in both programs is one of the limitations in this study. Furthermore, even though the perceived physical workload in both programs was the same, some participants did not perceive the game as exercising. Future studies should develop and evaluate VR games that feature similar movements as traditional workout videos. Finally, long-term studies should be conducted to test VR and video exercises without the potentially confounding novelty effect.

In summary, our results showed that although exercising with VR exergames, or using VR technology in general, is not a familiar concept for older adults, VR exergames are promising and can be an alternative option to exercise for this user group. We recommend using VR exergames for people who like experiencing new emerging technologies or alternatives and need or want an adjustable, dynamic fitness application. Furthermore, we recommend to dynamically set the difficulty of the game to match the skills of the users.

Chapter 7

Investigating Dynamic-Difficulty Adjustments With Older and Younger Adults

Figure 7.1: The range of motion calibration (left), a close-up of a soap bubble containing a butterfly (middle), and the view of the canoe and soap bubbles during one of the levels (right).

Our previous user studies (See Chapter 5 and Chapter 6) have shown the potential of exergames for psychological, physical and cognitive well-being of older adults, but, especially for the second user study, the difficulty of the exergames was not matched to the skill of the participants, with many participants wishing for a more physically demanding task. With this third user study, we addressed this problem by employing dynamic difficulty adjustments. We evaluated this with younger and older adults and matched the difficulty of the game to the skill of the player. Further, we compared the results and opinions of younger and older adults, aiming to identify transferabilities and differences between the user groups. This chapter is based on the following publication:

> **Kruse, L.**, Karaosmanoglu, S., Rings, S., and Steinicke, F. (2022). Evaluating difficulty adjustments in a VR exergame for younger and older adults: Transferabilities and differences. In Proceedings of the ACM Symposium on Spatial User Interaction (pp. 1-11). DOI: `https://doi.org/10.1145/3565970.3567684`

7.1 Introduction

While older adults have many similarities in terms of age-related fitness, each person has different needs with varying physical and cognitive abilities. Similarly, younger adults are a whole different user group, with possibly more experience with exergames and VR technology in general as well as differing requirements regarding cognitive-physical training activities. Therefore, to design and evaluate VR exergames, it is important to consider the needs of specific user groups to develop applications that are challenging, yet suitable and engaging for them.

Evaluating applications with special user groups can be challenging due to lack of access or availability. This is not only limited to older adults, but can also include other user groups with certain inclusion criteria, such as specific diagnoses or demographics. In particular, access to older people has been greatly limited during the pandemic in order to minimize risks for infection. However, applications still needed to be verified, and, thus, some researchers decided to test them with very small sample sizes or different user groups [167, 230]. For us, the unavailability of larger groups of older adults required us to test VR exergames with younger adults. While ideally developing and testing should be implemented with the actual user groups, transferability of the results between different user groups still needs to be explored. We expect to find both, similarities and differences, between these heterogeneous groups from which we can derive design guidelines for researchers and developers.

In this study, we examine the possibility of using an individual calibration of a VR exergame to determine a suitable starting difficulty. Especially for first-time players, it is important to maintain a balance - avoiding overwhelming them while encouraging their commitment and adherence. When using default values, it may be difficult to pick the correct difficulty from the beginning, deteriorating the experience, as it happened in our previous study (See Chapter 6).

Dynamic difficulty adjustment (DDA) is one way to automatically adjusts a game's features, e.g., speed or tasks, according to the user's skill [392]. Andrade et al. [11] emphasizes the importance of adjusting the game in an unobtrusive way, and of adapting the game difficulty to players' initial skill levels as quickly as possible. Therefore, we focus on the first step — the calibration of the initial difficulty and the resulting, general difficulty of the game which is shaped by this initial calibration.

DDA has been used and evaluated in several exergames, e.g., for people with Parkinson's disease [314] or in collaborative experiences [44]. Smeddinck et al. [314] developed an exergame where stars have to be collected using Kinect motion tracking. The difficulty of the application was adjusted on three different parameters: speed, accuracy and range of motion. Depending on the performance of the user, adjustments were applied in fixed, pre-determined steps once performance thresholds were crossed and their method resulted in a balanced game. Cantwell et al. [44] developed a collaborative walking game with individual difficulties according to the users' skill levels, facilitating intergenerational play and accounting for individual differences. Instead of just changing the speed of the game, the authors propose several difficulty adjustments, such as additional obstacles, or requiring larger movements, which are added when the users perform well. Pezzera et al. [250] propose to use a mixed-design; to include the history of the participant, their real-time performance as well as their emotional state to control the difficulty, and their approach showed promising results. It was also shown that automatically adjusting a game led to better player experience than letting the players do it themselves [12]. Based on prior works, and since we had no history of the participants available [250], we decided on using a simple DDA approach; we individually calibrated the initial difficulty level for each player, and then the game dynamically adjusted itself based on their performance, specifically the number of correctly collected bubbles.

To understand the effects of (i) different calibrated start difficulties and (ii) differences between two distinct user groups, we conducted two user studies with nine older adults ($M = 75.67, SD = 8.31$ years) and 30 younger adults ($M = 24.77, SD = 4.1$ years) who played a cognitive-physical VR exergame. The game requires players to sit in a canoe running down the river. In front of them, soap bubbles are spawned in 3D space, which have to be collected with the user's hand according to different cognitive tasks, following the dual-task paradigm [324]. The specific tasks for each level can be found in Table 7.1. The game was played for four rounds; a *Calibration* round as well as three rounds with (i) the exact difficulty calibrated in the first round as a starting difficulty (*Perfect*), (ii) 50 % higher (*Faster*), or (iii) 50% lower difficulty (*Slower*). We aimed to identify the participants' preferences for these different calibrations, but also which processing speed – the number of successfully completed tasks per second – they were capable of, and where to set suitable boundaries. In the *Calibration* round, the users are asked to touch successively spawning soap bubbles, in a speed that is pleasant for them. This is done in a stationary position on a virtual beach. We purposely told them to use a pleasant speed to achieve a good trade-off between speed and accuracy. We opted for a positive and encouraging approach in the *Calibration*, where the users could determine their own speed, as opposed to steadily raising the speed until a certain threshold is crossed, as it is done during the gameplay. This way, we wanted to lower their potential uncertainties and insecurities with this new technology and provide them with a warm-up [301]. In the other three rounds, the time the users needed to touch the bubbles in the *Calibration* is used for determining the start distance between them. Here, the users are sitting in a canoe, being moved towards the bubbles. Since playing games with time pressure has been shown to increase the sense of presence [53], and high-speed dynamics positively influence enjoyment [133], only a limited time was allowed to make a decision and if not reacted and hit in time, bubbles can be missed. In these rounds, the difficulty is further adjusted according to the player's performance, but since each round only takes one minute, the general difficulty depends on the previously calibrated starting value. We compare the successfully collected bubbles as well as the calibrated difficulty, and user experience, and analyze the connection between these factors for the two different user groups. Furthermore, we reflect on the possibilities of testing a game that is made for older adults with a different user group— namely younger adults. From these results, we improved the game and finally confirmed the changes and potential for this evaluation method in a follow-up study with 11 older adults ($M = 83.82, SD = 3.06$ years).

The general research question of this study ($\mathbf{RQ}_3$) was, how the difficulty of an exergame impacts user experience and performance, and which insights can be derived from evaluating this with different user groups (i.e., younger and older adults). To look at these aspects in detail, it is divided into three research questions:

- $\mathbf{RQ}_{3.1}$: How does the starting difficulty affect the users in terms of user experience (in particular, enjoyment, perceived workload, cybersickness, preference) and game performance?
- $\mathbf{RQ}_{3.2}$: Can a VR exergame developed for older adults be adjusted for younger users and still show high engagement and fun?
- $\mathbf{RQ}_{3.3}$: Which results can be derived from testing a VR exergame for older adults with a different user group (in this case with a younger adults)?

7.2 Methods

In this section, the application, the study as well as the used measurement methods are presented.

7.2.1 Application

For this study, we adjusted a game that was already tested with older adults [149]. It was developed with the human-centered design approach [320], involving both older adults as well as physiotherapists. The feedback from them was used to improve the game. The *Canoe Game* was developed in the game engine Unity3d version 2019.4.18 [327] with the Universal Render Pipeline and the Oculus Utilities SDK. It was built for the Meta Quest 1 and 2 [86] and offers both options: hand tracking and the use of controllers. For this study, we used controllers because of the more accurate tracking. The game includes simple, natural game concepts already familiar to the older adults and shows a high usability for this user group, as suggested in prior literature [167, 388]. Nevertheless, the game flow and tasks can be adjusted by changing the speed of the game to also be a fun and challenging experience for younger adults.

First, the ROM of the player is determined in 3D space. Six colored balloons are spawned above, in front of and next to the player, which have to be pushed away as far as possible. The two balloons in front of the player were added in this iteration of the game to determine the forward ROM of the player. The position of the balloons is then used to create a polygon shape (See Figure 7.1, left). Inside this shape, collectible objects can be spawned during the game. After all balloons have been touched, the environment is changed to a river with mountains to the sides, adapted from the GitHub project Boat Attack [209, 258].

The game features different levels of different types, combining training categories such as working memory, attention and processing speed [126], with some elements adapted from MoCA [232]. While there are more levels available in the original, full game, the types and levels used in this study are the following (see Table 7.1 for more details on game rules):

1. **Simple levels**: All bubbles (red and yellow), all bubbles (blue and violet), butterfly bubbles, fruit
2. **Coordination levels**: interchanging hands, right hand red - left hand yellow, right hand blue - left hand violet, two bubbles with the right hand - two bubbles with the left hand
3. **Alphanumeric levels**: lower number, only vowels, letters in word, even numbers

Level Type	Level	Description
Simple	All good (blue, violet)	Collect all bubbles. Several different colors of bubbles can be spawned.
	All good (red, yellow)	Collect all bubbles. Several different colors of bubbles can be spawned.
	Fruit	Collect all fruit. Several different fruit can be spawned.
	Butterflies	Free all butterflies from the bubbles by popping them. Only soap bubbles with butterflies inside of them spawn, which fly away once the bubble is touched.
Coordination	Right Hand Red - Left Hand Yellow	Collect all red bubbles with the right hand, and all yellow bubbles with the left hand. Red and yellow bubbles are spawned on both sides of the user's body, so cross-over movements are necessary.
	Right Hand Blue - Left Hand Violet	Collect all blue bubbles with the right hand, and all violet bubbles with the left hand. Blue and violet bubbles are spawned on both sides of the user's body, so cross-over movements are necessary.
	Interchanging Hands	Collect one bubble with the right hand, then one with the left one, and so on. Several different colors of bubbles can be spawned.
	Two Left, Two Right	Collect two bubbles with the left hand, then two with the right hand, and so on. Several different colors of bubbles can be spawned.
Alphanumerical	Lower number	Out of two bubbles presented at the same time, collect the one with the lower number.
	Even number	Out of two bubbles presented at the same time, collect the one with the even number.
	Only Vowels	Out of two bubbles presented at the same time, collect the one with the vowel.
	Letters in word "Kanu"	Out of two bubbles presented at the same time, collect the one the letter that is contained in the word "Kanu".

Table 7.1: The different levels and rules used in the game.

For our study, the game is played for four rounds, with three levels each. Each level is played for 60 seconds.

In addition to the original game, we adjusted it to add a difficulty calibration before the game starts. The goal of this is to achieve a more personalized game experience that fits the cognitive and physical abilities of each player.

In this Calibration round, the players are located on a small island. In front of them, within their ROM, one or two bubbles are spawned, according to the level (alphanumerical levels always spawn two bubbles that the user has to choose from). Once a bubble is touched, it disappears and the next one spawns. This game mechanic, in combination with hand redirection, was also tested in a user study with younger and older adults [114] and showed the potential to combine hand redirection with VR exergames. The game counts how many bubbles were popped correctly within one minute. This difficulty is then chosen as a start difficulty for the next three rounds.

In the rounds after the Calibration, the canoe moves along the river with a constant speed of 1m per second. No steering or acceleration of the canoe is possible. This was chosen to avoid the negative effects of cybersickness as much as possible. The general task of the game is to collect bubbles, which can be done by simply touching them with the controller. A popping sound and animation are played.

Difficulty can be adjusted by changing the distance between the spawning bubbles. The start difficulty determined in the *Calibration* is then multiplied with the scaling factors of the different conditions ($\times 1$ for perfect, $\times 0.5$ for faster, $\times 1.5$ for slower). It is then also dynamically adjusted in small-scale settings during the game, as suggested by Smeddinck et al. [314]. Every five seconds, the percentage of correctly collected bubbles compared to the total number of bubbles is evaluated. If it is above 80 %, difficulty is raised by 0.2 m. If the percentage of wrongly collected bubbles is above 20%, difficulty is lowered. The thresholds were also adapted from Smeddinck et al. [314]. This adjustment step is the same for all users and does not depend on the previously calibrated start difficulty. If a difficulty adjustment took place, the automatic adjustment is on cool-down for five seconds, since this is the time needed for the canoe to reach the bubbles with the newly created difficulty. While the game difficulty changes with a longer play time, one important factor that influences this is the starting difficulty that we determined in the *Calibration*. Since each round only takes one minute, the number of adjustments during the game play is limited. This is why the starting difficulty already needs to be close to the skill of the user, in order to avoid frustration or boredom.

Overall, an illustrative example for our DDA method can be as follows; For a person who collected 60 bubbles in the 60 seconds of the *Calibration*, the start difficulty would be one bubble per meter for the *Perfect* condition, one bubble every 0.5 meters in the *Faster* condition and one every 1.5 meters in the *Slower* condition. Then, after 5 seconds, as the first evaluation of the performance takes place, in case the participant hit all the bubbles, difficulty in the *Slower* condition will be changed to one bubble ever 1.3 meters, whereas it is one bubble every 0.8 meters in the *Perfect* condition.

At the end of the game, the players are rewarded with fireworks and confetti, similarly to the *Maestro Game*.

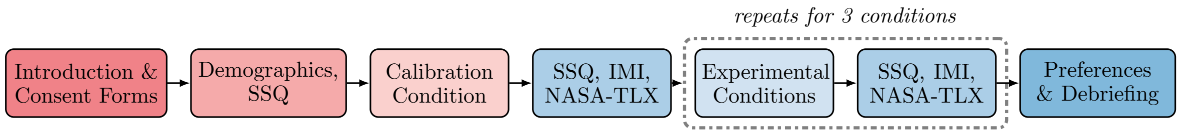


Figure 7.2: This figure shows the study procedure followed in our study.

7.2.2 Measurements

The study procedure can be seen in Figure 7.2. Before the study, demographic information was collected, including age, gender, and whether they studied or worked in the IT field. All questionnaires were administered in German.

After each condition, the participants answered the SSQ [132, 154], a short version of the IMI [281, 362]) and the NASA-TLX [113]. Furthermore, they answered the following two questions: "*How did you perceive the level of difficulty in this round?*" (on a scale from 1 (too easy) to 5 (too hard); adapted from [314]) and "*How much would you like to play another round like this?*" (on a scale from 1 (not at all) to 5 (extremely); adapted from [362]) on a 5-point Likert scale. After the last condition, they also answered some questions about their preference of game mode and game condition and gave their reasons for this choice. They were also able to make suggestions and give further feedback.

During the game, the number of correct and wrong bubbles, the level, difficulty and a movement summary of both hands were recorded.

7.2.3 Study Procedure

The study was conducted according to the guidelines of the Declaration of Helsinki, following the data protection guidelines of our university and the ethical guidelines of the German Informatics Society. An ethical assessment was made according to the regulations of the local university, which did not reveal any ethical concerns for the younger adults. The study part conducted with the older adults was reviewed and approved by the Local Ethics Commission of the Department of Informatics at Universität Hamburg with the application number 002/2022.

All participants signed consent forms before they started the study. They were free to leave the study or request deletion of their data at any time without providing a reason and without any disadvantages.

The study was run asynchronously; first, the younger and four mid-60 adults (one of them owned an HMD and assisted the others) participated remotely, then the HMDs were given to a senior living facility to run the experiment with older adults that required a little more help. A total of 27 Meta Quest (1 or 2) were given out to students for the period of two weeks and 3 participants used their own HMD. They conducted the study in their own home, without physical contact to the experimenters (See Figure 7.3 left). If required, they could contact the experimenters by email or phone. We decided on this approach because of the increasing COVID-19 cases and the resulting social distancing recommendations.

For the remaining older adults, the study was run with the help of a local senior living facility that provides homes with service for fit older adults, as well as full time care for older adults that require more help. They took care of participant acquisition and VR HMD distribution. These participants did not have as much experience with HMDs and technology, so we connected to them via Zoom to assist with putting on the HMD, starting the experience and with their questions. Furthermore, we observed them during the study in case of physical emergencies or problems. A caregiver from the senior living facility was with them during their participation (See Figure 7.3 right).

The study took around 45 minutes. The participants were in VR for around 15 minutes (four rounds, three minutes per round, ROM calibration). They took the HMD off after each round to answer questionnaires and were able to take a break as well.

Thirty students participated ($M = 24.77, SD = 4.1$ years old, 10 male, 20 female, all were studying a subject that was IT related, abbreviated as Py_{ID}). If required by their studies, they received course credit for their participation. Furthermore, 9 older adults between 67 and 85 years ($M = 75.67, SD = 8.31$ years, abbreviated as Po_{ID}) took part. Two of them were male and seven were female. None of them had worked in the IT field. All participants were able to lift both of their arms above their head. Participants were excluded from the study if they had a known history of epilepsy or an ongoing COVID-19 infection.

Before the study started, a consent form, demographics questionnaire and a pre-SSQ were filled out. Afterwards, the participants put on the HMD, and their ROM was calibrated by pushing the balloons as far away as possible.

Calibration Phase In the first round, the participants were located on a small island. Soap bubbles were spawned in front of them. Their task was to pop them by touching them with the controllers. This was played for three levels (60 seconds each) with different cognitive tasks in each level (one simple level, one coordination level and one alphanumerical level). Then, the number of bubbles that were touched in 60 seconds was calculated. This value was used to determine a starting distance between spawning bubbles in rounds 2-4.

Experimental Conditions In round 2-4 the player was sitting in a canoe that ran down a river. In front of them, there were bubbles which again had to be touched. Again, three levels were played per round. The distance between bubbles determined the difficulty. In these conditions, the difficulty was (i) exactly as calculated in the calibration (*Perfect*), (ii) 50 % lower than determined in the calibration (*Slower*), (iii) 50 % higher than in the calibration (*Faster*). Cox et al. employed a difficulty that was one third higher or lower in their study [53], but we were interested to see the effect of an even larger deviation and also account for the potential time needed to get used to VR and the task. Each round included a simple level, a coordination level and an alphanumerical level again. Starting difficulties for these three level types were calculated and used separately, because of the potentially higher cognitive demand. Every five seconds, the difficulty was evaluated and adjusted, if necessary. The order of the conditions and levels was counterbalanced. After each condition, a set of questionnaires was filled out (See subsection 7.2.2).

Figure 7.3: Left: An older adult playing the *Canoe Game* in a senior living home, and a caregiver assisting them. Right: An older participant conducting the study at home.

7.3 Results

Data was analyzed using frequentist statistics ($\alpha = 0.05$) in R, and the effect sizes were calculated with the *coin* package. Unless otherwise stated, data was not normally distributed, as tested with a Shapiro-Wilk-Test (p < 0.05).

We used Friedman's tests to evaluate differences between the four conditions (within factor; *Calibration*, *Slower*, *Perfect*, and *Faster*) for each user group. We further conducted Wilcoxon signed-rank tests for pairwise comparisons and employed a Bonferroni correction; all pairwise tests' p values are reported with this adjustment.

We note that we do not report between-participants factor comparison tests (younger vs. older adults) as our sample size, especially for the older adults, is low. Thus, we interpret our results in this regard by focusing on the tendencies in the descriptive data of both user groups. For detailed summaries of the questionnaire results, please see Table 7.2.

7.3.1 Cybersickness

Total SSQ scores for the four conditions were not very high, with a maximum score of 13.09 ± 18.18 for the younger adults in the *Faster* condition. Looking at both groups together, *Slower* reduced cybersickness by -0.67 ± 7.22, and *Perfect* by -0.38 ± 5.48. On the other hand, *Calibration*, which was always played first, increased cybersickness by 2.11 ± 5.55, while *Faster* increased it by 2.69 ± 8.96.

The total score was only significantly higher after the *Calibration* for the younger adults ($p = 0.03, V = 22.5, r = 0.363$). The other conditions did not lead to a significantly

Condition	Group	Pre SSQ	Post SSQ	Per. Workload	Int./Enj.	Competence	Pressure
Calibration	**Younger Adults**	7.11 ± 10.86	9.6 ± 11.9	24.97 ± 14.45	2.54 ± 1	2.73 ± 0.97	0.856 ± 0.82
	Older Adults	3.74 ± 5.29	4.57 ± 6.68	11.67 ± 14.19	3.48 ± 0.63	3.07 ± 0.55	0.56 ± 0.73
Slower	**Younger Adults**	9.6 ± 11.27	8.98 ± 12.33	20.06 ± 15.54	2.47 ± 0.9	3.36 ± 0.72	0.73 ± 0.95
	Older Adults	4.99 ± 6.74	4.16 ± 6.33	5.56 ± 4.77	3.22 ± 1.14	3.78 ± 0.44	0.07 ± 0.15
Perfect	**Younger Adults**	11.72 ± 14.29	11.72 ± 14.82	40.83 ± 21.15	2.62 ± 1.02	2.64 ± 1.14	1.62 ± 1.03
	Older Adults	6.65 ± 11.94	4.99 ± 3.48	15.65 ± 14.58	3.48 ± 0.65	3.3 ± 0.65	0.48 ± 0.77
Faster	**Younger Adults**	10.72 ± 14.26	13.09 ± 18.18	60.86 ± 15.84	2.83 ± 0.87	1.66 ± 0.95	2.49 ± 1.06
	Older Adults	4.16 ± 6.33	7.9 ± 11.74	26.67 ± 24.1	3.7 ± 0.42	3 ± 1	1.11 ± 1.47

Table 7.2: The table shows the descriptive values of the subjective measures used in the study across user groups. Scales for the measured factors are the following: SSQ: 0 (low) - 235.62 (high), Perceived Workload: 0 (low) - 100 (high), Interest/Enjoyment, Competence and Pressure: 0 (low) - 4 (high)

higher score for them (*Slower*: $p = 0.825, V = 56.5, r = 0.016$, *Perfect*: $p = 0.71, V = 67, r = 0.002$, *Faster*: $p = 0.325, V = 26, r = 0.132$). For the older adults, no condition produced significantly higher cybersickness (*Calibration*: $p = 0.595, V = 7.5, r = 0.244$, *Slower*: $p = 0.41, V = 11, r = 0.231$, *Perfect*: $p = 0.371, V = 3, r = 0.471$, *Faster*: $p = 0.136, V = 1.5, r = 0.501$).

7.3.2 Intrinsic Motivation

A short version of IMI was used to measure the Interest/Enjoyment of the players, as well as their Perceived Competence, and Pressure/Tension.

For both age groups, the highest enjoyment was observed in the *Faster* condition (4.7 ± 0.87 points for the older adults, 3.83 ± 0.42 points for the younger adults; possible maximum: 5 points). This is an important factor, since it is the self-report for motivation [280]. The *Slower* condition received the lowest enjoyment (4.22 ± 1.14 for the older adults, 3.47 ± 0.9 for the younger adults), which is still comparatively high. A Friedman test did not indicate any differences between the conditions for the younger ($p = 0.059$, $\chi^2 = 7.457$, $Kendall'sW = 0.083$) or older ($p = 0.3$, $\chi^2 = 3.686$, $Kendall'sW = 0.137$) adults.

The highest perceived competence was achieved in the *Slower* condition (4.78 ± 0.44 for the older adults, 4.36 ± 0.72 for the younger adults). Here, the *Faster* condition received the lowest score (4 ± 1 for the older adults, 2.66 ± 0.95 for the younger adults). A Friedman test indicated differences between the conditions for the older adults ($p = 0.016$, $\chi^2 = 10.292$, $Kendall'sW = 0.381$), but pairwise comparisons could not confirm this. For the younger adults, a Friedman test also indicated significant differences ($p < 0.001$, $\chi^2 = 49.235$, $Kendall'sW = 0.547$). They felt significantly more competent in the *Slower* condition compared to *Perfect* ($p < 0.001$), *Calibration* ($p = 0.006$), and *Faster* ($p < 0.001$). Furthermore, significant differences could be found between *Perfect* and *Faster* ($p = 0.002$) and *Calibration* and *Faster* ($p < 0.001$), with *Faster* resulting in less perceived competence compared to all other conditions.

For pressure, the *Faster* condition produced the highest pressure for both groups (2.11 ± 1.47 for the older adults, 2.49 ± 1.06 for the younger adults). In contrast, the *Slower* condition led to the lowest pressure (1.07 ± 0.15 for the older adults, 1.83 ± 0.95 for the younger adults). No significant difference was found for the older adults (Friedman: $p = 0.14$, $\chi^2 = 5.45$, $Kendall'sW = 0.202$). The younger adults (Friedman $p < 0.001$, $\chi^2 = 50.822$, $Kendall'sW = 0.565$) perceived significantly more pressure in the *Faster* condition compared to *Calibration* ($p < 0.001$), *Perfect* ($p < 0.001$), and *Slower* ($p < 0.001$), and more pressure in the *Perfect* condition compared to the *Calibration* ($p = 0.001$) and *Slower* ($p = 0.002$).

7.3.3 Perceived Workload

The highest workload was measured in the *Faster* condition (26.67 ± 24.1 for the older adults, 60.86 ± 15.84 for the younger adults), followed by the *Perfect* condition (15.65 ± 14.58 for the older adults, 40.83 ± 21.15 for the younger adults). The lowest workload was reported in the *Slower* condition (5.56 ± 4.77 for the older adults, 20.06 ± 15.54 for the younger adults). The *Calibration* received a workload score of 11.67 ± 14.19 for the older adults and 24.97 ± 14.45 for the younger adults.

A Friedman test indicated significant differences for younger ($p < 0.001$, $\chi^2 = 64.067$, $Kendall'sW = 0.712$) and older adults ($p < 0.001$, $\chi^2 = 16.618$, $Kendall'sW = 0.615$), but for older adults, no significant differences in post-hoc tests could be found for the total amount of workload in the conditions. For the younger adults, the *Faster* condition imposed more workload than all other three conditions ($p < 0.001$), and the *Perfect* condition received a significantly higher workload rating than the *Calibration* and the *Slower* condition ($p < 0.001$).

Regarding the question "*How did you perceive the level of difficulty in this round?*", the scores yield similar results, with the *Faster* condition being on average "*a little too difficult*" for the younger adults (4.07 ± 0.78) and slightly above "*just right*" for the older adults (3.22 ± 0.67). The *Slower* condition and the *Calibration* were rated as "*a little too easy*" by the younger adults (2.03 ± 0.81 and 2.37 ± 0.76 respectively), and received a rating of 2.44 ± 1.01 and 2.67 ± 0.71 by the older adults.

7.3.4 Game Performance

Game performance was recorded for each level. Table 7.3 shows the performance of the users in each game condition. This includes the mean distance between bubbles, the number of correct and missed bubbles as well as a movement summary. Two younger adults did not upload the game files correctly, so their data cannot be evaluated. Furthermore, we noticed that one younger adult made a mistake in the alphanumerical calibration level, always collecting the opposite of what their task was. Therefore, the difficulty for these levels was set to the easiest default value (one bubble every 4.8 seconds). Their data was therefore excluded from this analysis.

In the *Calibration*, the bubbles per minute were calculated for each level type; Simple, Coordination and Alphanumerical. We did this because we suspected that the different

Condition	Group	Level Type	Time between Bubbles (s)		Correct Bubbles		Missed Bubbles		Percentage Correct (%)		Movement Right (m)		Movement Left (m)	
			Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Calibration	Younger Adults	Simple	1.91	0.57	34.26	10.66	0	0	100	0	31.77	11.01	28.75	10.96
		Coordination	1.83	0.49	35.22	10.12	0	0	98.68	2.5	34.47	10.76	33.88	9.08
		Alphanumerical	1.81	0.49	35.11	7.76	0	0	95.9	3.48	32.46	11.98	24.02	8.72
	Older Adults	Simple	2.73	0.8	23.67	6.6	0	0	100	0	23.71	5.52	21.35	10.61
		Coordination	2.9	1.06	23.56	9.26	0	0	91.04	13.26	26.43	7.96	27.63	10.22
		Alphanumerical	2.79	1.18	24.44	8.75	0	0	83.97	19.21	27.93	9.09	21.35	10.61
Slower	Younger Adults	Simple	2.25	0.78	30.85	13.66	0.07	0.27	99.71	1.06	20.33	8.99	18.74	8.08
		Coordination	2.2	0.74	31.07	13.39	0.33	1.18	98.93	2.31	26.56	9.88	26.03	8.55
		Alphanumerical	2.17	0.7	30.3	9.08	0.37	0.97	98.66	3.42	21.98	6.57	18.75	6.77
	Older Adults	Simple	3.38	0.84	18.89	5.23	0.56	1.01	96.93	5.1	15.78	4.55	13.59	3.49
		Coordination	3.58	1.51	19.67	8.62	0	0	96.03	11.9	16.27	5.75	18.14	6.83
		Alphanumerical	3.13	0.98	20.89	6.33	0.11	0.33	99.63	1.11	16.41	6.97	15.8	6.23
Perfect	Younger Adults	Simple	1.43	0.74	54.67	31.16	4.41	8.21	95.04	9.63	32.84	17.52	31.26	15.74
		Coordination	1.31	0.48	52.07	18.89	2.33	4.39	95.25	6.13	41.81	15.94	41.16	15.74
		Alphanumerical	1.35	0.5	49.19	14.53	4.52	5.68	93.35	7.17	32.26	11.9	29.73	9.57
	Older Adults	Simple	2.17	0.83	31.44	11.64	0.11	0.33	99.57	1.28	21.38	6.51	22.49	7.8
		Coordination	2.45	1.28	32.22	18.67	0.33	0.71	96.92	6.25	28.49	15.01	26.62	14.41
		Alphanumerical	1.35	0.5	30.22	13	2.22	2.73	92.88	6.31	24.33	8.57	20.5	11.07
Faster	Younger Adults	Simple	0.62	0.17	102.85	24.85	28.04	11.36	79.5	6.55	56.82	16.54	54.74	15.61
		Coordination	0.76	0.16	82.63	17	18.37	10.62	78.95	6.79	62.49	11.11	62.65	13.17
		Alphanumerical	0.98	0.43	66.37	14.32	22.41	10.98	75.11	9.86	46.91	12.11	39.53	12.61
	Older Adults	Simple	1.09	0.49	64.78	25.55	12.89	15.43	87.39	10.34	39.59	10.72	35.64	17.38
		Coordination	1.19	0.37	54.33	15.07	9.22	7.98	83.14	9.32	45.28	14.22	43.86	14.77
		Alphanumerical	1.42	0.51	47.44	16.44	12.78	11.34	82.27	10.17	35.41	14.93	30.73	14.95

Table 7.3: The table shows the descriptive values of the game performance measures in each game level and user group.

types of levels were likely to require varying cognitive demands. For the Simple level, the younger adults collected one bubble every 1.91 ± 0.57 seconds, while the older adults collected one bubble every 2.73 ± 0.8 seconds. For the Coordination level, the younger adult's collection time was 1.83 ± 0.49 seconds and for the older adults 2.9 ± 1.06 seconds. Lastly, for the Alphanumerical level, the frequency was 1.81 ± 0.49 seconds per bubble for the younger adults and 2.79 ± 1.18 seconds for the older adults. Different to what we expected, there was no significant difference in the calculated collection time between the level types for older ($p = 0.68$) or younger adults ($p = 0.58$) as tested with a Friedman test.

During the game, the difficulty was evaluated and adjusted every five seconds. Figure 7.4 shows the percentage of correctly collected bubbles for each spawning time range that was played. It can be seen that first, the percentage is rising and then steadying at over 95 %. For the younger adults, the 95% correct line was crossed at around 0.875 seconds. That means that they were able to successfully collect a bubble every 0.875 seconds. For the older adults, it took a little longer, with the 95% range beginning at one bubble every 1.3 seconds. Regarding the level type, the participants in both groups were best in Simple levels, followed by Coordination levels and then Alphanumerical levels (See Table 7.1); therefore, different to the *Calibration*, the game performance decreased with higher cognitive demand during the game.

Regarding game performance, the total percentage of correctly collected bubbles for the older adults was 84.27 ± 9.83 % for the *Faster* condition, 87.92 ± 13.84 % in the *Calibration* and more than 95% for the remaining two conditions. The younger adults collected 77.85 ± 8.02 % correctly in the *Faster* condition, around 95% in the *Calibration* and *Perfect* condition and 99% in the Slower condition. The percentage of wrong bubbles was very low for all conditions, and this was similar for missed bubbles, except for the *Faster* condition. Here, older adults missed 11.63 ± 11.64 % of bubbles, while younger adults missed 22.94 ± 11.56 %.

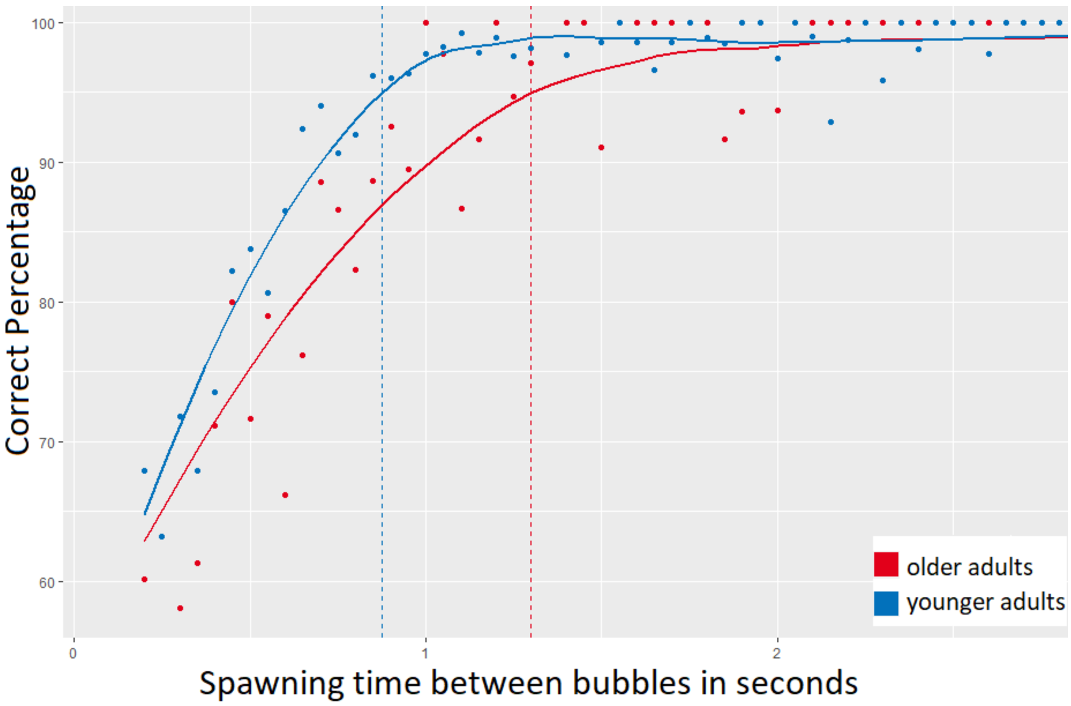


Figure 7.4: The percentage of correctly collected bubbles for each spawning time between bubbles. Fit with a Loess function.

The next performance metric we evaluated is throughput, which depicts the relation between precision (correct, wrong and missed objects) and speed. Comparing the speed-accuracy-trade-off is a well-known metric, e.g., for Fitt's Law [92], but this metric has to be adjusted to fit our game. We define this as the number of correct bubbles minus the number of missed or wrong bubbles per minute. Even though the percentage of correctly collected bubbles was lower for the *Faster* condition, its throughput was still higher than in the other conditions, with a total of 58.6 ± 21.71 bubbles for the younger adults and 42.18 ± 13.01 bubbles for the older adults. In comparison, the throughput in the *Perfect* condition was 30.26 ± 14.14 for the older adults and 47.57 ± 18 for the younger adults. It was even lower for the *Calibration* (older adults: 21.41 ± 10.14, younger adults: 34.17 ± 9.35) and the *Slower* condition (older adults: 19.41 ± 7.09, younger adults: 30.38 ± 11.94). This shows that the participants were capable of more than what was determined in the *Calibration*.

Furthermore, we analyzed the movement of the hands during the game. There was a higher amount of movement in the *Faster* condition, where more bubbles had to be collected. The coordination levels showed the highest amount of movement within their condition. In the comments, two younger adults reported that in the *Faster* condition, it was not possible to move as fast as the bubbles appeared (Py_{16}, Py_{28}). Two others mentioned that, in this condition, they just moved their arms without control or coordination to collect the bubbles (Py_2, Py_{33}). This indicates worse accuracy of movements, which should be important for physical exercise, also similar to the results for movement accuracy in the last study (See Chapter 6).

7.3.5 Qualitative Results

The game was well received by all participants. When asked how much they would like to play another round, the *Calibration* (older adults: 4.56 ± 0.73, younger adults: 3.5 ± 1.11 points; on a 5-point Likert scale from 1 to 5), and the *Faster* condition (older adults: 4.56 ± 0.53, younger adults: 3.43 ± 1.07) received the highest scores. The *Slower* condition received the lowest rating (older adults: 4 ± 1.32, younger adults: 3.23 ± 0.94), which still equals the willingness of "*fairly*" and "*moderately*" wanting to play again.

From the younger adults, eleven preferred *Faster*, eight preferred *Perfect* and seven liked *Slower* the most. Furthermore, the *Calibration* at the beach was chosen as the favorite condition by three participants. From the older adults, four liked *Faster* the most, three *Perfect*, one *Slower* and one could not decide between the *Calibration*, *Faster* and *Slower*. This person decided based on the levels they played, not based on the difficulty.

In general, the comments about this game were very positive, especially from the older adults. Po_{39} mentioned that this could replace the chair gymnastics they were normally doing and ten people mentioned that the game was fun (Py_2, Py_4, Py_{11}, Py_{13}, Py_{17}, Py_{19}, Py_{22}, Po_{26}, Po_{36}, Po_{37}). Eight younger adults proposed to add music to the game (Py_5, Py_6, Py_{13}, Py_{15}, Py_{20}, Py_{22}, Py_{27}, Py_{33}), which has also been shown to increase motivation [377].

Players reported that while in the boat, they could immediately see where the bubbles were, compared to the *Calibration*, where bubbles sometimes were out of the field of view (Py_2, Py_6, Py_{11}, Py_{13}, Py_{17}, Py_{18}, Py_{33}). Furthermore, several younger adults liked that in the boat, there was more action, more things happened, and it was more dynamic (Py_3, Py_6, Py_9, Py_{19}, Py_{24}). It also was a bigger challenge because they only had a certain time to react to bubbles and could therefore miss them (Py_5, Py_9, Py_{10}, Py_{11}, Py_{15}, Py_{16}, Py_{23}, Py_{28}), emphasizing that the moving gameplay is suitable for this group.

The older adults liked the setup in the canoe because of the nice landscape (Po_{30}, Po_{37}), although they mentioned not having enough time to look at it (Po_{30}). It was more challenging (Po_{29}), faster (Po_{36}), lively (Po_{25}) and included more action and different tasks (Po_{25}, Po_{26}) as well as more movement and concentration (Po_{38}, Po_{39}), compared to the *Calibration*. Most of them could imagine playing the *Calibration* to get used to the game and warm up (Po_{37}, Po_{38}, Po_{39}) and also because it takes the fear away (Po_{26}). Only one older adult mentioned that they would like to start directly in the canoe and with more power (Po_{30}).

7.4 Follow-Up Study

Due to the current COVID situation, this study only included a small number of older adults. This was due to the restrictions by the senior living facility and the enhanced hygiene protocols, preventing us from conducting in-person studies, which many older adults prefer, compared to a video call. More participants would enhance the quality of the data and enable further, statistical, comparisons with the younger adults. This further emphasizes the difficulties in conducting studies with small or special user groups. Furthermore, we evaluated a specific calibration method for a specific game, and transferability to other games and methods might be limited. We acknowledge that there are several other methods for DDA that future studies might consider: for example, Pezzera et al. [250]'s method accounts for additional variables for game difficulty such as the emotional state of players. To validate the conclusions we drew from the evaluation of two user studies, we were able to perform a follow-up study during the summer months, with some necessary improvements to the game that we identified based on the primary user study with young and older adults.

Improvements Our primary study results suggested changes in the game design, e.g., bubbles being out of the field of view in the calibration, resulting in a long searching time, distorting the calibration results. Therefore, the maximum arm ROM was changed to 65% in this condition, spawning the bubbles closer towards the front of the user. We further noticed that the calibrated values for the three different level types were very similar. Therefore, we decided to only administer one calibration round and play the experimental conditions with these results. Before the calibration, we introduced a short familiarization phase, where the players had to touch four bubbles to get used to the game mechanics as a tutorial.

After the calibration, we set the minimum distance between bubbles to 0.5 m (75% correctness rate of older adults), resulting in no more than two bubbles per second. This should prevent the game from getting impossible, but still provide a challenge for fitter players. For the game difficulty, we set it to a difficulty one third higher than what was achieved in the calibration – accounting for learning effects and following the suggestions of our study participants and Cox et al. [53].

Method In this follow-up study, we aimed to particularly focus on the enjoyment and perceived difficulty of the game with the improvements made. Additionally, we wanted to observe the older adults in-person, to see their immediate reaction, and identify possible safety and usability issues with the HMD.

11 older adults ($M = 83.82, SD = 3.06$ years) participated in this follow-up study (all female, seven had used VR before). They played four rounds with a duration of one minute each; the calibration, a Simple level, a Coordination level and an Alphanumerical level. Then, they answered the following questions: *How much fun was this game (no fun at all (1) - a lot of fun (5))?* and *How did you perceive the difficulty of the game (way too easy (1) - way too hard (5))?* Furthermore, they were able to leave comments.

Results All older adults rated the game difficulty as precisely right for them (3 ± 0). This confirms that the calibration method, with the changes we made, can help to adjust the game to the user's abilities. When asked about their perception of fun, the result was 3.73 ± 1.1, indicating a high feeling of fun for most players.

As a conclusion, with the results from our follow-up study, we were able to adjust this exergame for older adults. The suggestions from the younger adults, as well as the performance measures from the first group of older adults helped us, which emphasizes that testing a game with both user groups has its value. Younger adults can help to identify structural mistakes and can be considered as a pre-testing group before evaluating a game with older adults. But for evaluating performance and setting a suitable difficulty, we would advise to test with the actual user group.

7.5 Discussion

In this section, we discuss the findings of the study on the effects of the calibration method and the differences in the two user groups with the focus on the research questions.

7.5.1 Effect of the Starting Difficulty ($\mathbf{RQ}_{3.1}$)

The *Slower* condition showed high perceived confidence and low perceived pressure. It was the condition with the lowest number of mistakes, but also with the lowest number of collected bubbles, i.e., the lowest amount of exercise. The *Faster* condition on the other hand resulted in the opposite – lowest perceived confidence, highest pressure with the highest number of mistakes or missed bubbles. Nevertheless, the total number of bubbles was the highest in this condition, meaning that the participants performed the most exercises here.

When balancing cognitive-physical VR exergames, several factors have to be considered. First, the game should be engaging and enjoyable, subsequently leading to a high adherence in long-term usage. In this example, the faster conditions indicated these results. Furthermore, its physical consequences should be positive – not yielding severe cybersickness and resulting in therapeutically valuable movements. From observation and qualitative comments, the *Faster* condition resulted in reduced accuracy regarding movements; this should further be evaluated by movement specialists or with high-end tracking systems. Lastly, the game should be challenging, but also enjoyable. For future games, we propose to individually calibrate a game, but to incorporate some sort of in-game balancing system as well. Another idea would be to implement slower and faster phases of gameplay to give the players some time to relax, think and plan their next steps, instead of only adjusting the game with linear steps. It would also be interesting to evaluate different DDA methods (see [392]), such as machine learning combining prior performance knowledge of other participants with the player's own performance to achieve a good balance.

7.5.2 Adjustment of the Exergame ($\mathbf{RQ}_{3.2}$)

The older adults rated the game more positively than the younger adults, especially in terms of intrinsic motivation. Since the game was developed for the older user group, this finding is plausible. The younger adults might already have more experience with (VR) games, and, therefore, have different or higher expectations. Nevertheless, the game received many positive comments and suggestions from the younger adults as well. The questionnaire data further indicated their interest in playing the game again. This implies that, with the right calibrations of, e.g., game speed and task difficulty, this game can be played by multiple user groups. This opens up opportunities to also develop multiplayer versions of the game to create a social connection between the older adults and, for example, their younger family members, similar to the work of Cantwell et al. [44]. Here, the individual abilities of the users need to be integrated, but the game still needs to be entertaining and challenging for all players. With the knowledge gained from this study, we believe that calibrating a game that is suitable for both user groups is possible.

7.5.3 Transferabilities and Differences ($\mathbf{RQ}_{3.3}$)

Several transferabilities and differences between the two user groups have been identified: both in subjective measurements as well as in their game performance.

Fortunately, no strong cybersickness could be identified when playing this game with younger adults; the same could be confirmed with older adults. We advise to test for these negative effects with user groups whose symptoms are potentially not as severe, e.g., younger users, and to design the VR environments for older adults in a careful way [287].

Perceived workload was higher for younger adults than for older adults. We did not expect this finding, and the perceived workload for older adults could not have been predicted from only testing the game with the younger users. It might be because the game difficulty was objectively easier for the older adults as they were slower in the *Calibration*, or because they needed some more time to understand the game flow.

Regarding the game performance, we expected that younger adults would perform better and react faster than the older adults [372], especially with increased task difficulty [208]. Testing the game with younger adults revealed their limits with respect to their processing speed. This could have helped to adjust thresholds for the game speed for older adults as well, but after reviewing the data of the first group of older adults, further adjustments had to be made. For the older adults in the second study, whose mean age was a little higher than that of the first group, the performance plot again showed different thresholds. Therefore, we conclude that performance measures, especially with a cognitive-physical component, cannot be transferred easily between user groups.

The younger adults' feedback was very valuable for improvements, suggestions, and errors in the game flow. Since the younger adults in our study all had an IT background, they identified these issues and were able to communicate them. For example, the problem of not finding all bubbles quickly during the calibration, because some appeared outside their field of view, could have been identified with pre-tests with younger adults. Furthermore, they know what is theoretically possible in VR, whereas the older adults in our study had difficulties to think of game improvements due to a lack of experience with this kind of game or technology [67].

Lastly, when answering (standard) questionnaires, some questions were answered by the older adults with a different background in mind, e.g., the question regarding the favored round— here, we aimed at the difficulty of the round, not the specific levels they played in this round. However, this was not the case for younger adults, who answered the question as we intended. Additionally, the older adults in our study tended to answer more positively than younger adults. While the experience of older adults could be subjectively better than that of younger adults, it could also be that older adults did not want to complain or admit that they had problems with the game aspects out of courtesy or that the positivity effect occurred [205]. To mitigate this potential bias, we suggest testing games with different user groups and giving the participants enough privacy and data safety to reply openly. However, we also note that since almost all games are developed for a specific user group, the experiences of different groups may differ and it may be difficult to draw comprehensive conclusions.

7.6 Conclusion

In this study, we evaluated the effect of three different calibrated difficulties (*Slower*, *Perfect*, and *Faster*) on intrinsic motivation, cybersickness, perceived workload, preference, and game performance. We considered both, older adults ($n = 9$) and younger adults ($n = 30$) as user groups in our study. Most players evaluated a higher difficulty as more enjoyable, but this difficulty level also led to a higher perceived workload. During the gameplay, this caused faster and less controlled movements that could potentially impact the therapeutic benefits of the VR exergame. Hence, we conclude that for training purposes, VR exergames should be thoroughly tested with the end users and movement specialists. This could help to find optimal parameters for particular user groups without hindering VR exergames' potential therapeutic benefits.

Moreover, our findings successfully showcase using individually calibrated starting difficulty for a VR exergame. Through this, the game became suitable for the use of two different user groups: younger and older adults. However, the results also highlighted (non)-transferabilites between these groups. While the recommendations of younger adults helped to improve the game for older adults as well, adjustments of the game's difficulty was shown to be a non-transferable feature among user groups. Therefore, our results show the added value of an additional user group in the process, especially if the target group is not available. Nevertheless, the actual target group should always be involved in both the design and evaluation phase to ensure a user experience that fits their specific needs.

Chapter 8

Discussing Immersive Training for Older Adults

The three presented studies have shown some of the positive effects that our designed exergames can have on psychological, physical and cognitive well-being of older adults.

The first study demonstrated an increased well-being for older adults with dementia through exergames ($\mathbf{RQ}_1$). In this context, the potential of exergames to provide stimulating activities with long-term engagement and high user satisfaction were pointed out. Qualitative feedback emphasized the intervention's meaningfulness, providing a sense of accomplishment and distraction during the COVID-19 pandemic. This study provided the foundations for understanding the psychological, cognitive and physical benefits of exergames for older adults with dementia, and suggests further explorations of this research topic with different user groups as well as in different contexts. Nevertheless, limitations such as a small sample size and the absence of other training possibilities during the time of the study yield the need for future investigations.

Building onto the first work, the second study compared a traditional video-based exercise program to a VR exergame in order to see whether positive short-term effects like those of the first study could also be achieved through different media, e.g., through traditional video-based exercises, and whether a different user group, i.e. older adults without dementia, would enjoy these exergames. The users' feedback indicates that while VR exergames can be a suitable addition to traditional training ($\mathbf{RQ}_2$), the exergames should be aligned to the preferences and needs of the users. This could enhance its acceptance and provide a stimulating experience. It was suggested to include familiar movements into the exergames, and to implement exercises with a suitable physical impact in order to encourage long-term commitment, for example by tailoring the game difficulty to the users.

Consequently, the third study evaluated the effect of dynamic difficulties in a VR exergame that was evaluated with both, older and younger adults ($\mathbf{RQ}_3$). Findings indicate that a higher difficulty and intensity increased the enjoyment of users, but also their perceived workload, and it effected the accuracy of movements which might deteriorate the therapeutic value of the exercises. This study emphasized the importance of tailoring VR experiences to the end users, and pointed out the lack of transferability between younger and older user groups for some features, e.g., difficulty thresholds. Nevertheless, the opportunities of testing exergames with younger users were also pointed out, e.g., the possibility of identifying structural mistakes in the exergame prior to testing it with special user groups.

Together, these studies illustrate the potential of VR exergames for improving the well-being of older adults, underscoring the multifaceted approach we employed. They imply that user preferences, familiarity with exercise formats and tailored difficulty levels play an important role in the design of VR exergames.

8.1 Design Considerations

Further, our studies have shown several barriers that may discourage older adults from exercising with VR in the future, especially on their own. Three design considerations to remove these barriers will be presented in the following, as they also have a direct effect on future studies and on the independent interaction of older adults with VR technology: (i) The dependability of older adults on caregivers, (ii) the importance of feedback and encouragement employed in exergames, and (iii) the differing needs of different user groups.

8.1.1 Dependability on Caregivers

To begin, the first user study has shown that there is always a dependability of the older adults on their caregivers, especially for people with dementia. This ranges from being prepared to arrive at the location of the studies to adjusting the VR HMD, but also in being guided through the game flow. Caregivers already have an enormous workload [107], with a low number of caregivers for a rising number of older adults requiring their assistance. In this context, individual exercise activities are often not possible as the basic needs of all older adults need to be fulfilled with a higher priority. Family members and other informal caregivers, who could potentially assist the older adults during the VR exergames, are often interested to spend the time with their loved ones in a meaningful way [128]. This further emphasizes the importance of designing the exergames in a way that encourages meaningful conversations between the older adults and the assistants, or to find alternative ways in which the exergame can be experienced with minimal effort from external instances.

8.1.2 The Importance of Feedback and Encouragement

The second study has shown that both exercise methods lacked specific feedback and ongoing encouragement. While the position of the participants' arms was tracked in VR, and the musicians stopped playing if it deteriorated too much from the given path, no specific feedback on the movement accuracy was given. In the video-based exercises, this also led to a non-consistent execution of movements, which, in the long-run, could be detrimental for training progression and success. For long-term commitment, the tasks in the second exergame showed to be too repetitive and too physically easy for many participants. Here, the addition of an encouraging element could benefit the game, e.g., motivational language, rewards or a continuous progress report [25]. Further, looking at feedback from the other side – feedback voiced by the older adults – if the game is too easy, it is important to consider this for future iterations. A way to raise the difficulty of the exercise program dynamically should be integrated, and it should be considered which game elements are wanted and needed.

8.1.3 Different Users have Different Needs

In the last study, it was shown that different users have different needs and preferences regarding the game difficulty. This can be extended to different needs in caregiving, or supervision, or assistance. Further, it can depend on different factors such as experience with technology, physical or cognitive health, or even change on a day-to-day basis according to the participants' mood [58, 82], as it was also reported in our first user study (See Chapter 5). In order to enable all users to independently experience the beneficial nature of VR exergames, methods have to be implemented that balance the game, and cater to the needs of different users dynamically.

8.2 Future Work

Our research has shown the potential of VR exergames for older adults, but also pointed out limitations of our implementations. Related to the three design considerations discussed in the previous section, one future research direction is to improve the independence of older adults in playing and experiencing VR exergames without external help, integrating suitable methods to tailor the experience to their needs. To achieve this, we included intelligent virtual agents in our VR cognitive-physical training applications - computer programs mimicking human behavior. With the addition of artificial intelligence, these have the potential to dynamically answer the users' questions, guide them through the immersive exercise programs, and provide feedback and motivating encouragement – aiming to improve the user experience of our VR exergames, and further motivate the users' long-term adherence.

Part III

Intelligent Virtual Agents as a User Interface Element for Immersive Training

Chapter 9

Introducing Intelligent Virtual Agents for Immersive Training

9.1 Motivation

In order for exergames to be effective, usable and motivating, they need to show a high usability that is matched to the abilities and preferences of the user [382]. Aside from intuitive interactions [150], a well-designed immersive environment [133], and motivating, meaningful content [80, 82, 128, 133, 150, 307], the on-boarding of the game should be designed with a focus on engagement, providing a sense of trust and safety [133].

In this regard, it is important that the goals and objectives of the experience are well-explained and understandable [80]. The user group of 'older' adults is heterogeneous in itself, with adults in their 60's having different technological knowledge, experience or cognitive-physical fitness than older adults in their 90's [77], and the information should be provided in a suitable way, with currently required details, for all users. This information can be communicated through different channels, e.g., visual, auditory or through haptics [30], depending on the user's needs and preferences. The provided assistance could further be tailored to the users' knowledge and experience, e.g., by providing more advanced tips [25], or by adjusting the difficulty of language used. For people with dementia, it is recommended to give detailed instructions and repeat them, if necessary [150].

Often, caregivers, researchers or family members motivate older adults to try out VR [127, 336, 337], and they are the ones answering questions or assisting with technical difficulties, which raises concerns about additional workload being added onto their busy schedule [308]. Additionally, caregivers often face staff shortage and an enormous load of responsibilities [107]. Without their involvement, older adults would often be left alone with questions regarding the hardware usage or the game mechanics, which might restrict their access to the beneficial cognitive-physical training content. While, in our opinion, human contact is crucial for well-being, and should not be replaced, providing individual care for the administration of exergames is often simply not possible. Therefore, it remains an open research question whether and how the human involvement of caregivers and family members in this domain can be reduced, while still enabling older adults to experience VR.

9.1.1 Intelligent Virtual Agents

One approach to solve these issues is through virtual agents (VAs) or intelligent virtual agents (IVAs). These are computer programs that mimic human communication and social interaction [70]. IVAs can range from text-based chatbots to voice assistants, and to embodied agents. They have been employed in a variety of application fields, including mental health [37] or physical training [226], and they take on different tasks such as educating, motivating or monitoring users [165].

In prior research with older adults, IVAs were used to practice patient-doctor conversations. After training with an IVA how to communicate their pain, the participants were able to transmit more information with a real, human doctor [207]. In another paper, IVAs were used to show older adults how to perform easy balance exercises [225]. Aside from preferences on the representation of the agent, their results indicate that the older adults wished for the ability to repeat the instructions of the IVAs. Straßmann et al. have also evaluated the visual representation of virtual agents, with younger and older adults [322]. In their study, the participants interacted with health-related daily life assistants and their task was to fill out a health diary. Here, the agent's interaction was scripted and controlled using the Wizard of Oz methodology. Results indicate that the older adults rated the agents more positively and showed more bonding than the younger adults.

IVAs are not only represented in the shape of humanoid agents. Albaina et al. evaluated Flowie, a virtual coach in the shape of a flower that encourages older adults to walk more [7]. Their prototype was found to be encouraging and stimulating, but participants wished for more context-awareness of the virtual agent, e.g., to adapt the timing of the IVA's motivational cues. In our prior work, we compared a humanoid agent, an abstract text/voice representation and a virtual parrot in the aforementioned *Canoe Game* [187]. We found that younger and older adults preferred the parrot, but also that the personality of the user has an effect on their perception of the agent.

9.1.2 Artificial Intelligence and Large Language Models

Recently, OpenAI's ChatGPT has gained attention through its ability to produce human-like text output based on the user input it receives [244], starting a new era for communicating with IVAs. In the version *GPT-4o*, the multimodal large language model (LLM) accepts text, audio, and image input, and it shows human-level performance on several different benchmarks [2]. Regarding its use in VR exergames, LLMs can be prompted to communicate game rules and answer questions of the users regarding the content of the experience or assist with technical issues. One advantage is the LLM's ability to adapt its language, e.g., using more formal language for advanced users, or explaining difficult content in easier words for novice users. In VR, a GPT model was previously integrated as an assistant in an escape game, which was evaluated with younger adults [283]. Here, it helped the players to solve riddles, but limitations of the model regarding its knowledge of VR-specific contexts and several vague responses were observed. For older adults, a scoping review from 2022 found that only 1% of healthcare applications using AI with older adults featured VR technology [200].

While LLMs or voice-based systems have been evaluated for humanlikeness – for example, for their ability to pass the Turing test [146] – and ethical questions have been discussed [278, 302], applied research with older adults remains scarce. LLM research with older adults is often limited to basic evaluations, without integrating LLMs in already established systems. For example, with the involvement of older adults, GPT-4 has been evaluated as a screening test for mild cognitive impairment [375], and the authors claim that it achieves a higher detection accuracy than junior neurologists. Further, with voice-based systems, vocal characteristics like pauses or the behavior in error-handling situations were assessed with older adults [162], showing connections with cognitive decline.

9.2 Research Questions

Integrating IVAs successfully only works if they are accepted by the users. Therefore, it is important to find out how to design them, which prerequisites should be fulfilled to deploy them and which use cases they are most suitable for.

Chapter 10 presents a systematic literature review on embodied virtual agents in healthcare that aims to provide an overview over the users' preferences. We conducted this review on the whole body of healthcare literature on agents instead of the narrower field of cognitive-physical training to get a better picture of the roles of agents in this broader context, and due to the limited amount of IVA research in VR exergames. Further, diverse user groups and demographics, such as younger and older adults or adolescents and children were included in this review. Within the scope of this thesis, this section addresses the following research question:

- $\mathbf{RQ}_4$: How can intelligent virtual agents in the medical context be represented and which use cases are the most suitable?

In Chapter 11, we designed and evaluated an immersive cognitive-physical training for solving anagrams, which was explained and observed by different representations of an IVA, which varied in their level of visibility. In a user study that we conducted with younger adults, we evaluated the performance, social presence, and eye tracking information in the presence of each agent representation. In this study, we discussed the following research question:

- $\mathbf{RQ}_5$: Which agent visibility representation is the most suitable for immersive training applications?

In the final study, Chapter 12, the knowledge from all previous chapters was combined to integrate an IVA controlled by AI in the *Canoe Game*. Here, we observed the interaction between older adults and an IVA and interviewed them on their perception of the agent. This chapter explores the following research question:

- **RQ**$_6$: How do older adults perceive and use an intelligent virtual agent in a VR exergame?

In a final discussion, the contribution from these three works is reflected and design considerations for future research are pointed out.

Chapter 10

A Systematic Literature Review on Embodied Virtual Agents in Healthcare

Figure 10.1: An example of medical virtual agent in her office that we designed for illustrative purposes.

Medical virtual agents (MVAs) hold great potential to support users in achieving their health goals, especially at times or in regions where the demand for physiological and psychological therapy exceeds the capacity of medical services. For example, the agents could be employed in senior living facilities to compensate for the shortness of staff and relieve them in stressful situations. Further, optional physiotherapeutic training or cognitive therapy offers, which would otherwise not be possible, could potentially be conducted with the agents. To create an accepted agent for counseling, education and therapy, it is critical to understand the impact of factors such as the agent's visual representation and behavior to create a trustworthy human-agent relationship. Aiming to gather information about the preferences of users regarding IVAs, we conducted a systematic literature review with 59 papers about embodied virtual agents. This chapter is based on the following publication:

Kruse, L., Hertel, J., Mostajeran, F., Schmidt, S., and Steinicke, F. (2023). Would You Go to a Virtual Doctor? A Systematic Literature Review on User Preferences for Embodied Virtual Agents in Healthcare. In IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (pp. 672-682). DOI: `https://doi.org/10.1109/ISMAR59233.2023.00082`

10.1 Introduction

State-of-the-art IVAs combine models of natural language understanding, speech synthesis, and computer graphics to create versatile representations that can be customized to meet the needs of a specific use case [292]. Technologies of VR and AR allow for an embodiment of the agent, which has not only been shown to significantly increase an agent's credibility, trustworthiness, and social presence [293], but also has the potential to enrich human-agent communication through (subtle) facial expressions and gestures. The use of machine learning (ML) models can further provide agents with deep knowledge of physical and mental health, which can already lead to a diagnostic performance similar to that of a human expert [268]. As such, IVAs can be especially valuable when the need for physiological and psychological treatments exceeds the capacity of medical services [367]. According to the *WHO*, at least half of the world's population does not receive essential health services [369]. While we believe that MVAs are not an equivalent substitute for real doctors or therapists, they could potentially reduce the workload of medical staff and create new approaches in the field of telemedicine [366]. This potential became particularly apparent during the COVID-19 pandemic, in which VAs could have made a remarkable contribution by providing information and certain forms of therapies in a contactless manner and without requiring patients to be at a specific location [257]. Moreover, since therapies and their success highly depend on the individual patient [233], IVAs could potentially address the patient's personality and situation and use the resulting individual requirements to tailor the treatment accordingly [147].

For the practical use and mainstream adoption of virtual doctors driven by AI, it is critical to know when and under which circumstances patients would accept their consultation instead of visiting a real doctor. In this context, we are especially interested in how to represent these virtual doctors or therapists in terms of technology as well as their appearance and behavior. While the visualization of agents and avatars in VR/AR has already been investigated in a prior review [360], literature on medical agents in VR and AR remains scarce. We are therefore extending the technological range to include all kinds of display technologies, and subsequently reduce the scope of our review to a medical scenario.

AI agents in healthcare have been the subject of some prior reviews, e.g., focusing on effectiveness and usability with positive to mixed evidence for a high effectiveness in around two thirds of studies and good usability for the majority of presented studies [219]. However, the authors also reported barriers such as limited voice recognition accuracy, difficulties with forming personal connections with the agents, or repetitive dialogues. Another prior review investigated the user experience of virtual health assistants who provide health advice and support for patients [57]. They emphasized the design characteristics of an agent, such as its visual appearance or language features. For instance, they stated that in order to be accepted by users and to ensure long-term adherence to therapies and treatments, it is crucial to know how to represent these medical agents; this includes their visual appearance, the interaction modalities as well as their behavior. We extend this review for a broader representation of different kinds of healthcare agents, including virtual doctors and therapists. Furthermore, their analysis mainly focused on quantitative data, explaining how design characteristics influence user experience, but not necessarily reflecting the view of the users on these characteristics.

While prior literature reviews cover different aspects of virtual health agents or telemedicine, a thorough review of the existing literature focusing on the (potential) user preferences for embodied virtual healthcare agents is still missing. The tangible potential for VAs from a healthcare perspective [198] is only of value if users accept them as a point of contact for their medical matters [345, 368]. This review therefore aims to systematically examine the current literature on embodied VAs in this domain, evaluating in particular those works that base their findings on qualitative feedback from users. In this effort, the following research questions are addressed:

- **RQ**$_{4.1}$: How and with which technology can virtual doctors be presented?
- **RQ**$_{4.2}$: For which symptoms or special treatments could virtual doctors be particularly suitable and which ones have been investigated most in literature?
- **RQ**$_{4.3}$: Would users consult a virtual doctor or psychiatrist and what are the reasons for and against this?
- **RQ**$_{4.4}$: What conditions must be met to create social acceptance of virtual doctors and to increase the willingness of patients to consult them?

The contribution of our review is manifold: (i) An in-depth examination and preparation of current research trends related to MVAs, (ii) a thematic analysis of user feedback from prior literature, and (iii) a comparison of agents that are currently prevalent in the research and those that users envision in order to derive future research opportunities.

10.2 Methods

We performed a systematic literature review based on *Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)* [189, 300]. The review was preregistered in the Open Science Framework[1]. An overview of the process with corresponding paper counts after each step is provided in Figure 10.2.

10.2.1 Study Selection

We searched three databases *PubMed*, *ACM Digital Library*, and *IEEE Xplore* with the search terms listed in Table 10.1, which were collaboratively defined. It contains alternative expressions for (1) virtual agents, (2) healthcare, and (3) empirical evaluations. To be eligible for inclusion, papers must contain the specific search term in the title or abstract.

[1] `https://osf.io/hf7d4/?view_only=1a48362b3ab94750a93d7321a9c43d3a`

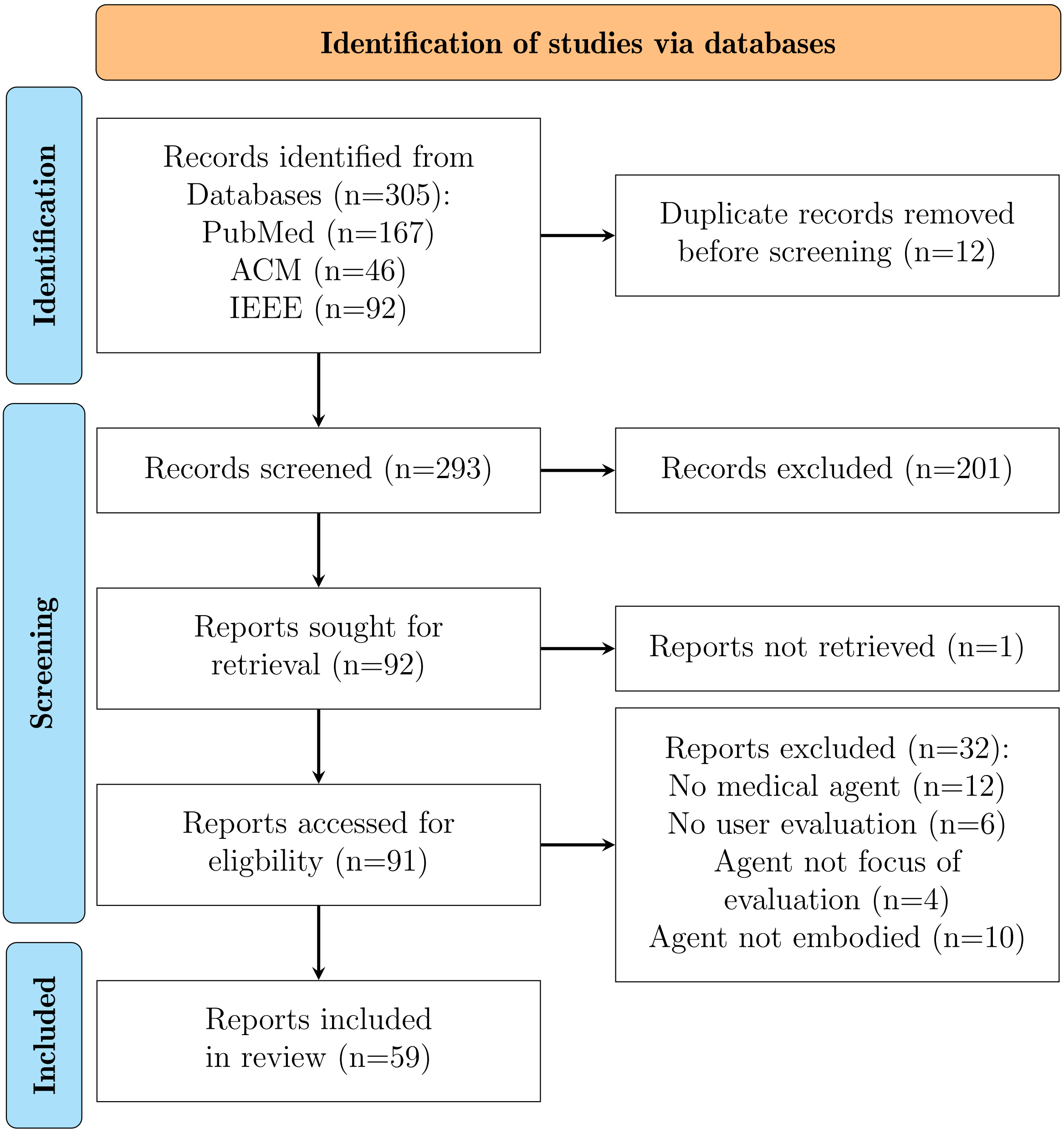


Figure 10.2: A flowchart showing the PRISMA process with all included and excluded paper numbers.

Category	Search Terms
(1) Virtual Agents	(“virtual agent”) OR (“virtual intelligent agent”) OR (“virtual doctor”) OR (“embodied artificial intelligence”) OR (“virtual therapist”) OR (“virtual coach”) OR (“virtual human”) OR (“virtual character”)
(2) Healthcare	(patient) OR (therapy) OR (diagnosis) OR (“medical treatment”) OR (health) OR (clinical)
(3) Empirical Evaluations	(empirical) OR (evaluation) OR (experiment) OR (study) OR (interview) OR (survey) OR (“focus group”)

Table 10.1: Terms used in our database search. All categories were joint with an ”AND” operator. Search terms had to be included in the title/abstract.

A total number of 305 documents were resulting from the initial database search. Twelve duplicates were manually removed based on the same title and authors. For the remaining 293 publications, the abstract, title and author information were downloaded from the databases and imported into Dovetail [73] for subsequent screening.

10.2.2 Screening

For initial screening, the research articles were divided onto four researchers (two researchers for each sixth of the data). We only included research articles that were written in English in our review. Screening inclusion criteria for the title and abstract were as follows: (1) The paper describes interactions between (potential) patients and a VA with a 2D/3D embodied representation. (2) The VA is processing the patient's health data with a focus on medical or therapeutic treatment. (3) The paper includes an empirical evaluation with human participants. (4) The paper is available online. (5) Literature reviews and other kinds of meta analysis, as well as workshops or talks, are excluded.[2]

Based on these inclusion criteria, custom tags were assigned using Dovetail. Each paper had to be accepted by two researchers. If the researchers disagreed on a paper, a tiebreaker scanned the abstract or content of the paper and decided on its (tentative) inclusion or exclusion. For the screening step, we conducted Cohen's Kappa test [95] to calculate the inter-rater reliability. For all pairs of the involved researchers, the results indicated a strong agreement [210].

Following the described process, 78 papers were accepted directly by the two assigned screeners, and another 14 papers were included after final approval by a tiebreaker. This led to a preliminary acceptance of 92 papers based on the review of the inclusion

[2]Compared with the protocol pre-registered on OSF, criteria (4) and (5) were added during the screening phase. After familiarization with the literature, we also collectively decided to include caregiving as well as life, fitness, and health coaches in the review if they fulfilled the other inclusion criteria and presented a therapeutic, health-related background.

criteria in the paper title and abstract. The selected articles were downloaded from the database websites. One article could not be located online and was therefore excluded from the analysis [236].

10.2.3 Full-text Eligibility Check and Data Extraction

The remaining 91 papers were divided among the us, with each paper being read by two researchers to reduce the subjective bias of each reviewer. Descriptive data for each publication, as well as findings related to the research questions, were recorded in an Excel spreadsheet. If publications included more than one study or participant population, we considered them separately and only extracted data that met our inclusion criteria. The data from the respective two researchers per paper were merged and conflicts were resolved by revisiting the paper.

Based on the full-text review, we collaboratively decided to exclude another 32 papers due to not including an evaluation (n=6), not including an agent matching our inclusion criteria (n=12), the agent not being the focus of the evaluation (n=4), or the agent not being described as embodied (n=10). This resulted in a final set of 59 papers that serve as the basis for our literature review on MVAs.

While the extracted qualitative data was further examined in a thematic analysis to investigate user preferences, user-independent information about the reviewed papers was used to create an overview of the state of the art of medical agent research. In iterative group discussions, we created dimensions to describe the characteristics of agents in the reviewed papers, as well as the underlying technical system and the application field. In a mixed (inductive and deductive) approach, we defined distinctive categories or investigated trends for these dimensions.

10.2.4 Thematic Analysis

For the final set of papers, extracted data was imported from Excel into Dovetail for further analysis according to the research questions. For this, we used reflexive thematic analysis [32], with the following steps: Familiarization with the data, generation of initial codes, searching for themes, reviewing themes, defining and naming themes and, finally, producing the report. The initial generation of codes was done by two researchers who, after familiarization with the data, reviewed it independently of each other. This resulted in a total number of 86 codes. Afterwards, all codes were discussed with three of the researchers; duplicate codes were merged and themes were defined, reviewed and named. This process resulted in 41 final tags, grouped into the themes that will be discussed in section 10.3 and section 10.4. After defining the final themes and corresponding codes, every paper entry was revisited in Dovetail and assigned fitting codes based on the data extracted during the full-text review. In a final step, these code assignments were validated by a different researcher.

10.3 Results

In the following, we will provide a holistic overview of the included publications, especially regarding their application field, the representation of the MVAs, and the methodology for user evaluation. As part of this, we will put a special emphasis on research questions $\mathbf{RQ}_{4.1}$, i.e., how and with which technology virtual doctors can be presented, and $\mathbf{RQ}_{4.2}$, i.e., which symptoms or special treatments have been investigated most in the literature.

The included 59 papers span a period from 1993 to 2022, with almost 90% of them being published in 2017 or later. Only one of the reviewed publications focused on the collection of requirements without utilizing any form of digital prototype but through a focus group [286]. The other papers covered a spectrum of VA implementations, from non-interactive screen mockups and videos (e.g., [93, 188]), through functional prototypes that could be tested in a laboratory setting (e.g., [37, 156]), to fully developed applications that could be tested at home (e.g., [161]).

10.3.1 Technology

The technology used to display the agents was not always specifically reported, and in many cases a computer or other screen technology was used. Sixteen papers featured an agent displayed on a mobile device, such as a smartphone or tablet. Furthermore, only five papers used immersive technology, such as AR [156, 225] or VR [37, 220, 248].

We also investigated the level of intelligence used to realize conversations between agents and users. Two papers report the usage of ML and describe the corresponding training process [242, 304]; however, in one of these papers the ML model was developed only after the user study had been conducted and thus was not evaluated by users [242]. One paper reports the use of "artificial intelligence", but without mentioning further details [105]. Multiple papers use some scripted dialogue systems to create pre-defined dialogues with fixed storylines (e.g., [23, 51, 78, 264]). However, most of the investigated papers do not explicitly report which technologies were used to create conversational agents. It should be noted that user studies frequently employed the Wizard of Oz approach, a setup which is operated by the experimenter but seems to be autonomous to the participants, to simulate intelligent behavior and natural conversations [196, 318].

10.3.2 Agent Representations

During our review, we observed a variety of representations for MVAs. Since some papers compared more than one agent or agent entity, and some others only discussed the theoretical idea of an agent, the numbers in this paragraph do not add up to 59. 51 papers describe humanoid agents, seven animals, two robots, and three agents that are represented by other entity types (plane, flower, fruit-people). Gender is explicitly reported as female in 29 papers and as male in 19 papers. The ethnicity of the humanoid agents is only explicitly reported in eight papers as Black [52, 238, 248, 348, 349, 354, 364, 385], seven papers as White [9, 52, 89, 248, 354, 385], one paper as Mediterranean [9], one paper as Hispanic [212], and one as Asian [9] while two papers present ambiguous ethnicities [267, 298]. It should be noted that in most of the papers that do not explicitly report the ethnicity of the agent, it has a light-skinned appearance.

10.3.3 Application Fields

Focusing on $\mathbf{RQ}_{4.2}$, the reviewed papers cover a wide range of application areas, especially with regard to medical conditions and the corresponding patient groups. To create uniform and medically accurate codes for the applications, we used the *International Statistical Classification of Diseases and Related Health Problems (ICD-11)*, which is maintained by the *WHO*.

Based on this classification, a large proportion of the papers concern the field of mental, behavioral or neurodevelopmental disorders. These specifically include mood disorders like depression [252, 344], developmental speech or language disorders [318, 354], anxiety or fear-related disorders such as public speaking anxiety [159, 220, 344], disorders specifically associated with stress such as the post-traumatic stress disorder [273, 330, 331], the autism spectrum disorder [326], and disorders due to substance use or addictive behaviors [24, 31, 242]. One paper mentions its application for mental health, without limiting its use case to a specific mental disorder [37]. Multiple papers consider factors that could negatively influence the (mental) health status, including stress [332], social exclusion or rejection [98], and psychological or emotional distress [89, 264, 298, 303, 304].

Five papers describe endocrine, nutritional or metabolic diseases, including diabetes [6, 161, 196] as well as overweight and obesity [14, 188]. Six papers investigate sleep-wake disorders [75, 76, 85, 251, 254, 333], and one paper has specialized in physical rehabilitation [263]. Different forms of cancer, classified as neoplasms by the ICD-11, are also addressed in the reviewed literature, including cervical [212], colorectal [52, 348, 349, 364, 385], and lung cancer [248]. Further specific target groups represented in our data are patients with injuries [299], incontinence [267], osteoarthritis [207], balance disorders [225], lung diseases [78] as well as chronic diseases in general [9]. 13 papers could not be assigned to a specific health condition although they investigate VAs in the context of general medical services [51, 61, 105, 156, 238, 272, 381], or to promote a healthier lifestyle in general [286, 322] and more physical activity in particular [7, 23, 93, 108].

Another 14 studies focus on the needs of specific age groups, i.e., children and adolescents [6, 9, 93, 108, 161, 188, 196, 318] or older adults [7, 51, 207, 225, 286, 322]. The latter group is particularly encouraged to pursue a healthy, more active lifestyle, while studies with younger patients often focus on gamified approaches, for example, to teach strategies for adhering to a defined treatment plan.

10.3.4 Agent Tasks

Separate from the application fields, we identified several tasks that were performed by the agents. During the inductive coding phase, we grouped these into the following categories, considering only the primary task of each agent unless separate use cases are described in the same paper:

- *Education (n=11)* comprises activities that primarily serve to provide information without in-depth examination of the user's health status. This covers agents that inform about certain diseases and thus want to encourage the user to take action, for example, by signing up at a real clinic.
- *Motivation (n=7)* includes methods to remind and encourage the user to follow healthy behaviors or adhere to a defined treatment plan.
- *Screening / Interviewing (n=6)* is defined as a systematic, often preventive, examination of the user, which comprehensively records their current state of health. While in real healthcare, screening – as opposed to interviewing – may involve additional medical tests, the two terms were used interchangeably in the reviewed literature on VAs.
- *Monitoring (n=7)* describes routinely tracking of one or more aspects of the user's health status, for example in response to a prescribed medication or treatment plan.
- *Diagnosis (n=3)* is usually preceded by screening of the patient, but adds another step by inferring a disease or health condition from the collected information (e.g., symptoms).
- *Treatment (n=25)* includes all activities that incorporate the patient's current health condition and apply therapeutic methods tailored to it. Agents are assigned to this category if they follow the same therapeutic steps as corresponding real-world medical professionals. As in real-world healthcare, the reviewed treatments differed in terms of their format (e.g., psychotherapeutic counseling for mental health vs. physiotherapeutic exercises for physical health) and their scope (e.g., brief therapeutic advice vs. cognitive behavioral therapy).
- *Training (n=2)* of the real doctor's appointment with a virtual doctor is proposed in two papers [51, 207]. The intention is to empower patients, particularly older ones, by teaching them how to better express their pain and how to actively participate in decision-making.

Multiple papers explicitly mention a hybrid approach where the VA and a real medical professional share responsibilities. Support from a real medical expert can be beneficial during the preparation phase of the actual treatment, for example by creating individual dialogues [267, 303] or treatment plans [354]. During treatment, human intervention was systematically used when the VA detected an acute crisis situation, such as suicidal thoughts [330, 333].

10.3.5 Agent Roles

In the scope of literature we reviewed, agents in the medical context presented different roles. Multiple agents represented the role of professional staff, i.e., doctors or therapists, especially if they perform tasks which are usually done by real professionals, like treatment [254, 264, 304] or screening [24, 105]. Besides, various agents deviated from such traditional roles and were portrayed as less professional, thus creating a spectrum of professionality, which ranged from virtual doctors on the one end to friend-like companions on the other. For instance, agents were referred to as coaches in multiple occurrences. While these could provide professional treatment, they did not necessarily have to have a medical appearance, for example, in terms of clothing [225].

Various other agents found in the literature could be described as personal assistants. They could also perform tasks like treatment, but often in a more personal way, for example, by supporting users in their everyday life by being available regularly or on demand [98]. Oftentimes, personal assistants were used for regular but less complex tasks, like monitoring sleep behavior [76], supporting self-management for patients with chronic diseases [6], or motivating healthy behavior [286].

Some agents were explicitly designed to represent the role of a companion or friend, mainly if their target group consisted of children or adolescents. Here, empathetic humanoid agents or playful animals were investigated to encourage young people to follow health and well-being advice [9] and to set self-determined physical activity goals [108]. In the context of physical activity, a study's results also suggested that a friendly agent that creates a parasocial connection could increase children's intrinsic motivation [93]. Furthermore, agents designed to be perceived as friends were utilized to educate children playfully [196]. Even though agents for adult patients were usually not explicitly designed as friends, they could still be perceived as companions by users, which was positively mentioned [78, 251] and even resulted in the desire for agents to be represented as friends [238].

Besides the spectrum of medical professionality, some publications reported the desire of users for agents to take on different roles. For instance, during a requirement analysis for agents to support teens regarding weight management, participants indicated that the agents could serve as teachers, describing this role as a knowledgeable entity that could teach about healthy habits and behaviors [188]. Also, members of a church community could imagined agent personas as members of the church staff [238].

10.4 Results of the Thematic Analysis of User Feedback

Focusing on the analysis of user feedback using thematic analysis, we especially intend to address our research questions $\mathbf{RQ}_{4.3}$, i.e., reasons for and against using MVAs, and $\mathbf{RQ}_{4.4}$, i.e., conditions that must be met for users to accept MVAs. Due to the early state of reviewed research, we mostly report preferences on the representation of MVAs here. All the subsequently presented results are based exclusively on empirical findings

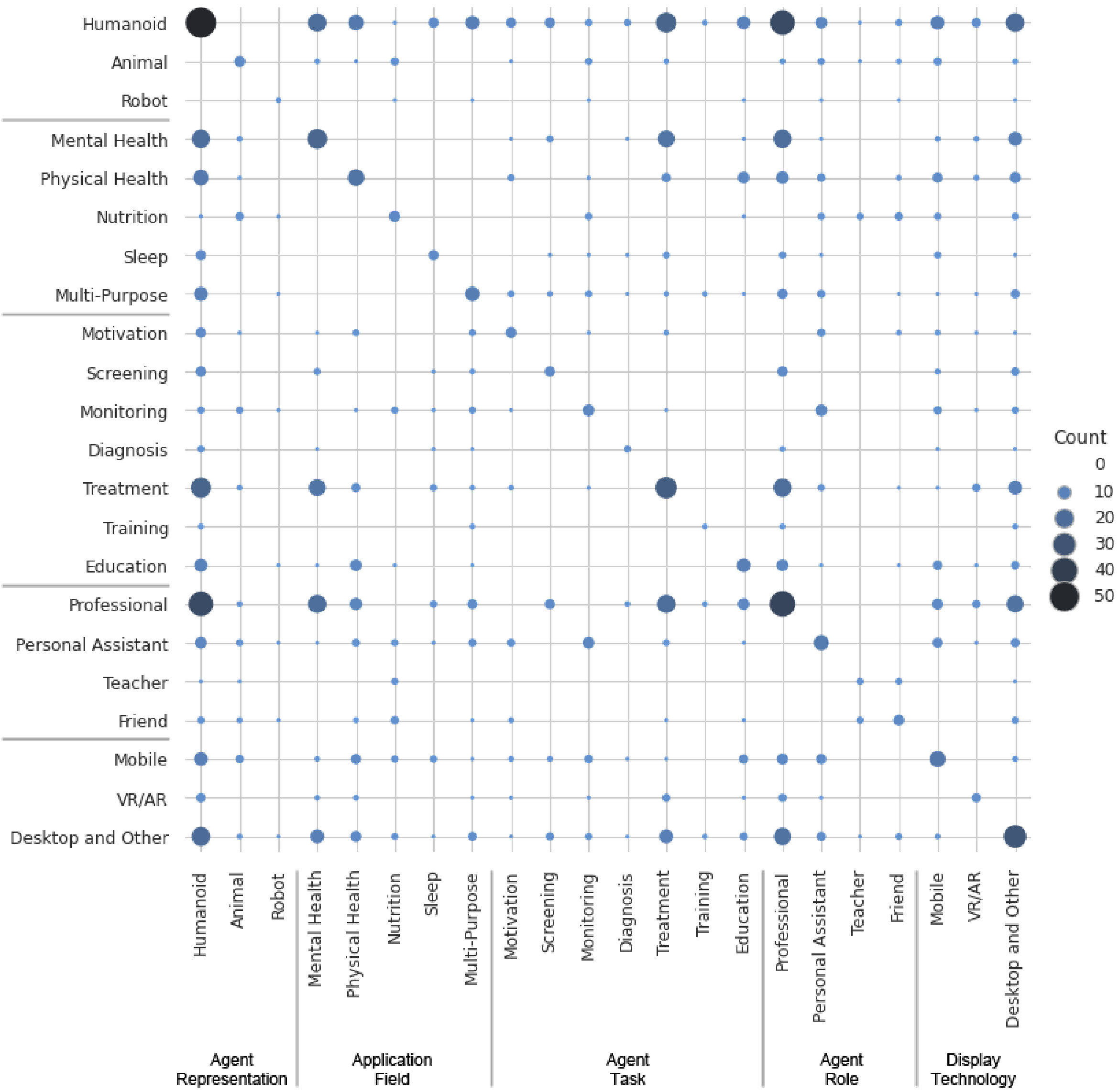


Figure 10.3: Plot showing the frequency of combinations of representations, application fields, agent tasks, roles and technology.

from the user evaluations that are described in the reviewed papers. Non-validated hypotheses of the authors, for example as part of the motivation or interpretation of the study results, were explicitly excluded from the thematic analysis.

10.4.1 Positive Aspects of Medical Virtual Agents in Comparison to Real Doctors

Availability and Accessibility

Two of the most commonly stated reasons for using MVAs were *availability* and *accessibility*, which can provide various benefits for different agent roles. VAs who take over tasks that would usually be performed by a doctor or therapist can be useful when a real doctor's visit is not possible. Study participants perceived a virtual doctor as particularly beneficial for patients who are unable to see a real doctor due to time or financial constraints [264]. Furthermore, the burden of travel was mentioned as a reason why people do not seek human care [333]. Participants who tried a virtual doctor while on a waiting list appreciated the possibility to get advice more quickly than from a specialist, and could even imagine using an agent to bypass the need to consult a human doctor [267]. When presented with the concept of a virtual therapist, patients in a psychiatric ward described dissatisfaction with medication being the primary form of treatment and expressed the desire for more conversational therapies [37] – both aspects that could potentially be addressed by an MVA. Participants also acknowledged the ability of agents to provide support during the day and just in time, whereby patients can easily consult them on demand [14].

Customization and Tailoring

Participants frequently expressed the desire to be able to customize agent characteristics or praised such a feature if presented agents included it. They positively mentioned the ability to change the agents' appearance, for example, regarding visuals, age, and gender [14, 37, 78, 225, 286, 299, 344, 348], for example to meet personal preferences. Besides visual attributes, participants praised agents which offered multiple input and output modalities [78, 364], pointing out the diverse needs of different users [344].

In addition to the ability to actively customize agents to personal preferences, tailoring of agent attributes was also evaluated as a positive aspect of virtual doctors. For example, in contrast to a real doctor, an agent's behavior may be tailored to the user's cultural background. In the context of a church community, participants reported that they would be more willing to listen to and follow advice from agents if they reflected their own character, i.e., by using language with a matching level of spirituality [238]. In other studies, participants suggested tailoring the agents' conversational style to match the users' needs, for example, by adjusting the pace of interaction [298]. Additionally, the agents' behavior could be tailored to how long patients have already been in treatment [242].

Furthermore, participants voiced the wish to interact with agents that are portrayed as peer professionals, that is, professionals with similar experiences as the patient. Patients with psychological disorders endorsed the idea of an agent represented as a person with similar mental health problems but trained to be a therapist, indicating that the simulated similarity could have a relaxing effect [37]. Correspondingly, individuals with spinal cord injury reported their preference for the VA to be in a wheelchair [299].

Confidentiality

In multiple papers, the inherent characteristic of agents – in contrast to human beings – to be non-judgmental was positively mentioned by participants, especially for agents in the field of medical problems that require patients to disclose sensitive information, such as depressive disorders [252], or alcohol misuse disorder [31]. Participants expressed that they felt more comfortable sharing information with agents, even though they knew that this information would be viewed by humans later [24]. A pilot study also suggested that agents could be used as communication coaches to increase the amount of information patients share with their doctors. Older adults reported more pain information to their real doctor when they practiced the communication with an agent before [207].

Quality of Information

Participants commended various aspects regarding the quality of information provided by medical agents. They perceived the agent's information about colorectal cancer as accurate, unbiased and relevant, and acknowledged that agents have additional access to information [349]. Furthermore, participants expressed that information delivered by agents was based on facts and studies and thus reliable [31]. Also, the source of information used by the agent can influence the perceived quality. For instance, by disclosing a local healthcare provider as the source, the agent's information was perceived as trustworthy by participants [385].

10.4.2 Negative Aspects of Medical Virtual Agents in Comparison to Real Doctors

While most papers present a generally positive view on the agents, in a few examples, the users disliked the agent and its therapy due to *interpersonal reasons*, describing it as weird [9, 344], hard to relate to [344] or not suitable for everybody [37, 332, 344]. Users criticized that VAs are simply *not humans* [298] and cannot manage *crisis situations* [31]. Also, the fear of *replacing human doctors or cutting their funding* was voiced [37]. Besides this, reasons and factors against using virtual doctors were mostly related to limited technical implementations, including *wrong or missing information or treatments* [6, 7, 51, 161, 344], *technical difficulties* with voice detection [31], *limited interactivity* of the application or dialogues [51, 344, 348], and *limited realism* of the agent [264, 344, 348, 349, 364].

10.4.3 Prerequisites for Consulting a Medical Virtual Agent

To mitigate some of the negative aspects and fears of the users, we have compiled a set of formal requirements from the literature that MVAs – and the application they are embedded in – must meet to actually be consulted by users. First, such applications must fulfill high *privacy and data safety* standards with controlled access to data and no commercialization [37, 225, 286]. Furthermore, they should be *accessible* [37, 78, 286], *affordable* [333] and *validated* by credible institutions or therapists or by using evidence-based data [7, 14, 344, 348, 385]. Another crucial aspect is the *control* of application features, such as notifications [161], and that the choice in treatment decisions is left to the user [31, 385].

10.4.4 User Preferences for Agent Representation

To address our research question $\mathbf{RQ}_{4.4}$, we collected preferences of users for the representation of MVAs, including its *entity type*, *embodiment*, or *realism*.

Entity Type

Only a few papers compared multiple entity types of an agent, with a machine-like character being perceived as more likable than a humanoid one in a healthy lifestyle intervention [322] and inanimate objects as well as animals being less preferred than humanoid agents in social training for people with autistic traits [326]. For children, two studies examined their preference of agent type and found that they favored a physical robot over its virtual representation for diabetes education [196], as well as a physical toy lamb over a virtual lamb or a human on a tablet screen for linguistic assessments [318].

Embodiment

Findings on the effect of embodiment differed for use cases and between participant groups. Kim et al. concluded that a real caregiver provided the optimal user experience, but an embodied virtual assistant was also a viable option for patient care environments [156]. It was also found that a video-taped real person was more effective in a medical interview than a VA [381]. In another study, a virtual doctor increased adherence intention and general medical consultation enjoyment more than a recording of a real doctor [61].

Comparing an embodied agent to a voice or text interface resulted in similarly mixed results. In a therapy setting, the bond between an embodied agent and the user was stronger than with a text-based agent [220], and embodied agents provided higher social presence and engagement than voice-only agents [156], which is in line with previous research results in other contexts [155, 293]. Also, an embodied agent led to a stronger intention to learn more about colorectal cancer screening compared to text-only [385].

On the other hand, an embodied agent was perceived as less competent than a voice-only character as a health-related daily life assistance [322], it was harder to share personal information with an embodied agent compared to a voice-only agent [156], and some participants suggested to remove the face and voice of the agent altogether and to only have a text interface in a digital mental health platform [344].

Realism

Several factors went into the question of how realistic an agent should be portrayed, and again findings from the literature were mixed. According to Dai and MacDorman [61], realism does not necessarily lead to a higher usage intention or enjoyment of digital doctors. While some older adults had the highest usage intentions for a cartoon-styled agent [322], a realistic representation was preferred by others [225] since a cartoon style can make the user feel patronized and childish [299]. Wilson-Howard et al. [364] report an early, cartoon-like version of their agent to be not human-like enough, therefore leading to diminished credibility. Venning et al. [344] also reported that their agent tried "*too hard to be human*". Younger users, such as students, reported the lowest usage intention for a cartoon-styled agent [322], while adolescents had very diverging opinions on how their weight management agent should look like, including preferences for human-like characters, cartoons, animals, or even fantasy characters [188].

10.4.5 User Preferences for Agent Appearance

Visually, users commented on the *ethnicity*, *age*, *gender*, and *appearance* of MVAs.

Ethnicity

Only one study asked about preferred ethnicity in relation to the participants' and found that of eleven teenagers, eight preferred the same ethnicity, one preferred a different ethnicity, and two did not have a preference [9]. From these participants, nine were White and two were from Oceania. There were also some studies which evaluated the effect of agent ethnicity on certain measures. For Black study participants, race-matching positively influenced message relevance [52] and the participants' lung cancer risk perception [248]. It also positively but not statistically significantly influenced intention to be screened for colorectal cancer [52]. This was not the case for White participants. Cooks et al. [52] explain this in the following way: "*For groups that have not endured these historical inequalities, there may be less uncertainty surrounding clinician credibility, thus reducing the impact of this factor within these populations*", which further highlights the importance of having non-judgmental, adjustable medical agents that all user groups feel comfortable with.

Age

The age of agents was commented on in only three papers. Preference was related to the age of the participants in two of these papers; adult participants rated agents close to their age (e.g., in their 40s) as trustworthy for social skill training [326], whereas of eleven adolescents, seven preferred characters of the same age and four preferred older characters for living with chronic health conditions [9]. Participants in a study by Vilaro et al. [348] preferred agents that were not too young and not too old because they were concerned that a younger doctor might not have enough knowledge and an old one might not be up to date on recent medical developments.

Gender

Aside from the wish for customization of gender [78, 286, 348], a few papers reported a preference for a specific gender, with female agents being favored over male ones (out of eleven participants, six preferred female, one male, four did not care) [9] and a female agent being preferred by male participants [326]. Mostajeran et al. [225] found that for physical training, a realistic female representation was preferred over a realistic male one, and a cartoon female over a cartoon male. The gender of the agent also had direct effects on the interaction of students, who were less inclined to interact with a male virtual human than a female one [89], however, this did not extend to the expertise perception of the agent. Finally, five participants commented on the male gender of a doctor in a study by Constantin et al. [51], wondering why it did not reflect the reality of doctors in the UK, where the majority is female.

Appearance

In some papers, the participants commented on the look of the agent, for example, their clothing style [31, 238, 349, 364, 385]. Some participants found that casual clothing was incongruent with the role of the agent – in this case someone they would turn to for guidance in their faith [238] – and that the agent should dress accordingly, for example, with a white coat or medical attire [349, 364]. On the other hand, some commented that the VA was too clinical [344] or that the introduction of the agent as a doctor set user expectations to look and behave like one, which were not met [385]. Furthermore, it was found that attractiveness of the agent was a significant predictor for screening intention for colorectal cancer [52].

10.4.6 User Preferences for Human-Agent Interaction

Regarding human-agent interaction, users specifically commented on the agent's *personality*, *behavior*, and *communication*. For these categories, a distinction can be made between the users' opinion on MVAs, which highly depends on the evaluated implementation, and the users' wishes towards them, which can imply important personality traits.

Attitude and Personality

Only a few papers reported specific wishes in terms of the agent's *attitude* towards the user or its *personality*. Participants in a study by LeRouge et al. [188] wished for their agent to provide empathic support and guidance for weight management. Vilaro et al. [348] reported that users desired a friendly and likable agent, and participants in a study by Mostajeran et al. [225] would have preferred a friendlier rather than a neutral expression from the tested agent. Furthermore, for some adolescents, it was important to be able to identify with the agent and that agents are "*unafraid to be themselves*" [188].

Behavior

Regarding the agent's *behavior*, study participants noted the importance of positive reinforcement versus criticism (for weight management) [188]. Participants in the same study also indicated that the agent should ask them questions to gain better insight into their individual eating behavior. Related to weight loss, users preferred a coach that supports habit formation, prompts daily rehearsals, and brings a sense of joy and happiness to the participants [14]. Some participants wanted the agent to be reactive and not proactive, in order to not be too intrusive, but at the same time convincing and motivating [286]. The authors also stated that the agent should be respectful and not force the users to do something. A participant in a study by Constantin et al. [51] suggested to let users choose whether they would like to lead the dialogue or if the agent should do so.

Communication

Participants' recommendations for the human-agent *communication* were to tailor the agent's language to both the treatment and the user [242]. Motivational messages outperformed ones without such content [331], and personalization of the messages had a positive effect on how much people felt heard and taken seriously [331]. Furthermore, the speed and clarity of speech is important, with participants requesting a slower or faster speech [212, 225, 348] or adjustable subtitle speed [349]. The speech needs to be clear and well-articulated [225], especially for older adults. Multiple modalities, such as audio and subtitles, were also useful for study participants and improved their user experience [348, 364]. Participants even particularly suggested adaptations for individuals with visual or audio impairments [37]. Moreover, the option to have only text instead of voice output for a mental health platform was suggested [344]. On the other hand, for older adults, a preference for using voice input over clicks was reported [51], as it decreases physical effort. A preference for a human voice was reported by Shamekhi and Bickmore [298], but a synthetic voice was also rated as acceptable. While the voice in some studies already sounded human [31, 78], in others, the participants wished for a more natural sounding voice [225, 286]. In multiple studies, the limited interactivity due to scripted dialogues was pointed out [6, 37, 51, 242]. This was especially the case when multiple-choice menus instead of chat or voice interfaces were used, as they were not natural and flexible enough to shape the discussion and ask all the participants' questions. Finally, it was also suggested to add the option to repeat instructions [225].

10.5 Discussion

In this section, we will discuss the implications and limitations of current research in relation to our research questions and point out potentials for future work. These will be connected with considerations for VR/AR technologies. Furthermore, we will reflect on the limitations of this review.

10.5.1 Technology ($\mathbf{RQ}_{4.1}$)

Our analysis of previous literature has revealed that embodied MVAs are a fairly new concept in empirical research, which has only gained increased attention within the past five years. Many of the reviewed research prototypes are still subject to certain technical limitations, such as communication problems or limited realism of the agent, which can introduce bias in the research results. For example, audio-visual limitations could result in a more negative perception of the agents, since they are very prominent in some studies [31, 264, 344, 348] and disguise the efficiency of the treatment or intervention. While advancements in algorithms (such as natural language processing) and graphics hardware are expected to alleviate these technical constraints, thorough user testing of the prototypes remains important.

In our review, we observed that only three papers acknowledge the potential of using ML or AI approaches to enhance the VA's responsiveness to user input, in particular by recognizing the user's health status [304]. The still low prevalence of corresponding systems may be attributed to various aspects, including the non-deterministic dialog flow (and thus a potentially reduced comparability between study sessions), the complexity of training custom ML models, high computation times, which are reflected in a correspondingly higher reaction time of the agent, and last but not least the risk of erroneous outputs. While the first aspect is inherent to ML-driven dialogues, the latter have seen tremendous development in recent years due to improvements and the increasing availability of both speech recognition models and powerful pre-trained language models such as GPT-4 [243] or OPT [389]. Therefore, an increase in the use of ML models can be expected in research as well, which would open up interesting possibilities for fully automated, flexible medical agents. Nevertheless, the danger of erroneous output persists, which can have grave ethical consequences [277], e.g., when a misdiagnosis is given or when the safety of a patient cannot be guaranteed during an emergency situation [199].

In our literature search, we also found that 16 papers do not specify the intended display technology, which may be in part because they are still in an early stage of development. In addition, the papers that do describe a specific technology mostly focus on (desktop) screens (n=24) or mobile devices (n=16). There are only five papers considering AR [156, 225] or VR [37, 220, 248] displays, and none of the 59 considered papers was drawing a comparison between different output technologies. The few VR/AR studies have already indicated a positive therapeutic effect of immersive virtual environments, for example, by giving patients the feeling of being in another place [37]. In addition, we know from previous research on interactions in VR/AR that immersion has a positive impact on telepresence [240], which may also affect other measures related to human-agent relationships. While prior research has shown that preferences on agent visualization remain similar between technologies (e.g., AR/VR) [360], we still believe that the results from desktop/mobile setups should be confirmed in future user studies to account for the technical differences in VR/AR, e.g., stereoscopic vision, resolution or color fidelity. Furthermore, the suitability of VR/AR for the medical context has to be investigated, for example regarding availability and usage context – it is still relatively expensive and most potential users will not have these systems already at home. Furthermore, using them in public places is not very common due to privacy and safety reasons. Finally, the accessibility of these systems needs to be tested, especially in the medical context. This includes consideration of suitability of VR/AR for specific health conditions, such as epilepsy, or for people with pacemakers, for whom the usage is not recommended.

10.5.2 Usage and Tasks of Medical Virtual Agents ($\mathbf{RQ}_{4.2}$)

The reviewed papers cover a wide spectrum of medical applications, including distinct health conditions such as chronic obtrusive lung disease and diabetes. One surprising finding of our review was that the majority of papers included at least some representatives of their actual target group in the evaluation (n=41).

Within these medical applications, the agents' tasks can mainly be characterized by being non-decisive and requiring only limited professional competencies, for example, educating users about the consequences of certain diseases or motivating them to adhere to healthy habits. VAs that diagnose specific health conditions are considered in only three papers [61, 75, 252]. Based on the present body of data, it is not possible to decide beyond doubt whether this is a result in itself suggesting that VAs are specifically suited to assist real medical personnel in low-threshold tasks and are accepted as such, or whether current studies are just biased in that direction by omitting more complex tasks. This is especially interesting in combination with VR/AR technology. In non-medical contexts, most VR/AR agents are used in educational scenarios, (e.g., as a museum guide [249] or an instructor for repairing scenarios [41]), or as assistants [266, 358], which is in line with our findings for medical VAs. While several theoretical arguments support the first explanation, we found some subjective views that VAs have more diverse and potentially more reliable sources of information than real medical staff, and that they are non-biased. Moreover, two of the related papers indicate high acceptability and/or trust of an agent diagnosing either major depressive disorders [252] or excessive daytime sleepiness [75]. Similar positive results on trust are reported by Persky et al. [248], who explored an agent that calculates and communicates an individual's cancer risk, which can be considered as a pre-stage to diagnosis. Based on these preliminary results, further empirical studies are necessary to draw a definitive conclusion about the responsibilities users would assign to VAs in medical settings.

While the design of the agents to act as professionals is most prominent, only five papers designed the VA to act as a peer or companion, and all of these were aimed at children and adolescents. However, adult participants also repeatedly praised companion-like behavior of the agents they interacted with. Furthermore, participants expressed their wish for agents as peers, i.e., by representing individuals with similar physical or psychological attributes, but none of the agents presented in the reviewed literature fulfilled this request. Thus, we see the usage of personalized agents as an opportunity to strengthen the social bond between patients and medical characters, in contrast to mirroring the traditionally fixed roles of doctors and therapists.

10.5.3 Reasons for and Against Visiting Virtual Doctors ($\mathbf{RQ}_{4.3}$)

Participants expressed a desire for customization and tailoring of MVAs. They valued the ability to personalize agent characteristics, including appearance, age, gender, and communication style. Tailoring the agent's behavior to cultural background, conversational needs, treatment duration, or representing peer professionals with similar experiences was also seen as positive. Especially the control over input and output modalities is an interesting use case in our opinion, as it allows patients to utilize the VA according to their specific needs. It empowers them to engage with the agent in various settings, such as their home environment, where they can employ a natural verbal communication style, or in public spaces, where they can opt for a written chat or multiple-choice responses. This flexibility not only ensures privacy but also enables users to confidentially receive medical advice tailored to their preferences.

10.5.4 Conditions to Create Social Acceptance of MVAs ($\mathbf{RQ}_{4.4}$)

Due to the early stage of (comparative) research in many of the papers, or the different research focus, we were unable to answer this research question as intended. Instead, we report the preferences of users regarding the representation of medical VAs and found that they show mixed results in almost all of our measured scales. While Weidner et al. [360] concluded that realistic visuals are generally more suitable, studies in the medical domain show that some user groups, e.g., children or some older adults, prefer cartoonized representations. Similarly, embodiment did not always lead to the best results in terms of competence or openness. Our analysis indicates that a realistic representation might not always be the most suitable and that preferences are dependent on the use case and user group. In this scope, it is also interesting to look at VR/AR agents again. In order to increase their realism, special care has to be taken, due to the three-dimensional nature of VR/AR content. Users in these setups are often able to change their perspective and move in space – therefore the agent has to respond realistically to these changes and it is possible that the agent moves out of the field of view of the user. This differs from 2D displays, where the perspective of the user is mostly static, which might be easier to implement. We also hypothesize that, independent from technology, some *comfort functions* could be implemented to improve the user's experience, e.g., customization of the agent, the option to slow down or repeat conversations, or the option to switch between communication modalities. Nevertheless, due to the lack of comparison studies, it might also be interesting to look at VAs that were not tested in the medical domain to receive a broader picture of agent preferences of the general population.

10.5.5 Limitations of the Literature Review

This literature review is associated with some limitations. For instance, we only searched three databases: ACM, IEEE and PubMed, since these are the three largest ones for our research field. Nevertheless, other databases might yield new results. In our literature search, we came across a variety of terms describing the same concept, for example, *virtual agent / human / coach / therapist / clinician / medical specialist*, but also *interactive coach*, *digital health agent*, or *computer-animated doctor*. While our search terms were sufficient to capture all of these expressions, there may be other ways to reference the targeted topic that consequently prevented inclusion of the respective papers. Similarly, we only covered publications in English. Including different languages might be a way to better reflect cultural differences in agent preferences. In our initial search results, we also found some agents who were only text- or voice-based. While we specifically looked at embodied agents in our review, chatbots or voice-agents might take on different roles than embodied ones, with potentially more powerful AI, different technologies or advanced application areas. It would be interesting to see whether embodied agents will also advance in their responsibility in the future. During the eligibility check, we came across papers that had to be excluded due to the formulated criteria, but nevertheless showed some overlap with the investigated research topic and therefore might be worth considering in future work. In particular, consultation with a real doctor via digital platforms shares similarities with interacting with an MVA in terms of a subset of factors, both positive ones (e.g., increased availability) and concerning ones (e.g., cautionary handling of data access) [214]. Therefore, findings from the field of telemedicine might be applicable to MVAs as well.

Although we made an effort to reduce subjective bias to a minimum by involving at least two researchers per paper in the screening and data extraction stages, and by performing subsequent steps such as thematic analysis in a group discussion, there is still a potential influence of the researchers and their individual backgrounds on the results. In particular, all participating researchers work in the field of human-computer interaction (with three of them having prior experience in the health domain) and therefore may have a different perspective on the topic than, for example, a medical expert.

10.6 Conclusion

In our systematic literature review, we analyzed 59 research papers on embodied virtual agents in healthcare, with a focus on the user preferences in terms of agent representation, behavior as well as underlying implementation structures. We found that researched agents predominantly took a humanoid form and were most commonly employed for some kind of patient treatment, including basic therapeutic counseling, full conversational therapies, but also physiotherapeutic treatments. Other investigated tasks include education, motivation, and screening. In most papers, VAs were depicted in a professional role and presented on a non-mobile screen. On the other hand, agents acting as personal assistants were most often used with mobile devices. This may underscore the novel role that MVAs could embody, extending beyond the scope of real doctors. Many positive attributes of agents were identified, such as being available, accessible, and non-judgmental, but also some negative aspects, such as their technical limitations and inability to act adequately in certain scenarios, e.g., crisis. To address preferences of different user groups, the importance of customization was pointed out. Finally, we identified the following research gaps and design considerations, which are especially important for the VR/AR community:

- investigating and testing the diversity of display technologies: their availability, accessibility and usage context,
- considering the diversity of user groups: including consideration of suitability of VR/AR for specific health conditions,
- exploring the roles of agents: assistive or decisive,
- ensuring (behavioral) realism of MVAs: especially in 3D space.

Our systematic literature review can serve as a basis for researchers in this field to facilitate design decisions for medical virtual agents in future projects.

Chapter 11

Comparing Agent Visibility in Cognitive-Physical VR Training

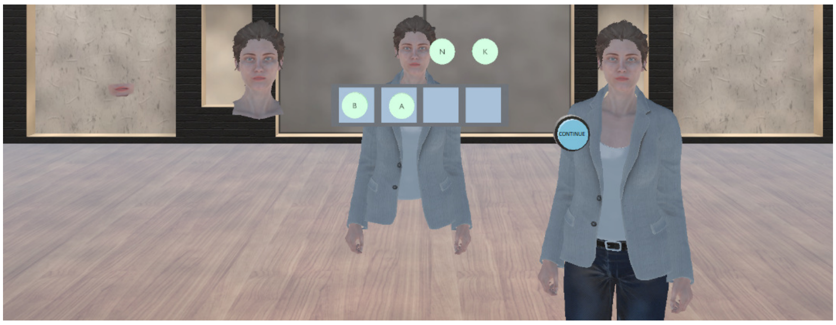


Figure 11.1: The agent representations in front of a half-solved anagram, from left to right: Voice (invisible), Mouth-Only, Head, Upper Body, Full Body.

Our systematic literature review (See Chapter 10) showed that opinions on IVAs differ depending on the users, the use case, and the context. Therefore, it is important to evaluate the representation of agents depending on a chosen task and application. As a first attempt, we integrated an IVA that differed in terms of visibility into a VR exergame where anagrams had to be solved. Further, again due to the unavailability of the actual target group, we evaluated the application with younger users, aiming to identify structural differences in the agent representation which might be transferable to older adults as well (See Chapter 7). The goal of this user study was to identify whether the visibility of an agent has a significant effect on the task performance in immersive cognitive-physical training, and which representation is preferred by users. This chapter is based on the following publications, where the first one is the full paper and the second one is a reduced poster version:

Kruse, L., Mostajeran, F., and Steinicke, F. (2023). The Influence of Virtual Agent Visibility in Virtual Reality Cognitive Training. In Proceedings of the ACM Symposium on Spatial User Interaction (pp. 1-9). DOI: https://doi.org/10.1145/3607822.3614526

Kruse, L., Mostajeran, F., and Steinicke, F. (2023). High Levels of Visibility of Virtual Agents Increase the Social Presence of Users. In IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) (pp. 843-844). DOI: https://doi.org/10.1109/VRW58643.2023.00264 (Poster)

11.1 Introduction

VAs have been integrated into various applications to play a variety of roles such as virtual coaches [227], assistants [155], and health professionals [347]. They come in different shapes and representations, ranging from text [384] and voice-interfaces only [259] to fully-animated realistic 3D reconstructions of real humans [347]. A recent systematic review has also identified various configurations of showing different body parts of avatars and agents in research, e.g., hands-only or hands and torso [360]. Previous research has shown that exposure to and/or interaction with VAs could result in *realistic*, i.e., similar-to-reality, social responses in humans. For instance, VAs can be used to elicit emotions such as social anxiety [223], embarrassment, irritation, and self-awareness [312] that humans usually demonstrate in response to other real humans.

Several studies have also shown that similar to performing tasks in the presence of real humans, the mere presence of a virtual agent could influence individuals' performance. This can be observed in the form of an improvement in task performance, which is known as *social facilitation* or a deterioration of the performance known as *social inhibition* [383]. For example, Miller et al. [218] observed an increased performance (i.e., social facilitation effect) for easy tasks and a decreased performance (i.e., social inhibition effect) for hard tasks in the presence of a virtual agent in an AR setting.

On the other hand, it has been suggested that the visual representation of virtual agents may have an effect on users' psychological responses to them. For instance, Nowak and Biocca [237] observed that users show different social responses to different representations of an IVA with varied levels of anthropomorphism (i.e., the extent to which a virtual agent looks like a real human). Another study showed that the users' confidence in an agent's abilities increases when it has a humanoid representation and demonstrates social behaviors [155]. Also, the visual representation of an agent may play a role in users' attitude towards them. For example, in a remote collaboration scenario, Yoon et al. reported that a realistic full body representation of a virtual avatar was preferred by users over an upper body or cartoon style representation [380].

In a recent literature review by Weidner et al. [360], the authors collected (among other) information on the preference of users on body part visibility of avatars and agents. Avatars here are visualizations controlled by real humans, where agents are completely guided by computer programs. For avatars, it is an established visualization to only see incomplete humans, e.g., the user's head and/or hands, or head and torso, which are also widely accepted. However, the preference for agent visibility, which was only covered in 17 papers identified by Weidner et al., may differ. Thus, our study aimed to explore this aspect; in particular the level of body visibility on users' social presence or task performance. Therefore, we designed and conducted an experiment to study participants' perception of social presence towards a virtual agent with varied levels of visibility. In addition, we studied participants' task performance in the presence of each representation.

For this, we implemented a virtual female agent in five different levels of visibility ranging from the lowest level (i) voice-only, (ii) mouth-only, (iii) head, (iv) upper body, to the highest level (v) full body. In a within-subject user study, participants solved anagrams in three levels of difficulty (i.e., easy, medium and hard) in the presence of each representation of the virtual agent in a VR setting. This paper serves as an extension of our previous work [173], additionally presenting eye tracking and more detailed performance results. Our research question was to find out which agent visibility representation is the most suitable for immersive training applications $\mathbf{RQ}_5$.

11.2 Related Work

Oh et al.[240] reported in their systematic literature review that one of the features determining social presence (i.e., the feeling of being present in the virtual world with a real human) is the visual representation, including the degree of visual realism, of the communication partner. According to Oh et al., visual realism comprises of photographic, anthropomorphic and behavioral realism. In their definition, photographic realism describes how realistic an IVA appears, anthropomorphic realism determines to which extent an IVA looks like a real human, and behavioral realism describes the degree of similarity of an agent's behavior to an actual human behavior.

There are a few studies which have varied visual realism to study its effects on humans. For example, to visualize the avatar of a remote collaborator, Yoon et al. [380] compared two different levels of photographic realism (i.e., (i) cartoon and (ii) photo-realistic) in three different levels of anthropomorphic representation (i.e., (i) full body, (ii) upper body and (iii) head plus hands). They observed that a realistic full body representation was preferred over an upper body or cartoon style representations. In their study, participants rated the full body representation with the highest social presence scores, while stating that the other representations could be also suitable for other scenarios. In the context of training at home, realistic representation of VAs have been reported to be preferred by older adults over cartoon representations [226, 227].

Some studies have made comparisons between embodied and voice-based VAs. For instance, Kim et al. [155] showed that when a VA has a human body representation and shows social behaviors, users feel more confident in the agent's ability to perform tasks compared to a voice-only agent. On the other side, they found sharing personal information with an embodied VA to be harder than sharing them with a voice-only agent. In addition, participants of some other studies reported a higher sense of social presence and engagement with embodied VAs compared to voice-only representations [157, 222, 293].

There is only little previous research about the effects of different representations on task performance. Beun et al. [22] demonstrated that the presence of a realistic virtual woman's head leads to significantly higher scores in a memory test compared to a text-only representation. However, they could not observe any significant differences between this realistic head representation of a woman and a cartoon gorilla representation. Another study [224] compared five different levels of visibility of a VA in terms of their effects on solving anagrams in an AR setting. In their preliminary study, they found that a more visible representation leads to higher social presence and higher percentage of correctly solved anagrams.

Several other studies have also reported performance increase and decrease in the presence of VAs. For instance, Walker et al. [357] observed that participants provided answers to experiment's questions slower but with fewer mistakes when they were asked by a virtual face on a monitor compared to when they were given through text. In another study by Rickenberg and Reeves [269], participants performed a task on a website either alone or in the presence of an animated character on the website. The results showed a social inhibition effect when the character was perceived to be capable of monitoring the participants' actions.

Furthermore, Park and Catrambone [247] also observed these social effects in relation to not only the presence of a VA, but also the difficulty level of the task. In their experiment, participants performed three tasks in two levels of difficulty (easy and difficult) and three social settings, i.e., (i) alone or in the presence of (ii) a real human or (iii) a VA on a computer screen. The results demonstrated social facilitation for easy tasks and social inhibition for difficult tasks in the presence of the real human as well as the VA. In another study [191], social facilitation was observed in terms of speed when performing an easy task, meaning that the easy tasks were performed faster in the presence of a VA (here, the face of a digital human) compared to performing the task alone. In addition, in a study by Zanbaka et al. [386], participants solved math problems in two levels of difficulty (easy and hard) either (i) alone, (ii) in the presence of a real human, (iii) a projected VA or (iv) a VA in VR. The social inhibition effect was observed when participants performed hard tasks and were in the presence of real or virtual humans. However, no social facilitation effect could be observed, probably due to a ceiling effect. In another study by Miller et al. [218] in an AR setting, participants solved cognitive tasks either (i) alone or in the presence of (ii) an embodied VA. As a result, a social facilitation effect could be observed for easy tasks and a social inhibition effect was reported for hard tasks in the presence of the VA.

Therefore, previous studies have shown that it is possible to evoke social facilitation and inhibition effects in the presence of VAs. However, different visual representations, in particular the level of visibility of the agents, have not been considered as an influencing factor. In this work, we aim to assess the social presence felt by the individuals towards different levels of visibility of a VA in a VR setup, since research indicates that VR can elicit higher feelings of social presence compared to a desktop setup [106]. In addition, we compare participants' task performance in the presence of each representation to study their possible social facilitation or inhibition effects.

11.3 Methods

A total of 30 participants between 19 and 34 years of age (26.33 ± 3.96 years) took part in the study (16 women, 13 men, one preferred not to say). They were recruited through our university study recruitment system and received course credit. Twelve participants were frequent VR users, 15 had used VR a few times before, and three had never used it before.

In accordance with prior work, we established the following hypotheses.

- **H**$_1$: Higher levels of visibility lead to higher perceived social presence of the virtual agent.
- **H**$_2$: The participants look for a longer time at the more visible representations of the virtual agent.
- **H**$_3$: The time each participant spends looking at each representation of the virtual agent is positively correlated with their perceived social presence.
- **H**$_4$: Higher levels of agent visibility lead to an increase in users' performance in easy tasks including (**H**$_{4.a}$) higher correct answers and (**H**$_{4.b}$) shorter response time (i.e., social facilitation effects).
- **H**$_5$: Higher levels of agent visibility lead to a decrease in users' performance in difficult tasks indicated through (**H**$_{5.a}$) lower correct answers and (**H**$_{4.b}$) longer response time (i.e., social inhibition effect).

11.3.1 Application

The virtual environment was implemented in Unity3d version 2020.3.34 with the Pico UnityXR SDK v2.0.4 and TobiiXR SDK 3.0.1.179. We used Character Creator to design the IVA and the iClone plug-in to add a speaking and idle animation to it. For the study, a Pico 3 Pro Eye HMD was used, which has a resolution of 4K with 3664×1920 pixels and a field of view of 98° horizontally and 90° vertically.

There were five different agent representations: (i) voice-only, (ii) mouth-only, (iii) head, (iv) upper body, and (v) full body (see Figure 11.1). Each of the embodied representations had the same animations and used the same voice, and they followed the user with their gaze, similar to Ferstl et al.'s work [91]. Body parts were hidden by simply changing part of the materials to invisible ones. For the mouth and head, idle motions of the body were reduced so they would remain in a stable position. We included a representation of the mouth to provide a source for the speaking sound, although we acknowledge that this choice does not align with realistic or commonly used representations. Nonetheless, for the sake of completeness, we decided to incorporate this condition in our study. Our agent was designed as a virtual woman in order to resemble common voice-agents, such as Siri and Alexa. In the VE, the IVA was located to the right side of the user (see Figure 11.2). No body representation of the user was provided. The only representation they saw of themselves were the models of their controllers.

Figure 11.2: A collage of a participant in VR solving an anagram (left) while the full body agent (right) watches them.

During the experiment and in the presence of each virtual agent's representation, the user's task was to solve 12 anagrams. They had 40 seconds to solve each word. An information sign showed a timer in front of the user. The letters of the anagrams were displayed on round objects in front of the user. They had to be grabbed with a controller and moved to their correct place on a word board (see Figure 11.2). When the user dropped the letters on the board, they were registered at that position, but they could still be moved on the board or exchanged for other letters. Once the user was satisfied with their word, they could press the continue button. Alternatively, if the 40 second timer was up, the word automatically switched to the next one.

11.3.2 Measurements

During our study we employed several measurements, including some related to the performance of the user at solving anagrams as well as subjective measurements through questionnaires.

Social Presence

After each condition, we measured social presence with the Social Presence Survey (SPS) by Bailenson et al. [16], and the *Social presence - Actor within medium (parasocial interaction)* subscale of the Temple Presence Inventory (TPI) [195]. SPS is used to measure how much participants perceive a VA to be an actual, conscious and aware human. The *Social Presence as a Social Actor within a Medium* sub-scale of the TPI tests for parasocial interaction where users respond to social cues presented through a medium.

Eye Behavior

Using the Tobii Eye Tracking provided by the VR HMD, we recorded the fixation duration (FD) of the areas of interest (AoI) in the study. The AoIs included the agent's face/mouth and body, and the FD is the amount of time that the AoIs were looked at. We included eye tracking as prior research suggests positive correlations between FDs on agents and their willingness to continue working with them [391] and because different appearance fidelity of agents can influence the user's visual attention [351].

Cognitive Performance

In order to study the social facilitation and inhibition effects on cognitive performance, we employed anagrams in the local language in three different difficulty levels: (i) easy, (ii) medium and (iii) hard. We evaluated the easy anagrams in a web-based pre-study with 14 participants, where we presented them with 40 4-letter-words taken from a crossword puzzle website. The participants had 20 seconds to solve each anagram, by writing down the correct word in LimeSurvey. After the pre-study, two words had to be removed because they had more than one possible solution. Furthermore, one word was removed because less than 50 percent of participants solved it correctly. For this study, the words that 100% of the participants solved were also not used. This left us with 20 words for this study. For medium difficulty, we used ten of the easy anagrams from Hoffmann et al. with a frequency class of 10 and 11 [129] and ten of the easy anagrams from Voss et al. with a probability of correct answer between 50 and 90 percent [353]. We chose these as a medium difficulty instead of an easy one, because in our pretesting, our participants needed too long to solve these. As hard words, we also used the word pool of Voss et al. by picking out 20 of their hard words [353]. All words used in the anagram solving task can be found in Appendix C. During the study, the letters had to be physically dragged to their correct place within the word. This served as the physical training part of our study. For our data collection, we recorded the number of correct words, and the time needed to complete each word.

User Preferences

At the end of the study, we also asked some qualitative questions about participants' favorite and least favorite agent representation, and what they would choose if they had to do cognitive-physical training again in the presence of an IVA.

11.3.3 Study Procedure

After arriving at the lab, participants filled out and signed the informed consent form. Thereafter, they filled out the demographic questionnaire and put on the Pico HMD. The Pico eye tracking was calibrated for each participant by running the Tobii User Calibration sequence. After that, we instructed them to start the application when they were ready.

At first, participants were familiarized with the cognitive task (i.e., anagrams) by going through a tutorial. In this tutorial, participants had to solve two simple four-letter anagrams by dragging the letters to their correct position. In order to prevent additional social effects, the experimenter left the experimental room prior to each VR exposure.

Afterwards, the first condition started. The order of the conditions was counterbalanced with a Latin square. In each condition, the IVA (in its corresponding representation) explained the task by saying: "*Welcome to this round. Your task will [again] be to solve the following anagrams*". Then, the agent moved on to an idle animation and observed the participant. Participants had to solve four easy four-letter-words, then four medium six-letter-words, and finally four hard six-letter-words. For each word, they had 40 seconds. Once they pressed the continue button or when their time was up, the application switched to the next word. After all words were presented, the application prompted them to take off the HMD and answer the questionnaires regarding social presence and enjoyment on the computer. They were also allowed to take a short break. Then, they started the next condition. After finishing all conditions, they answered additional questions about their favorite and least favorite IVA representation and could leave additional comments. In total, the study took around one hour, with 20-25 minutes in VR (4-5 minutes per condition).

11.4 Results

This section presents the results for social presence (see Table 11.1), the eye tracking data evaluation (see Table 11.2), the performance measurements (see Table 11.3), and the user preferences (see Table 11.4) and explains the statistical evaluations that we conducted.

First, Shapiro-Wilk tests were conducted to test for normal distribution for all data. Despite the partially non-normal distribution of the data, we decided to report the analysis based on parametric tests in order to not switch between statistical tests. For this reason, repeated measures ANOVAs were selected as the statistical analysis method. Moreover, previous studies have shown that ANOVA is robust against violations of the normal distribution [27, 115, 291]. The significance level was set at 0.05. As an effect strength, the generalized eta squared (η_g^2) was reported. Thereby, a value of 0.01 represents a small effect, 0.06 a medium effect, and 0.14 a large effect [48, 181]. The p-values of the pairwise comparisons were corrected using the Bonferroni method.

11.4.1 Social Presence

As mentioned above, we employed two questionnaires to measure different aspects of perceived social presence. The participants' responses to these two questionnaires are depicted in Table 11.1 as well as in Figure 11.3 and Figure 11.4, with significant differences marked in the graphics.

Condition	**SPS**	**PSI**
Voice-only	-1.68 ± 0.17	1.9 ± 0.14
Mouth-only	-0.92 ± 0.14	2.03 ± 0.15
Head	-0.18 ± 0.17	2.73 ± 0.18
Upper Body	0.03 ± 0.12	3.28 ± 0.19
Full Body	0.19 ± 0.2	3.22 ± 0.22

Table 11.1: Mean and SE results from the Social Presence Survey and the Parasocial Interaction subscale (PSI) of TPI.

As it can be seen, the feeling of social presence increases with an increase in the level of agent visibility. The only exception in this pattern can be observed between the upper body and full body representations in terms of their perceived Parasocial Interaction. Here, the upper body received the highest mean value.

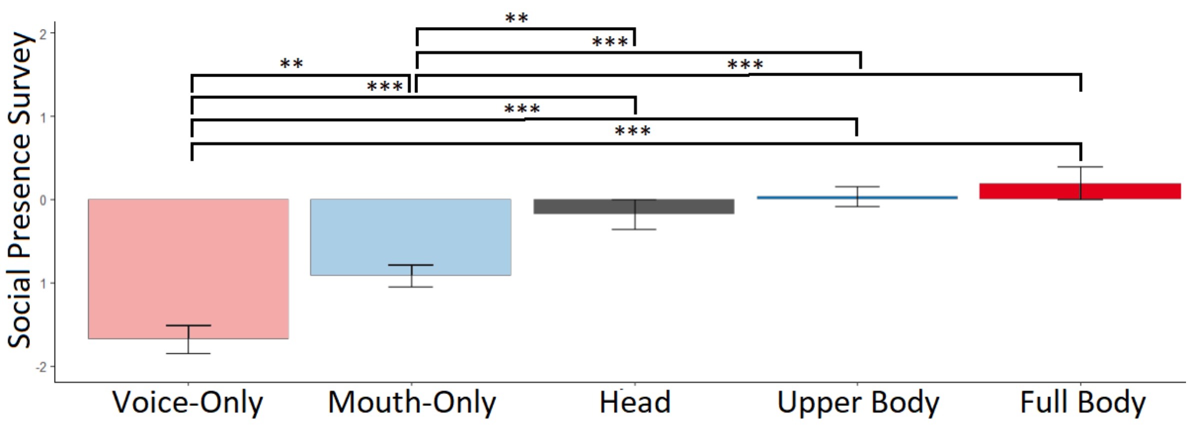


Figure 11.3: Social Presence Scale results for each condition. Bars show the mean and error bars the standard error. Significant differences are marked: *** $p <= 0.001$, ** $p <= 0.01$

We performed a repeated measures ANOVA on both social presence measures which indicated a significant main effect of the virtual agent's representations on SPS ($F(4, 116) = 35.9917, p < 0.001, \eta_g^2 = 0.3935$). Pairwise comparisons with Bonferroni corrections reveal that voice-only and mouth-only conditions had significantly lower SPS values than all other conditions (voice-head: $p < 0.001$; voice-upper body: $p < 0.001$; voice-full body: $p < 0.001$; voice-mouth: $p = 0.0055$; mouth-head: $p = 0.0013$; mouth-upper body: $p < 0.001$; mouth-full body: $p < 0.001$); with voice-only receiving the lowest rating overall. No significant differences could be observed between head and upper body ($p = 1$), head and full body ($p = 0.2176$), and the upper body and full body ($p = 1$) conditions.

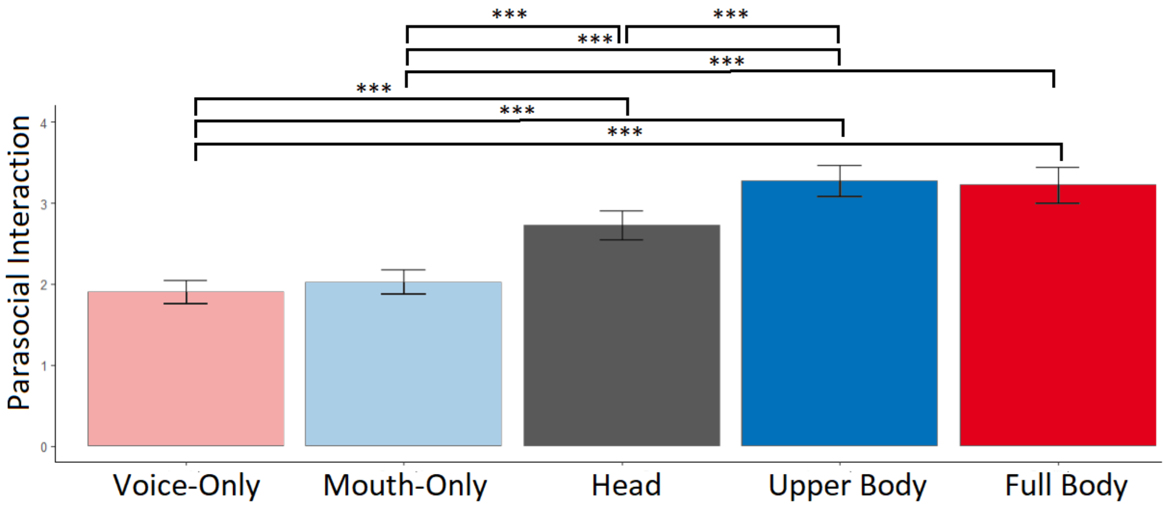


Figure 11.4: Social Presence results for each condition for parasocial interaction. Bars show the mean and error bars the standard error. Significant differences are marked: *** $p <= 0.001$, ** $p <= 0.01$

In addition, we observed a significant main effect of the virtual agent's representations on Parasocial Interaction ($F(4, 116) = 31.6124, p < 0.001, \eta_g^2 = 0.2662$). Here, the difference between voice-only and mouth-only was not significant ($p = 1$), but all other conditions received significantly higher scores compared to the voice-only and mouth-only condition (voice-head: $p = 0.001$; voice-upper body: $p < 0.001$; voice-full body: $p < 0.001$; mouth-head: $p = 0.001$; mouth-upper body: $p < 0.001$; mouth-full body: $p < 0.001$). Furthermore, the upper body condition received significantly higher values in Parasocial Interaction than the head ($p = 0.001$) condition. The difference between head and full body ($p = 0.0511$) as well as upper body and full body ($P = 1$) were not significant. Thus, the results of social presence measures (i.e., Social Presence Survey and Parasocial Interaction) support H_1.

11.4.2 Eye Behavior

It can be seen in Table 11.2 that the representations with a higher visibility were looked at for a longer period of time, especially during the explanation (around 46% of the time over all conditions). During the solving of the anagrams on the other side, the participants spent below 2% of their time looking at the agent, and focused mainly on the task instead. This effect is even stronger for difficult words, where participants spent an average of below 0.5% of their time looking at the agent, whereas they focused the agent for around 1% of their time in the easy conditions.

The results of a repeated measures ANOVA revealed a significant main effect of the agent representation on both, the time that the agent was looked at during the introduction of the task when the agent was talking ($F(3, 87) = 0.8262, p = 0.001, \eta_g^2 = 0.0893$) and while participants were solving the words and the agent was silent ($F(3, 87) = 9.4024, p < 0.001, \eta_g^2 = 0.0866$). Pairwise comparisons showed that during the introduction of the task, the full body representation was looked at for significantly longer time than the head ($p = .008$) and the mouth-only ($p = 0.01$) representations, but not longer than the upper body representation ($p = 1$). The upper body representation was looked at for significantly longer than the mouth-only representation ($p = 0.041$). No significant differences could be observed between the other conditions (head - mouth: $p = 1$, upper body - head: $p = 0.385$). During the solving of the words, the only significant difference could be observed between full body and head ($p = 0.034$) representations. Therefore, H_2 can be supported.

Finally, we observed a significant positive correlation between the time spent looking at the virtual agent and their rated social presence (Parasocial Interaction: $R = 0.44, p < 0.001$, SPS: $R = 0.51, p < 0.001$), supporting our hypothesis H_3.

Agent	**Σ FD (Total)**	**Σ FD (Explanation)**	**Σ FD (Solving)**
Mouth-Only	2.06 ± 3.77 sec	7.08 ± 4.71 sec	0.38 ± 0.68 sec
Head	2.33 ± 4.17 sec	7.98 ± 5.04 sec	0.44 ± 0.76 sec
Upper Body	3.27 ± 4.25 sec	9.78 ± 2.93 sec	1.11 ± 1.54 sec
Full Body	3.36 ± 4.46 sec	10.19 ± 3.52 sec	1.08 ± 1.27 sec

Table 11.2: The amount of time the different agent representations were looked at and the users' preferences. During the study, there were two phases: The explanation phase where the agent verbally explained the task (*Explanation*), and the test phase, where the participants solved the anagrams and the agent watched them (*Solving*).

11.4.3 Cognitive Performance

During the study, the number of correct words, and the time needed for solving the words was recorded. Table 11.3 shows the mean and standard errors for the number of solved words and the time needed to do so. The table is divided by word difficulty, since we wanted to test for differences among these. For correct words, the easy word comparison of one participant was removed because there was a tracking error in one of their conditions, preventing them from touching the letters; therefore the first three words were not solved, which would have resulted in a bias.

On the remaining data, a 3 (task difficulty: easy, medium, hard) x 5 (virtual agent's representations: voice-only, mouth-only, head, upper body, and full body) repeated measures ANOVA was performed. No significant main effect of the agent's visibility could be observed on the number of correct words $F(4, 116) = 0.6652, p = 0.617, \eta_g^2 = 0.0046$) or the time needed to solve all words ($F(4, 116) = 0.98, p = 0.421, \eta_g^2 = 0.0062$). Thus, H_4 and H_5 cannot be confirmed. However, a significant main effect of the task difficulty could be found ($F(2, 58) = 211.9678, p < 0.001, \eta_g^2 = 0.5352$), with easy words being solved significantly more often and faster than medium or hard words ($p < 0.001$), and medium words being solved significantly more often and faster than hard ones ($p < 0.001$). This confirms our design decision for the anagram tasks in three different levels of difficulty. We also tested for gender effects in the number of correctly solved words, but contrary to what was proposed by Bond and Titus [29], no significant differences were found in our work ($F(2, 27) = 1.3487, p = 0.277, \eta_g^2 = 0.0908$).

Condition	Difficulty	Solved Words	Time for Words
Voice-only	easy	3.62 ± 0.12	10.49 ± 0.49 sec
Mouth-only	easy	3.52 ± 0.14	10.89 ± 0.57 sec
Head	easy	3.72 ± 0.1	10.81 ± 0.54 sec
Upper Body	easy	3.79 ± 0.09	11.71 ± 0.55 sec
Full Body	easy	3.76 ± 0.09	11.8 ± 0.61 sec
Voice-only	medium	3.07 ± 0.19	13.88 ± 0.65 sec
Mouth-Only	medium	2.87 ± 0.18	13.51 ± 0.62 sec
Head	medium	2.87 ± 0.21	15.09 ± 0.69 sec
Upper Body	medium	3.03 ± 0.18	16.2 ± 0.79 sec
Full Body	medium	3.07 ± 0.16	15.12 ± 0.76 sec
Voice-only	hard	0.93 ± 0.17	27.33 ± 1.77 sec
Mouth-Only	hard	0.87 ± 0.17	22.24 ± 1.47 sec
Head	hard	1.17 ± 0.18	26.1 ± 1.2 sec
Upper Body	hard	0.7 ± 0.15	25.63 ± 2.23 sec
Full Body	hard	0.83 ± 0.17	20.91 ± 1.62 sec

Table 11.3: The number of words solved per condition (four words were the maximum) as well as the time needed to solve the words in seconds (time for words that were not solved correctly was not included). Mean and SE.

11.4.4 User Preferences

At the end of the session, we asked the participants about their favorite representation, their least favored one and which one they could imagine doing long-term cognitive-physical training with. The results can be seen in Table 11.4.

Regarding the favorite agent representation, it was quite balanced between voice-only, upper body and full body. Many people liked the voice-only condition, for example since it provided the least distraction from the real task (P_9, P_{11}, P_{17}, P_{25}, P_{28}, P_{30}), whereas the two embodied conditions provided a feeling of a real person being there

Agent	favorite	Least favorite	Training favorite
Voice-only	10 (+1)	4	8 (+8)
Mouth-Only	1 (+1)	15 (+2)	0 (+3)
Head	2	3 (+1)	1 (+3)
Upper Body	7(+1)	2 (+2)	4 (+4)
Full Body	8 (+1)	4	4 (+8)
Mixed	2	2	13

Table 11.4: This table contains the users' preferences regarding the favorite and least favorite representation, as well as the agent they would choose if they would do cognitive-physical training on a regular basis. Mixed results, where participants chose two or more options, are written in brackets and in the last row.

(P_5, P_{10}, P_{19}, P_{20}) and they were the most natural (P_{15}, P_{18}, P_{23}) and human-like (P_{24}, P_{27}). Some people criticized that the full body agent's legs sometimes collided with the ground or that the proportions did not appear correct (P_6, P_{21}, P_{24}).

The least favorite agent was the mouth-only representation, because it was "*creepy*" (P_3, P_9, P_{24}, P_{27}), unnatural (P_8, P_{20}, P_{25}, P_{28}) and uncanny (P_{10}). The head representation was also described as irritating and distracting (P_4, P_{18}), and weird (P_1, $P_1$4). Some participants also commented that the upper body and full body representations were distracting them from their task (P_{11}, P_{13}).

Regarding the question about their preference for future cognitive-physical training, most of the participants had mixed opinions. Many stated that it depended on the kind of task and duration (P_4, P_5, P_8, P_9, P_{12}, P_{15}, P_{19}, P_{28}). For sessions with a longer duration, some participants stated the benefit of embodied agent to not "*feel alone*" (P_4), and that some company would be nice (P_9). In the comments, the voice-only representation was pointed out most positively. More than half of the participants valued the voice-only condition because it was the least distracting (P_9, P_{13}, P_{17}, P_{25}, P_{30}) and one participant stated that they could visually focus on the task with this agent representation (P_8).

11.5 Discussion

In this section, we report on the implications of our research, especially in relation to our hypotheses. Furthermore, we reflect on limitations of our study design as well as future research directions.

11.5.1 Social Presence

Our results revealed that the perceived social presence increased with higher level of visibility of the IVA, which is in line with prior research on avatars [360, 380] and confirms our hypothesis H_1. An especially large difference could be observed between the two conditions without eyes, namely voice-only and mouth-only representations. They showed significantly lower social presence than the other conditions, indicating that the interaction with them did not feel as human-like as with the other agent representations. In addition, only the upper body and full body representations were in a positive range of the social presence survey, implying that only the visually rich representations seemed conscious and human-like to the participants. This is similar to a user study with Furhats, where the representation of the robotic heads led to low humanlikeness scores [5]. However, both social presence measures did not present very high ratings for any of the agent representations. This might be because of the limited interaction the users had with the agent representations. During each condition, the agent only explained the task at the start and did not speak to the user again afterwards; instead it only observed them while they solved the anagram task. A systematic review also suggested that during games or when users have to focus on the task, the representation of an agent or avatar loses relevance [360]. We suppose that this might also be the case here, where users had to focus too much on their task to pay attention to the agent.

11.5.2 Eye Tracking Behavior

During the explanation of the task, participants looked longer at the IVA when it was represented with a higher level of visibility, which is in line with our hypothesis H_2. We found a significant positive correlation between the amount of time the participants looked at the IVA and the social presence scores (H_3). The difference in viewing times might be attributed to better capturing of the users' attention. For once, the size of the agent, with larger representations providing more visual details, might contribute to this. In an AR study using a human-sized agent and a miniature version of it, the larger agent was also looked at for longer amounts of time, but the smaller one was looked at more often [358]. It has also been shown that faces are almost always in focus [221], and a more in-depth examination of the eye tracking data would provide additional information on where exactly the users looked – into the face of the agents or onto their body, if it was visible. Furthermore, the gestures that were used by the two most visible representations might have caught the attention of the users [117].

In future work, it could also be tested how cognitive load differs between the representations through psychophysical measurements like blinks, saccades or pupil dilation [164, 305]. This would provide additional information on task performance and information processing.

11.5.3 Task Performance

We could not find any significant differences for task performance in the presence of the five representations of the IVA. Thus, in contrast to our priori hypotheses (H_4 and H_5) and previous research [63, 224], we could not observe any social facilitation and inhibition effects on the anagram solving task.

This might be due to several reasons. First, several other papers could also not find differences in task performance depending on the visualization of their agents [203, 358], which might imply that the visualization does not have a large effect on this measurement. Furthermore, the difficulty of the task we used might not have been suitable to show these effects. For the anagrams we chose, most of the easy ones could be solved by all participants, and most of the hard anagrams were too hard to be solved at all, especially in the 40 seconds time that the participants had per word. This might have led to floor/ceiling effects [386]. Compared to our pre-study, where we only tested the easy words, participants needed slightly longer to solve the words (pre-study: 7.59 ± 1.97 seconds, agent study: 11.25 ± 5.83 seconds), which we attributed to the physical task of moving each letter to its right place. For the hard words, even the users who solved them needed more than half of the given time. In future work, the amount of time allowed for participants to either solve the anagrams or to skip the word could be increased.

11.5.4 Limitations and Future Work

Furthermore, our study design shows several limitations.

First, we only tested one specific visualization of an agent. In future studies, several additional conditions could be included, such as comparing different genders of the IVA, comparing a real human to an IVA, a different species of agent, or no agent/human at all. In our design, we only showed one female agent to the users and did not match or manipulate the gender. Prior research has shown that in the medical or therapeutic context, female agents are largely favored over male ones [9, 225], especially by male participants [326]. The gender of the agent also affected students, who were less likely to interact with a male virtual human than a female one [89]. This is why we decided to use a female agent only, but in future studies, the gender should be varied to test for these gender effects.

Finally, the chosen cognitive-physical task might not be able to reveal any social facilitation or inhibition effects. This might be due to the limited interaction with the agent, who only took on an observatory role. However, a verbal or conversation-based task in which the IVA takes on a more visible role could cause a difference. One possible choice for a task might be an island survival task, which relies on collaboration and communication. This has been tested before in an asymmetric extended reality collaboration between two users, but even here, the authors found no difference in social presence between their avatar representations [111]. They assume this might be because the task mainly relied on audio. Nevertheless, the increased interaction with the conversational partner might effect social presence. In the case of our study design, if the correct word would have been verbalized to the IVA, as done by Mostajeran et al. [224], the IVA would have been more in focus, which might have a positive influence on social presence. Furthermore, including interactive conversations before the start of the task, for example using LLMs like ChatGPT [243] or the Wizard of Oz methodology might further increase realism and strengthen the connection between the agent and the user.

11.6 Conclusion

In this work, we presented the results of a user study to examine the effects of different levels of agent visibility, in particular, on perceived social presence and cognitive task performance in VR. The IVA was represented as voice-only, mouth-only, head, upper body, and full body of a virtual woman. Thirty participants performed anagram tasks in three levels of difficulty (easy, medium, and hard) in the presence of each IVA's representation. In order to understand the user's gaze behavior towards different representations and its relation to the self-reported social presence measure, their eyes were tracked during exposure to the representation of the IVAs. After each exposure in VR, they evaluated the IVA through questionnaires addressing their social presence. They were also given the chance to select their preferred representations.

The results provide important implications for the design of future IVAs for innovative technologies and setups, such as AR/VR systems or new working or learning environments. Our study suggested that when users perform cognitive training, IVA's level of visibility may not have a statistically significant effect on their performance. Eye-tracking data further emphasized that the visual attention of the users during the task was not on the agent, but on the task itself. This could be a reason for its low influence on task performance.

While the IVA was not paid attention to during the solving of the task, the participants spent around half of the time looking at the agent during the introduction of the task. Here, the higher the level of visibility, the longer participants looked at them. This is also in line with our findings for social presence, where higher levels of visibility had a significant positive effect on the perceived social presence.

Preferences on the agent representation were mixed, and comments further implied that the representation should depend on the use case and training duration, with voice-only interfaces providing the least distraction from other visual tasks, and upper body or full body representations inducing the highest feelings of not being alone. This is an important result for the visual design of IVAs in a majority of use cases, such as training, exergames or educational scenarios. Furthermore, the integration of more intelligent or interactive virtual agents could help to create a higher feeling of social presence.

Chapter 12

Investigating the Use of AI Assistants in a VR Exergame for Older Adults

Figure 12.1: An older user playing the *Canoe Game* with the AI agent helping them.

The last two chapters have shown that most IVAs are still in an early state of research, with only few of them employing a real AI system for controlling conversations. With this study, we were trying to bridge this gap by integrating an IVA communicating through GPT-4o into the *Canoe Game* (See Chapter 7). Our literature review has shown that the visual design of an IVA depends on the use case (See Chapter 10), and our study on agent visibility underlined this, with participants' opinion on their preferred representation ranging from the voice-only agent to a full-body agent (See Chapter 11). In the last study, none of the agents resulted in a high feeling of social presence, which is why we investigated alternatives over the humanoid agents [187]. Here, a parrot was the preferred option by younger and older adults for the *Canoe Game*, which is why we integrated it to visualize the IVA. In this user study, we were interested in identifying barriers and potentials for AI-guided IVAs in VR exergames for older adults. With this, we aimed to strengthen the independence of older adults in using these technologies without depending on caregivers or researchers. The work from this chapter is based on the following publication:

Kruse, L., Rings, S., and Steinicke, F. (2025). My Focus was on the Game: Investigating the Use of AI Assistants in a Virtual Reality Exergame for Older Adults. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA). DOI: `https://doi.org/10.1145/3706599.3719808`

12.1 Introduction

In order for exergames to show the desired effects, it is crucial for them to be designed well. Especially for first-time VR users, controls of the application, rules, or safety notices should be explained and communicated in an easily understandable way. In this regard, reading a lengthy instruction sheet and the insecurities of (new) users with the technology can be a usage barrier, which we aim to break down by employing IVAs. In most prior work, conversations with the agents were scripted, and only a small subset used some sort of AI to enable realistic, natural conversations between the user and the agent [165]. This leaves room to research the impact of natural, verbal conversations between IVAs and older adults.

AI has made huge advancements in the last years [260, 343], with LLMs like the generative pre-trained transformer 4 (GPT-4) [2] transforming its view of the public. With LLMs, natural conversations with IVAs are made possible. These agents can potentially take on a variety of tasks and roles, from assistants [143] to idea generators or conversation partners [271]. In our case, they take on the role of a tutorial agent that explains the game rules to the players and comments on their performance.

When talking about AI and IVAs, most of the time, younger adults are asked about their opinion, excluding the view of older adults in using the technology. Nevertheless, this user group has their own opinions on this topic, with their own barriers, but also their own potentials for using it to extend their daily life and it is important to ensure a more inclusive and comprehensive understanding of how these technologies can be used across all age groups.

With this work, we aimed to get a first glimpse into using IVAs in VR exergames played by older adults. Our aim was to see how an IVA that is communicating verbally using GPT-4o and text-to-speech (TTS) is perceived by the older adults in terms of user experience and humanlikeness, if it raises their motivation to train and which barriers the users are facing. Furthermore, we were interested to analyze the conversations the older adults had with the IVA, identifying potentials, but also errors of the IVA that lead to misunderstandings. With this knowledge we aim to provide guidelines on designing IVAs using LLMs in exergames for older adults.

Our research question for this study was:

- $\mathbf{RQ}_6$: How do older adults perceive and use an intelligent virtual agent in a VR exergame?

12.2 Methods

This section details the application adjustments we made for this version, as well as the measurements and the study procedure.

12.2.1 Application

As an application, we used the *Canoe Game*, which has been evaluated with older adults for a variety of different research questions [149, 169, 187]. In this VR exergame, users sit in a virtual canoe that runs down a river and have to touch soap bubbles with their hands, according to different rules. We adjusted the original version to feature only three levels: A level where all bubbles should be collected, one where red bubbles need to be collected with the right hand, and yellow bubbles with the left hand, and finally a level where bubbles that have a number above them and bubbles with a letter above them have to be collected alternately. The game starts in augmented reality (AR) passthrough mode allowing users to seamlessly transition into the virtual environment. This helps them adjust their vision and become familiar with the VR controllers, ensuring a smooth and easy on-boarding experience. When users feel comfortable, a button press starts a transition to the VR experience, allowing them to take control over the game flow, which we found to be perceived positively in our prior work [170]. They then have to calibrate their sitting position by pressing and holding both trigger buttons for three seconds, followed by a calibration of their range of motion, which is done by pushing four balloons as far away as possible. After this is done, they can press a button on their right side to start the canoe, and subsequently collect the soap bubbles. At the end of the game, a highscore panel shows them how many correct bubbles they collected, rewarding them with positive feedback and featuring competitive interactions between players [25].

To integrate the IVA, we added an animated 3D parrot (See Figure 12.4), as suggested in our prior work, to fit to the tropical island scenario [187]. In that user study, a full-body humanoid agent was compared to a text-interface and to an animated parrot. Results with younger and older adults showed that the parrot was the preferred representation for this exergame, which is why it was integrated in this iteration. The parrot uses the OpenAI model *GPT-4o* with the OpenAI Text-to-Speech voice “Fable” for speech generation in German. Users are able to communicate with the agent naturally through their own voice. The spoken words are then transcribed using the Google Speech-to-Text API and sent to GPT-4o as a user prompt. Voice input is only recognized while the IVA is not talking itself. We implemented this to avoid interference caused by concurrent voice input. A sign in the virtual world indicates to the participants when they are allowed to talk (a mouth, with a green border), or when they can only listen to the parrot (an ear, with a red border). This was explained and shown to them beforehand (See Figure 12.2).

The interaction with the AI agent takes place in the form of a single coherent chat, which is started off with a system prompt that details the expected interaction and medium of

Figure 12.2: A user is shown the icons indicating to the participants that they can currently talk (left), and that they can currently only listen (right).

communication, ensuring the agent's answers can easily be transformed to audio using TTS. The prompt also contains contextual information about the game in the form of the scene, task, expected events and instructions the agent should communicate to the participant as a step-by-step guide. It is the agents task to guide the users through the game and explain the steps detailed above. In addition, the agent can answer the user's questions about the game, its controls, e.g., which buttons to press, or its benefits. The agent also gives the players a compliment if they perform well in the game, or a reminder of the game rule if they are still making mistakes. A flow diagram of the game, as well as the explanations of the agent are shown in Figure 12.3. After certain changes of the game state, or actions performed by the participant, these are added to the end of the chat history in the shape of a written description, including the contents of the next expected response. For example, if the calibration is successfully finished, this is communicated to the agent by sending a prompt stating that the user is done, and the next step can begin. The agent cannot directly access the state of the game, e.g., verify whether what the user communicated is true – at this point, it is only text-based. A translation of the exact prompts given to the agent in each situation can be found in Appendix D.

12.2.2 Measurements

To assess how well the game and the IVA are received, we used several measurements. First, we were interested in the user experience, which we assessed with the User Expe-

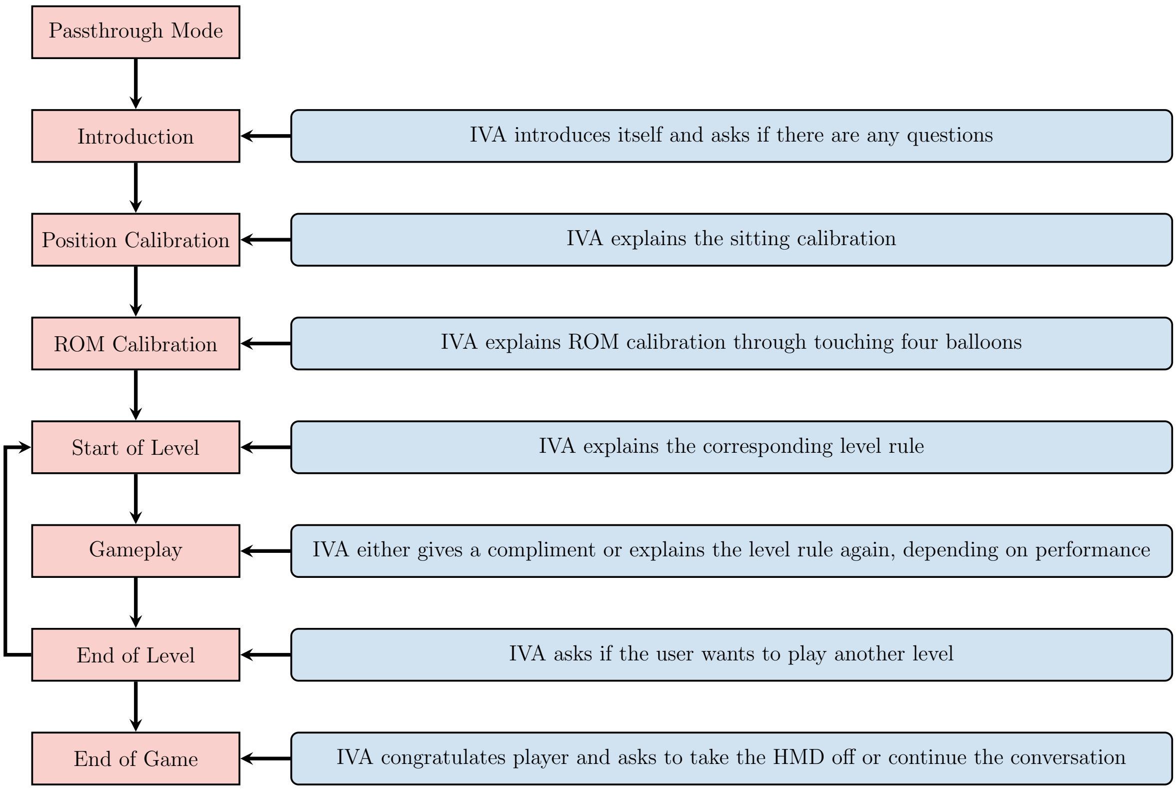


Figure 12.3: Gameflow (red nodes) as well as input of the IVA provided in different situations (blue nodes).

rience Questionnaire (UEQ) [184].

Next, we were interested in how motivated the participants were to play the game. For this, we used a short version of the Intrinsic Motivation Inventory [280, 362] (IMI). Here, we were especially interested in the Interest/Enjoyment subscale, which is considered as the self-reported measurement of motivation.

Third, the participants answered the Godspeed Questionnaire (GQS) [19] to get a better picture of their perception of the virtual agent.

Finally, we performed a semi-structured interview with the participants, where we asked them about their experience with the IVA, their wishes, fears and if they could imagine using an AI agent in real life. The interview questions can be found in Appendix D. Additionally, we logged the prompts that the participants sent to the IVA, as well as the agent's responses.

12.2.3 Study Procedure

We conducted the study in cooperation with a local senior living home, the Hospital zum Heiligen Geist Hamburg, on their premises. Upon arrival, participants signed a consent form and received a small briefing about the goal of the study. We did not tell them anything about the VR game itself, but rather explained that they could ask the virtual agent what the game was about, and how to play it. We also gave them a presentation of how to talk to the agent, showing them the visualizations of an ear for listening and a mouth for talking (See Figure 12.2).

For the study, we used a Meta Quest 3 HMD with two controllers. It was connected to a 4G+ mobile network. The participants then put on the HMD and started in the passthrough mode, where they had time to adjust the HMD and controllers. The parrot introduced itself and asked the participant if they had any questions. Then, it continued to guide them through the calibration and the game flow. After each level, the participants could decide to either continue playing, or end the game. At the end of the game when the scoreboard was shown, the IVA congratulated the player, informed them how many bubbles they collected and told them they could either take the HMD off, or continue the conversation with the parrot.

Afterwards, they filled out the three questionnaires mentioned in subsection 12.2.2 and a semi-structured interview was conducted with the participants. At the end, they were debriefed and thanked for their time.

12.3 Results

Eleven older adults participated in the study. All of them were able to give legal consent and did not have a diagnosed cognitive impairment. They were between 77 and 91 years old, with an average of 84.54 ± 3.73 years. Nine identified as female and two as male. Two of them had used an AI or voice assistant on a computer before. Seven participants had used a VR headset in the scope of user studies before, and the other four had experience with an immersive projector-based gaming system called the ExerCube [317]. The results for all questionnaire data can be found in Table 12.1.

Figure 12.4: A visualization of the virtual parrot.

Scale	Mean and SD
UEQ Attractiveness	1.83 ± 0.96
UEQ Perspicuity	1.16 ± 1.06
UEQ Efficiency	0.93 ± 1.03
UEQ Dependability	0.93 ± 1.15
UEQ Stimulation	1.59 ± 0.69
UEQ Novelty	1.36 ± 1.11
IMI Interest/Enjoyment	3.64 ± 0.44
IMI Perceived Competence	2.48 ± 1.07
IMI Perceived Choice	3.06 ± 0.80
GQS Anthropomorphism	3.85 ± 0.71
GQS Animacy	4.16 ± 0.67
GQS Likeability	4.33 ± 0.74
GQS Perceived Intelligence	4.22 ± 0.88
GQS Perceived Choice	3.61 ± 0.84

Table 12.1: Mean and SD for all questionnaire data we collected during the study.

12.3.1 User Experience Questionnaire

The UEQ was evaluated using the evaluation tool provided on the authors' website [296] and compared to the benchmark data (See Figure 12.5). It shows good scores for attractiveness, stimulation and novelty, but also below average results for the three subscales perspicuity, efficiency and dependability.

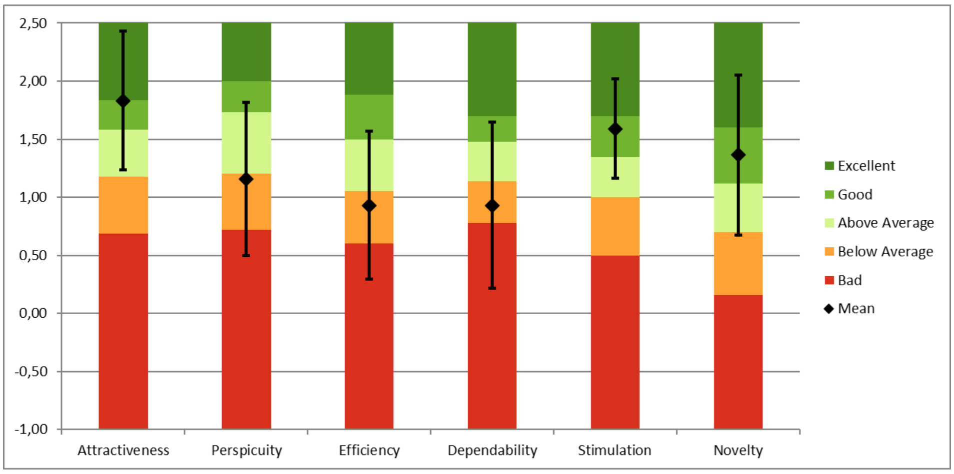


Figure 12.5: The results of the UEQ compared to the Benchmark data.

12.3.2 Intrinsic Motivation Inventory

The interest/enjoyment subscale of the IMI showed a high self-report for motivation (3.64 ± 0.44). Nevertheless, it is lower than in a study performed with the same game without the IVA previously [169], where this score ranged between 4.22 and 4.7, depending on the condition. Furthermore, our participants rated their perceived competence quite low (2.48 ± 1.07) - even lower than the older adults rated it in the *Faster* condition in [169] (4.0), which was meant to be overwhelming.

12.3.3 Godspeed Questionnaire

The GQS showed relatively high values in all subscales. Compared to the parrot in our prior study [187], our parrot received slightly higher values for animacy and perceived intelligence, but slightly lower ones for likeability.

12.3.4 Interview Evaluation

To evaluate the semi-structured interviews, we first transcribed them and then performed an inductive thematic analysis on the data to find common themes [32, 33, 35]. In the following, we report the main findings, which relate to the themes '*Communication*', '*Focus*', and '*Possibilities of AI*'.

Communication Most of the participants rated the communication as very positive. Participants found that the IVA was patient (P_{11}), understandable (P_4, P_5, P_9), and appreciated that instead of trying out which buttons work, they had someone to ask their questions ("*If someone is there, one can ask*", P_{11}). One participant said that the game was too complicated for new players, but she explicitly said that she did not blame the parrot and his explanations for this (P_3). Nevertheless, some participants also encountered technical issues in the communication. For example, at some point, the agent did not respond anymore or a reply from GPT took too long to get through (P_2). Also, for one participant, the agent kept repeating sentences, which was perceived negatively (P_7).

Focus Many participants reported that their focus was not on the IVA, but on the game, even though the briefing informed them about the goal of the study. P_1 reported that they did not talk much to the agent, because they were just focusing on the game. P_{10} also said that during the game, their attention was on the task and if the agent had talked, this would have bothered them. This was also confirmed by P_6, who paid little attention to the possibilities of talking to the IVA, but rather followed the instructions of what the parrot explained. They were focused on something new and were not interested to find out facts like, for example, why the game was developed.

Possibilities of AI Before the study started, the possibilities of asking the agent questions were explained to the participants, e.g., that they could ask questions about the game, but most older adults did not use this feature. In the interview, after being reminded of this possibility, P_1 mentioned that they should have asked the agent something. After giving P_3 a specific example, they said that if they had known this before, they could have used the AI in a different way, but they simply did not know. P_{10} also said that if they had more insight into the topic of artificial intelligence, it would be easier for them to deal with it, but they were not interested to invest too much time in learning about it. P_8 also said that they did not know which other possibilities there were, and instead of getting a list of all possibilities, they would let themselves be surprised.

Finally, when asked if they would have been willing to play the game without the researcher present, but only with the parrot, all eleven participants confirmed that they would.

12.3.5 Interaction Evaluation

During the study, all prompts were recorded, along with observations from the leading researcher. On average, each user sent 5 ± 2.17 prompts to the agent with a total of 55 prompts over the eleven participants. Interestingly, the number of prompts sent by each person is positively correlated with their perceived competence, as tested with a Spearman's correlation ($t = 2.4988, df = 9, p - value = 0.03393, r = 0.64$).

The evaluation of the conversation content through Thematic Analysis resulted in two interesting topics: '*Misunderstandings and Hallucinations*', and '*Timing and Initiative*', which will be presented in the following paragraphs.

Misunderstandings and Hallucinations Out of the total of 55 prompts, 15 were statements like "*okay*", "*ready*", "*done*", or "*thank you*". From observations, prompts like "*done*" sometimes led to misunderstandings, resulting in the IVA continuing to explain the next game rule, even if tasks were not successfully finished yet. For example, this happened if the calibration button was only touched for a short time, or if not all of the balloons were touched yet. The participants believed they were done, communicated this, and this false belief in turn was transferred to the IVA (P_6: "*Was this enough?*", IVA: "*Yes, this should be enough. Lets continue with the second step*").

Another twelve prompts were questions about the game, for example which button to press (P_6), how to reach the bubbles (P_{11}), what comes next (P_9) or the inquiry to repeat what the agent just said (P_2). Further, they asked for confirmation on if they should perform an action now (P_2, P_7, P_{11}). Here, another limitation became visible - hallucinations. When users asked for a confirmation on their action by using deictic communication such as pointing, the IVA tended to confirm their question, even if the participants meant the wrong action. For example, P_6 asked if "*the black round start button*" was the correct one to start the game, whereupon the IVA confirmed it:

"*Yes, exactly. Press the round black start button to start the canoe*". In this case, the participant was looking at the trigger button on the controller and the correct button was spatially on their right side and colored in red. A similar error happened for P_2, who commented that she had "*a finger back [on the trigger button] but nothing [was] moving*". The IVA then explained that the start button was on the right side of the screen and that they should point the controller onto the button and then press the trigger button. Again, this was not correct, since the button worked by simply pushing it away with the controller, similar to a real-world button.

Timing and Initiative Most of the questions occurred during the calibration of the game. Another eleven prompts followed immediately after the initial greeting, where the users were asked if they had any questions about the game. Here, only two participants asked a question regarding what the game was about (P_1), and what the IVA was planning to do with them (P_8). All others stated that they currently did not have any questions. During the gameplay, the participants did not ask questions anymore, but rather voiced some comments about the gameplay, for example that the bubbles were "*too high*" (P_5), "*I can't do this*" (P_{11}), or reminding themselves of the game rule "*No, that's left, red right, left [yellow]*" (P_5). After the game ended, the parrot told the users that if they wanted, they could continue talking to him. Similarly to the start, only two participants used this opportunity. One user asked the parrot how many bubbles they missed (P_9), while another one told him that the topic of AI was very unfamiliar to them (P_{10}), which was followed by a short explanation of AI by the IVA. The other users ended the game or took the HMD off. Another interesting observation was that often, participants made understanding comments while the IVA was talking. In order to prevent interference, these comments or interruptions were not recorded at this point, because the IVA only listened for input while it was not talking.

12.4 Discussion

In this section, the results from the questionnaires, interviews, and interactions with the IVA will be discussed and put into context. Furthermore, guidelines for future research are presented.

12.4.1 Questionnaires

While the design of our study, as well as the relatively small number of participants, challenged statistical evaluations of our questionnaire results, we chose to present the findings and comparisons with prior studies, along with a discussion on observable trends.

The UEQ shows below-average results for perspicuity, efficiency, and dependability. We suspect that this might partly be due to the delay from asking a question to receiving the verbal answer. Due to network restrictions of the senior living facility affecting access to the GPT API server, we were limited to use a mobile network instead of a faster WIFI connection. Furthermore, the silence timeout we used to detect that the participants were done talking added an additional short delay. These delays could lead to a more unnatural and less efficient perception of the conversation.

The dependability score of the application was also negatively impacted by the technical limitations, such as long delays, or sometimes no reply at all due to network errors. While these issues will be resolved with advanced versions of GPT and faster network connections, current programmers should account for them. One idea could be to use animations or filler words that subtly indicate that an answer is currently being generated [182].

While the GQS showed relatively high scores for likeability, they are lower than the ones experienced in a prior study without the use of AI. We note that in both studies, the number of participants was relatively low and the different ratings can be due to interpersonal reasons. Nevertheless, there might also be some reasons for the lower likeability score. For example, in normal conversations, humans often make confirming sounds or use words and behaviors showing their interest in the conversation, e.g., back-channeling [350]. While our participants showed this behavior in the conversation with the IVA, the IVA was neither sensing, interpreting nor showing the same behavior when the older adults were talking. This could have further made the conversation with the IVA feel less natural and reduced the older adults' sympathy for the IVA. Nevertheless, the parrot did show slightly higher values for perceived intelligence compared to the pre-scripted assistant from prior work, reflecting its use of a responsive AI. Additionally, prior versions of the game did not feature any negative feedback, whereas the agent this time reminded the player of the rules if mistakes were made.

As one participant noted, the IVA's dialogue during gameplay could be distracting and may have increased the pressure to perform well. This might also be reflected in the IMI score, with a low perceived competence, and, compared to prior versions, also a lower interest/enjoyment. Interestingly, the perceived competence was positively correlated with the number of prompts sent by the participant. This might indicate that asking the IVA for help increased the feeling of competence of the players. Nevertheless, these results should be taken with caution and further comparison studies are needed to confirm these observations.

12.4.2 Qualitative Discussion

While the qualitative evaluation showed great potential for the use of IVAs, e.g., having the opportunity to ask questions, the main focus of the participants was on the game itself, and not on the IVA – similarly to the results we found in the study about agent visibility (See Chapter 11). In our case, the IVA guided the conversation and explained the next step automatically when an action, for example the calibration, was completed. In future research, it would be interesting to see how the conversation would be shaped if this behavior was different, e.g., if the participants had to actively inquire about each step to receive more information. Nevertheless, we suspect that this might further decrease the perceived efficiency of the game. After all, the IVA only served as a helper and, for the older adults, was not the main part of the experience. As researchers, we are interested in the results of a user study, however, the older adults were most likely more interested in experiencing the game content and benefiting from the immersive training. One older adult also voiced a concern about the research itself, stating: "*How much sense does such an invention or [...] scientific endeavor make at all? [...] Is it worth investing all the time and money into such a project, or could [...] things be invented that [...] are more helpful?*".

In this regard, we can confirm that all players were able to navigate through the game and play it until the end, without input from the researchers. This is also confirmed by the interview data that stated that all players would be happy to play only with the IVA, without a researcher present. Nevertheless, it remains an open research question whether in reality, this is feasible and also safe. Especially for movement-intensive games, we believe that there should be a human observer present to intervene in case of a (medical) emergency, at least until AI capabilities are advanced enough to ensure the safety of the older adults at all times.

12.4.3 Technical Discussion

The largest technical limitation we encountered during the study was the latency in the responses from the agent, which resulted from the slow internet connection we experienced while at the facility. This influenced all the connections for interacting with the agent, namely the STT, AI and TTS. While the connection speed is the main cause, in each of these steps, improvements are possible and necessary to result in smooth, natural communication.

Further, we encountered several hallucinations of the agent. While LLMs are prone to hallucinate unintended outcomes, there are different ways to prevent this. First, we only observed extrinsic hallucinations, which are hallucinations that cannot be verified from the source input [145]. In our case, this was observed with the color of the button (original: red, hallucination: black), which we did not include in our initial prompt to the agent, and the interaction with the button (original: pressing, hallucination: raycasting). In future implementations, pre-testing might be able to identify these issues and prompts can be adjusted. Another idea would be to mitigate the hallucinations through modeling and interference methods [145], such as reinforced learning or fine-tuning. The IVA also tended to agree with the user when they asked a question, without having the corresponding information. Rychert et al. integrated a prompt into their agent that stated to "*never assure the user that they can do something with a specific object*" [283], which might also help to decrease the certainty with which the agent suggests actions in our implementation. Further, if information is not provided to the agent, it could visually and audibly show signs of uncertainty, as it was also done in prior research [294]. Further, the latest advancements of GPT feature vision abilities. Including this could help to make the intentions of the users more clear, e.g., by detecting which objects their gaze or attention is on, and could also account for the imprecise speech of the users, e.g., "*Should I press this button?*". It could also help to include state awareness, e.g., by always sending the state of the game and of interactable objects to the IVA in each prompt [283]. This could help to reduce errors from users falsely claiming to be done with a task.

12.4.4 Guidelines

With the knowledge gained from our exploratory user study, we propose the following guidelines for designers and programmers planning to integrate IVAs into their applications that are made for older adults.

1. Make the Agent More Approachable Often, the older adults were not aware of all the possibilities the IVA provided, or they had hesitations to ask questions. One idea would be to make the agent more approachable, e.g., with a friendly introduction, emphasizing that it is in the game to help the players and that they can always ask questions. Further, it could give examples on what to ask. This could, on one hand, lower the barrier to ask questions in the first place, and on the other hand, it would help players to discover more of the ways they can use the IVA.

2. Use Specific Prompts Prompts and information given to the AI should be as specific as possible, with as much information as possible. While we aimed to provide all information necessary for the gameplay, which was even more than prior versions of the game communicated, we could still observe hallucinations where the IVA filled in missing knowledge with suggestions of the user, e.g., touching the wrong button. Through careful playtesting with the end users in a user-centered design approach, common misconceptions can be identified and prompts can be altered.

3. Implement State Awareness In our implementation, the whole conversation was guided through user prompts. Here, the LLM did not make a difference between prompts that were sent by our system, and prompts that originated from the verbal conversation with the older adults. This led to misunderstandings if the participants falsely communicated the successful execution of an action. We, therefore, propose to implement state awareness in the IVA, e.g., through the ability to access the state of the game flow [283], or by integrating new advancements such as vision abilities.

4. Consider Whether to Really Use AI While using AI has many advantages, there might be situations where predefined dialogues surpass the non-deterministic communication of IVAs. Posing the same question several times can lead to different outcomes, and some of those might not be the desired ones, e.g., wrong or misleading information. Designers should carefully consider whether there are situations where exact words are needed to understand the instructions or context – which could be implemented without the integration of AI. Furthermore, environmental and ethical considerations of using LLMs [21] should be kept in mind when using them in large-scale implementations.

12.4.5 Limitations and Future Work

Our study design shows several limitations that should be addressed in future work. First, we only evaluated one condition, namely the exergame with the integration of an IVA using GPT-4o. This was done to get a first glimpse into the usage of IVAs by older adults, and to inform us about possible barriers and difficulties. In future studies, within-subject studies should be conducted to compare different implementations with each other, e.g., an LLM-based IVA with a pre-scripted dialogue system. Furthermore, our sample size was relatively small. While larger sample sizes could help to confirm the patterns we found in this paper, larger sample sizes could also diversify the perspective and interactions of our users on this topic. For example, it would be interesting to also include older adults who have prior experience with LLMs, or those who have no experience with (VR) exergames at all. Finally, technical limitations such as a long delay should be minimized, or ways to bridge the waiting time should be implemented. Moving on with this topic, we plan to evaluate how prior training in using LLMs might benefit the interaction between older adults and the IVA. Furthermore, we plan to investigate the interaction between IVAs and older adults in the context of different usage scenarios, e.g., physical or cognitive therapy.

12.5 Conclusion

In our exploratory user study, we evaluated the feasibility of an IVA in a VR exergame that assisted the players and guided them through the gameflow. In interviews and through questionnaires, we found that the communication with the IVA was perceived as positive, the participants' questions could be answered understandably and all of them stated that they would also play the game alone with the IVA, without a researcher being present.

Nevertheless, technical limitations such as a high delay and hallucinations deteriorated the experience. Especially wrong information being communicated by both, the IVA and the users, led to misunderstandings. State awareness is needed to reduce these occurrences. Further, we found that the older adults in our study rather focused on the game, and the conversation with the IVA only played a subordinate role. We propose that the possibilities of AI should be explained and demonstrated to the older adults beforehand, informing them about possible use cases.

While it is easier to use pre-scripted tutorials, IVAs present a possibility for further tailoring the agent's behavior and communication to the needs of the individual users. In the future, this might enable older adults to access technology such as VR exergames more independently, empowering them to ask questions and fully enjoy the benefits these programs offer.

Chapter 13

Discussing Intelligent Virtual Agents for Immersive Training

In this part, our research focused on intelligent virtual agents in the healthcare context and in immersive training applications. Each paper emphasizes the importance of a good user experience for the successful interaction with the IVA, highlighting how different user preferences can influence engagement and user experience.

In the first paper, a systematic literature review was conducted that analyzed 59 papers on embodied virtual agents in healthcare ($\mathbf{RQ}_4$). Most agents were represented as humanoid agents, with different visual preferences depending on the use case and user group. They were often employed for treatment of various conditions, e.g., for conversational therapy or physiotherapy, but also served as motivators or educators. Positive attributes of IVAs were identified, such as their availability and accessibility, but technical limitations and the inability to perform well in crisis situations diminished the view of participants on IVAs.

The second paper presented a study evaluating one aspect of agent representation – agent visibility ($\mathbf{RQ}_5$). In this study, participants performed an anagram solving task in VR in the presence of five different agent representations, ranging from a voice-only agent to a full-body agent. The study found that a higher visibility positively influenced the perceived social presence, but it did not have an effect on the cognitive-physical task performance. Further, preferences of the users on the agents depended on the task and duration. Importantly, the agent representation should support the users in their task, and not distract them. Finally, eye tracking data showed that the agent was only looked at during the explanation of the task, while the users focused predominantly on the anagram solving task afterwards.

In the last study, the knowledge gathered on IVAs was used to integrate an interactive assistant into the *Canoe Game*. It was able to explain the game rules and answer questions of the players about the game or its controls using GPT-4o. In a user study, the interaction with the agent was evaluated and older adults were asked to share their insights, as well as barriers and preferences they had about the agent ($\mathbf{RQ}_6$). The study found that the agent successfully guided the players through the exergame, and all older adults stated that they would also play the game only with the IVA and without a researcher in the room. Nevertheless, technical limitations such as hallucinations of the LLM or a long delay deteriorated the experience. Further, participants were more interested to play the game instead of interacting with the agent, again emphasizing the subordinate role that assistants play in exergames.

The three works demonstrate that research in this direction is still relatively new, with advancements in recent years dynamically changing the technical possibilities, and also shaping the view of the public on IVAs and AI. Our studies show that integrating IVAs into exergames, while being promising and technically feasible, needs the involvement of the end users to determine a suitable representation and interaction modalities. Further, the agent's role as assistants and not as the main content of the application was pointed out in both of our studies, which is also underlined by literature.

13.1 Design Considerations

This part of this thesis has demonstrated that, while IVAs show enormous potential for assisting users in immersive training applications, several challenges were observed, and there is not one universal solution for the design of IVAs. Therefore, we drew three design considerations for future studies employing IVAs: (i) To choose a suitable visualization depending on the user group, context and task, (ii) to note that the attention of the users is not necessarily on the agent, and (iii) to consider whether a fully realistic and natural agent is needed for VR exergames.

13.1.1 Suitable Visualizations Depend on the Task, Context and Users

Both the systematic literature review from Chapter 10 and the user study on agent visibility presented in Chapter 11 showed that the preference of users regarding the representation of the agent differ. In prior literature, this varied from users wishing for text-only agent interfaces [344], to preferring cartoon models of humanoid agents [322], and to wanting realistic representations of doctors in formal clothing [349, 364]. Here, it depended on the use case, the situation and the user group. In our study on agent visibility, this was also pointed out in the preferences of the users, with a similar number of them preferring the voice-only, half body and the full body agent. When asked which one they would prefer for long-term training, their opinions shifted towards the voice-only agent, but almost half of the participants could not decide on a single visualization for this question. This is similar to another research paper, where the preferred communication modality with an agent depended on the usage context [163]. For our last study, we decided on using a parrot representation, as we found that this was the preferred option for this specific exergame [187]. Nevertheless, when employing an IVA in a different context, a parrot might not be the most suitable representation. This indicates that more research in this area is needed, and conclusions drawn from our – context-dependent – exergames might not be transferable to every VR exergame scenario.

13.1.2 The Attention of the Users is Not Always on the Agent

In both of the user studies we conducted, we realized that the attention of the users was not predominantly on the agents, but on the cognitive-physical task they had to fulfill. While for us, the effect of the agent was what guided the research, this was not necessarily the most interesting part of the experience for our study participants. In the agent visibility study, we recorded the participants' gaze, and it was shown that they only looked at the agent during the explanation of the task – during the task execution, their visual attention was mostly on the interactive elements. In the last study, our participants also noted that they just wanted to play the game, instead of finding out more information about it, e.g., its cognitive-physical benefits. In both studies, the agent only served as a method to understand the game rules, without the participants' desire to form a deeper connection with it.

13.1.3 Reconsidering the Need for Hyperrealistic Agents

All three works have shown that the most realistic agent is not necessarily the best option. In the literature review, sometimes cartoonish agents were preferred over realistic ones [322]. The visibility study also indicated that a voice-only interface might be superior to a realistic full-body agent in some cases, which might be unrealistic in real-world co-located scenarios. Our prior work also confirmed this for the agent used in the last study, where a speaking parrot was preferred over a humanoid agent [187], which is also not easily transferable to real-world cases. This indicates that, in some scenarios, visual realism might not be the most important factor to a successful human-agent-interaction.

Additionally, in the final study, the realistic communication skills of the agent were barely used - only few older adults used the opportunity to ask questions to the IVA. Further, even though the LLM showed potential in communicating game rules differently, depending on the users' skills, conversations between participants were often very similar. In this case, it is worth considering whether integrating an LLM is worth it. Positive aspects include the possibility to engage in realistic conversations, and the availability of immediate verbal help for user questions. But on the downside, the non-deterministic nature of LLMs makes it hard to control what the IVA will say in all situations; leading to misunderstandings and hallucinations. For future research, it should be considered to only use an LLM when necessary, e.g., for intermediate questions, and employ fixed, pre-recorded rules in other situations. This does not only have benefits in terms of reduced costs and environmental considerations [21], but also helps developers to remain fully in control of their application.

Part IV

Conclusion

Chapter 14

Summary

In this thesis, different research in the domain of XR exergames and IVAs was presented. The first study showed that engaging older adults with dementia in playing a VR exergame over the course of several weeks improved their well-being, and provided a meaningful, engaging activity, especially during the COVID-19 pandemic ($\mathbf{RQ}_1$). In the second study, we concluded that VR exergames for older adults are comparable to traditional, video-based training, but the acceptance and engagement depends on the chosen tasks, difficulty and user preferences ($\mathbf{RQ}_2$). The third study highlighted differences between the user groups of younger and older adults in their preferred difficulty level, but also indicated that an increased difficulty led to a higher enjoyment among both user groups, while at the same time deteriorating the movement accuracy ($\mathbf{RQ}_3$).

Intending to improve the independence of older adults in engaging with VR exergames on their own, the design and integration of IVAs was explored in the following three works. Our systematic literature review indicated that embodied IVAs in healthcare are designed in various ways, they are employed for a variety of different tasks, and we point out different design recommendations ($\mathbf{RQ}_4$). In a user study, we showed that an increased visibility of IVAs enhances perceived social presence, but it did not have an effect of task performance. Here, we highlight the need for IVAs to be designed to assist users without distracting them ($\mathbf{RQ}_5$). The final study showed that older adults appreciated the integration of an IVA, but technical limitations impacted their experience, and participants showed a preference for engaging with the game content over interacting with the IVA ($\mathbf{RQ}_6$).

The general research question of this dissertation was how to design immersive cognitive-physical training applications, especially for older adults. Overall, our research has shown that, in order to design effective cognitive-physical training in VR, a multitude of considerations has to be taken into account, and that there is rarely one determining factor. In this thesis, these considerations ranged from the employed technology (Chapter 6) to varying difficulty levels (Chapter 7), and design considerations when integrating IVAs in exergames (Chapter 11 and Chapter 12).

The exergames we developed were a source of joy and lively discussions. Our study participants often volunteered to be included in future studies, and emphasized the potential of the exergames. Additionally, caregivers noted the engagement of the older adults who participated in our study sessions, especially during the COVID-19 pandemic. This emphasizes that even in exceptional circumstances, VR exergames can help to entertain and keep older adults fit.

Chapter 15

Design Recommendations

To give researchers working on the design and evaluation of VR exergames for older adults directions for future developments, the following section will present design recommendations derived from this thesis.

In Chapter 8, we already proposed some design considerations leading to the integration of IVAs in the following part:

- The dependability of older adults on caregivers.
- The importance of feedback and encouragement in exergames.
- The differing needs of different user groups.

Further, in Chapter 13, considerations for the integration of the IVAs were proposed:

- To choose a suitable IVA visualization depending on the user group, context and task.
- To note that the attention of the users is not necessarily on the agent.
- To consider whether a hyperrealistic IVA is needed for VR exergames.

Reflecting on these design considerations and the work conducted as a whole in this thesis, the following implications can be drawn.

15.1 Include the Actual End Users in the Design Process

Our user studies have shown that it is essential to integrate the end users in the design process. This helps to obtain their opinion on preferences, to observe possible barriers and to ensure that the developed applications have a good user experience. Throughout this thesis, applications and elements were sometimes evaluated with younger users due to unavailability of larger numbers of older adults during the COVID-19 pandemic. While the younger adults helped to identify flaws in the implementation and design, they are not a suitable substitute for fine-tuning the game due to differences in aspects such as technical expertise and cognitive-physical fitness. Further, preferences for the game context, wishes for music, or physical accommodations, highly depend on the users and might substantially differ between diverse demographic groups.

Additionally, it is worth reevaluating whether the interfaces we design are wanted and needed by the older adults, and whether each element contributes to a positive user experience. As an example, the IVA integrated into the exergame in Chapter 12 was not evaluated as positively as we had expected. The older adults' engagement with the IVA was limited, and UEQ scores for perspicuity, efficiency, and dependability were below average. Furthermore, one older adult raised doubts about whether the integration of AI makes sense in this context. Therefore, early focus groups and brainstorming sessions with target users are important to quickly identify these issues. Here, showing early prototypes or demonstrators is encouraged to foster the engagement of the users, and move beyond theoretical discussions about novel, unknown concepts.

15.2 Designing Engaging, Enjoyable Tasks for VR Exergames

The exergames often showed a consistently high enjoyment, confirming that the tasks and settings were suitable for most participants. For example, enjoyment remained over 4 (out of 6) during the long-term study of *Memory Journalist* [167], and IMI scores remained over 4 (out of 5) for the *Maestro Game* [168], and in the *Canoe Game* [169]. Only in the last study of the *Canoe Game* it was lowered to 3.64, which is still comparatively high (See Chapter 12). Therefore, it can be concluded that all of our exergames showed to be motivating, and that the condition, e.g., the difficulty in [169], only played a minor role. Hence, we encourage researchers to design the tasks and setting together with the end users to create engaging experiences and identify settings that are less suitable early on [149].

15.3 Choosing the Correct Difficulty of VR Exergames

When designing VR exergames, special care should be placed on delivering a suitable difficulty, or providing the possibility to adjust the difficulty to the user. The second game, *Maestro Game*, was originally developed for people with dementia, but could not be evaluated with them due to the COVID-19 pandemic and the risks associated with their participation. Therefore, older adults without dementia evaluated the game, but it proved to be too easy for most of them, leading to a lower intention to use this game for long-term training. While older adults in prior focus groups enjoyed the exergame [270], it was not deemed suitable for more athletic players, and even not perceived as sports.

In the first evaluation of the *Canoe Game*, the effects of different difficulties could be shown, with a higher movement intensity leading to a higher enjoyment [169]. Nevertheless, this also reduced the movement accuracy, which is important for obtaining healthy training effects and preventing injuries. Therefore, a suitable balance between these two factors should be maintained.

15.4 Considering the Safety of VR Experiences

Particular attention should be given to the design of safe experiences. This ranges from data safety, to physical safety, but also to mental well-being, e.g., avoiding experiences that trigger unwanted negative effects [359].

First, data safety, especially in combination with third party services, should be considered. In VR, movement and other nonverbal data can be collected, and these could potentially be used to determine the users' future behavior [15], therefore identifying users. Further, developers of VR applications could potentially get insights into the physical environment surrounding the users via the use of the passthrough mode, or the correspondingly drawn boundary. Regarding IVAs, their AI abilities are still largely based on proprietary LLMs, such as OpenAI's GPT. This company is based in the US and while they are committed to protecting their users' privacy and developing a safe AI system [285], different laws in different countries, such as the General Data Protection Regulation (GDPR) in EU countries, can apply. Further, special care should be taken when handling sensitive data, e.g., in the healthcare context [278].

Physical safety encompasses the necessity of creating applications that prevent physical injuries, such as falls, or warn the users if their behavior is likely to cause such injuries. For example, in Chapter 7, a higher difficulty led to faster, uncontrolled movements, which could possibly result in injuries. Moreover, when immersed in VR, the users' awareness of their real surroundings and their proprioception might be impaired, especially if there is a conflict between visual and proprioceptive senses. Additionally, symptoms of cybersickness can occur. As a way to prevent this, intelligent sensors could be used to track the movement of the players and information should be provided if the movements are likely to deteriorate the well-being of the users. Additionally, environmental dangers such as cables should be eliminated before starting the VR experience [150].

Finally, the mental well-being of the participants should be kept in mind. This means that experiences should not trigger negative effects, e.g., remind users of traumatic events that happened in their past [359]. While this should be prevented as much as possible through careful design of the applications, suitable strategies should be developed to react to these situations if they occur, e.g., peacefully ending the VR experience [167]. Further, the games should not cognitively or perceptually overwhelm participants. While this could be solved by developing games that can be customized to the users, again, stop markers should be developed to smoothly end the experiences.

Chapter 16

Future Work

This thesis shows that there are many remaining research directions in the domain of exergames for older adults and IVAs, and for the combination of these two topics. In order to develop motivating exergames that foster long-term adherence, these should be addressed in future research.

16.1 A Suitable Choice of Technology

While the exergames investigated in this thesis all used immersive VR through HMDs, other display technologies might be easier to handle for older adults. For example, CAVE environments like the ExerCube [317] have been integrated into senior living facilities and show a considerable success in engaging the older adults to exercise [120]. Here, weekly training sessions are offered, and high score lists document and display the progress of the older adults and foster competitiveness and adherence. On the other hand, these systems are not suitable for individual, home-based training due to their size and price, where HMDs show some advantages.

16.2 Exploring Exergames Outside the Research Context

For the future of XR exergames, it is important that they are available and affordable for all users. For example, even if a standalone HMD like the Meta Quest costs below 500€, this is still a significant barrier for people with limited income, e.g., people on a pension – and it does not automatically include customized games or therapies. One therapy application that is currently on the market is Cureo, a VR application for treating sensorimotor or cognitive impairments, such as strokes [55]. It is available with a session-based license, with around 5€ per session. While the therapy is paid for by one insurance company in Germany [56], users largely have to pay for it themselves. One possibility for financial support in Germany are so-called DiGas ("Digitale Gesundheitsanwendungen") [62]. These are digital applications in the health domain that show a positive effect for patients and are paid for by insurance companies. While this concept is still relatively new, there is already one DiGa available that uses VR to treat agoraphobia, highlighting the potential for XR technology to be included in this DiGa register [142].

While medical certifications of VR exergames are out of the scope of this thesis, the idea to work towards this in the future is intriguing. One factor for getting a product medically certified is that it is low-risk, and that it improves the users' health or their ability to live with a diagnosis [140]. In the scope of this thesis, we were not able to conduct multiple long-term studies. An exception is the first study, which was, however, only conducted on a small scale. Therefore, the effectiveness of the exergames could not be evaluated on long-term physical or cognitive effects. Prior research has also found that the long-term effects and the adherence to XR exergames has been poorly studied [148], and this shows the necessity of conducting such studies.

16.3 Confirming the Effects Found During the Pandemic

Especially during times of social isolation, an imminent need for training opportunities for older adults was observed, and this need could not be fulfilled by traditional, face-to-face methods. Nevertheless, the pandemic situation also led to limitations in the studies; namely small test groups, dropouts due to sickness and lock-downs and unavailability of special test groups such as people with dementia in later studies. These limitations negatively influenced the statistical power of our user studies. Further, since other training programs or social activities were canceled during the pandemic, our user studies might have been evaluated more positively than they would during normal circumstances - simply because they were some of the limited events available during that time. Therefore, it is important to confirm these findings under normal circumstances.

16.4 The Future of IVAs

For future research, it remains an interesting research question how to further improve the independence of older adults in using XR technology. Our approach of integrating IVAs to improve the independence of older adults indicated that the older adults would also exercise alone without the IVA. On the other side, technical limitations decreased our confidence in this technology at this point (See Chapter 12). Furthermore, while we can largely control the content within our application, external factors such as the HMD hardware and operating system, especially standalone HMDs, remain out of our control, since the integration of an IVA in these domains is – currently – not intended by the manufacturers. This hinders IVAs to also help the users with the preparation of the hardware or with starting the application. Nevertheless, advancements of AI assistants in everyday life look promising, e.g., the integration of AI assistants in the Meta Smart Glasses [215]. Following this trend, we can expect to see the integration of IVAs in further areas of our life, and their usage is likely to also become more common, and more natural over time.

Advancements in LLMs are quickly progressing, which will enable more natural conversations, quicker responses and improved reasoning skills of LLMs. These again need to be evaluated with end users. Already now, OpenAI integrated vision abilities into GPT-4o, meaning that the model can analyze images and answer questions about them. This, in combination with scene and state awareness, can help to eliminate communication errors such as wrong beliefs about the state of VR experiences.

Moreover, IVAs should be integrated and investigated in different use cases, which foster a more active communication of the human users. For example, conversational games where the agent and the user solve riddles together, or reminiscence activities, where the agent asks interactive questions about the user's past could be investigated. Finally, the interactive abilities of IVAs could also be used for conversational therapy or psychotherapy in order to improve the psychological well-being of users, i.e., older adults [176].

16.5 Ethical Considerations

In all of these cases, ethical considerations need to continuously be taken into account. LLM technology is currently not in a state that is without errors, e.g., biases, hallucinations, and lack of knowledge about real-time events [21, 136, 234]. Future work should evaluate the safety of the developed systems, and also investigate whether ethical guidelines are followed. For example, the (AI) systems should be thoroughly tested in long-term evaluations while also evaluating critical situations with experts. When working with vulnerable user groups, such as people with dementia, special care should be taken, and their interests should take precedence over those of the researchers.

Additionally, researchers should ensure that the data collected within the VR systems is securely stored. Aggregated or irreversibly anonymized data are not considered "personal data", and, therefore, lack legal protection. Nevertheless, there is the potential to infer highly sensitive information from this data [355], and it can be used by machine learning algorithms to identify users, even if users try to deliberately change their behavior [179]. There are some positive effects in inferring information about, for example, the procession of diseases like dementia, or the training effects obtained through the exergames. Still, this medical information needs to be protected and user should retain control about their data.

16.6 Summary

There are still many research gaps that should be filled in the future. An emphasis should be put on the people using the systems, and their lived experiences. Further, technical advancements will facilitate the development of faster, more capable systems, which should be evaluated for feasibility, safety and for ethical implications with the end users.

Chapter 17
Conclusion

This thesis investigated the effect and the user experience of VR exergames. In three studies with older adults, different aspects of these games were evaluated, i.e., (i) their effect on the well-being of older adults, (ii) their comparability against traditional video-based exercises, and (iii) the ability to transfer information about suitable difficulties of an exergame between user groups. In these three studies, we found that VR exergames can have a positive effect on the users, even in exceptional situations such as the COVID-19 pandemic. Furthermore, we found that VR exergames are comparable to traditional video-based training, but that they need to match the skills of the users. Finally, the last study showed that users prefer faster-paced games. It highlights the potential of testing exergames with different user groups for identifying structural mistakes in the games, but also showed the differences of user groups in terms of fitness and preferences. In all of these studies, caregivers or researchers had to assist the older adults in playing the exergames; however, this reliance on assistance is not sustainable in the long term due to the high workload of caregivers and for financial reasons.

Aiming to improve the independence of older adults in experiencing the VR exergames on their own, we further investigated the concept of intelligent virtual agents. Here, starting with a systematic literature review, we found that the preference of users on IVAs largely depends on the use case, the context, and the user group. In a user study, we investigated the impact of different representations of an IVA in terms of body visibility during cognitive-physical training in VR. Here, we found that more visible representations led to a higher feeling of social presence, but no effect on task performance could be observed. Further, social presence ratings remained rather low for all agents, and the attention of the users was on the cognitive-physical task – not on the agent. In a final user study, we combined the knowledge gathered in previous work into an exergame. Here, we integrated an IVA guided by GPT-4o, and investigated the interaction and the opinions of older adults with the IVA. We found that while the older adults would also play alone with the IVA, and without a researcher or caregiver present, technical limitations and hallucinations of the LLM deteriorated the experience. Additionally, similarly to the previous study, the agent was not in focus for many of the older adults, whose goal it was to play the exergame.

Future work should aim to improve the interaction between the users and the IVAs, or investigate other ways to improve the independence of the older adults. Finally, studies to investigate the long-term impact of VR exergames on the psychological, cognitive and physical well-being of older adults should be conducted. This would help to verify the enormous potential that exergames show for improving the well-being of older adults.

Bibliography

[1] Jennifer Abbasi. "Social isolation—the other COVID-19 threat in nursing homes". In: *Jama* (2020).

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[3] *Ageing*. 2024. URL: `https://www.who.int/health-topics/ageing` (visited on 11/08/2024).

[4] National Institute on Aging at the National Institutes of Health. *What is Mild Cognitive Impairment?* 2024. URL: `https://www.alzheimers.gov/alzheimers-dementias/mild-cognitive-impairment` (visited on 04/08/2022).

[5] Isabella Ågren and Annika Silvervarg. "Exploring humanlikeness and the uncanny valley with furhat". In: *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. 2022, pages 1–3.

[6] Rieks op den Akker, Randy Klaassen, Kim Bul, Pamela M. Kato, Gert-Jan van der Burg, and Pierpaulo di Bitonto. "Let Them Play: Experiences in the Wild with a Gamification and Coaching System for Young Diabetes Patients". In: *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*. PervasiveHealth '17. Barcelona, Spain: Association for Computing Machinery, 2017, pages 409–418. ISBN: 9781450363631. DOI: `10.1145/3154862.3154931`. URL: `https://doi.org/10.1145/3154862.3154931`.

[7] Inaki Merino Albaina, Thomas Visser, Charles A.P.G. van der Mast, and Martijn H. Vastenburg. "Flowie: A persuasive virtual coach to motivate elderly individuals to walk". In: *2009 3rd International Conference on Pervasive Computing Technologies for Healthcare*. 2009, pages 1–7. DOI: `10.4108/icst.pervasivehealth2009.5949`.

[8] George S Alexopoulos, Robert C Abrams, Robert C Young, and Charles A Shamoian. "Cornell scale for depression in dementia". In: *Biological psychiatry* 23.3 (1988), pages 271–284.

[9] Bayan Alsharbi and Deborah Richards. "Using Virtual Reality Technology to Improve Reality for Young People with Chronic Health Conditions". In: *Proceedings of the 9th International Conference on Computer and Automation Engineering*. Iccae '17. Sydney, Australia: Association for Computing Machinery, 2017, pages 11–15. ISBN: 9781450348096. DOI: `10.1145/3057039.3057080`. URL: `https://doi.org/10.1145/3057039.3057080`.

[10] Shilpa Amarya, Kalyani Singh, and Manisha Sabharwal. "Ageing process and physiological changes". In: *Gerontology*. IntechOpen, 2018.

[11] Gustavo Andrade, Geber Ramalho, Hugo Santana, and Vincent Corruble. “Extending reinforcement learning to provide dynamic game balancing”. In: *Proceedings of the Workshop on Reasoning, Representation, and Learning in Computer Games, 19th International Joint Conference on Artificial Intelligence (IJCAI).* 2005, pages 7–12.

[12] Dennis Ang and Alex Mitchell. “Representation and frequency of player choice in player-oriented dynamic difficulty adjustment systems”. In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play.* 2019, pages 589–600.

[13] Claudia Armbrüster, Marc Wolter, Torsten Kuhlen, Will Spijkers, and Bruno Fimm. “Depth perception in virtual reality: distance estimations in peri-and extrapersonal space”. In: *Cyberpsychology & Behavior* 11.1 (2008), pages 9–15.

[14] Rikke Aune Asbjørnsen, Jøran Hjelmesæth, Mirjam Lien Smedsrød, Jobke Wentzel, Marianne Ollivier, Matthew M. Clark, Julia E. W. C. van Gemert-Pijnen, and Lise Solberg Nes. “Combining Persuasive System Design Principles and Behavior Change Techniques in Digital Interventions Supporting Long-term Weight Loss Maintenance: Design and Development of eCHANGE.” eng. In: *JMIR human factors* 9 (2 2022), e37372.

[15] Jeremy Bailenson. “Protecting nonverbal data tracked in virtual reality”. In: *JAMA pediatrics* 172.10 (2018), pages 905–906.

[16] Jeremy Bailenson and Jim Blascovich. “Interpersonal Distance in Immersive Virtual Environments”. In: *Personality & social psychology bulletin* 29 (2003), pages 819–33. DOI: 10.1177/0146167203029007002.

[17] Karlene Ball, Daniel B Berch, Karin F Helmers, Jared B Jobe, Mary D Leveck, Michael Marsiske, John N Morris, George W Rebok, David M Smith, Sharon L Tennstedt, et al. “Effects of cognitive training interventions with older adults: a randomized controlled trial”. In: *Jama* 288.18 (2002), pages 2271–2281.

[18] Bundesfachstelle Barrierefreiheit. *Das Barrierefreiheitsstärkungsgesetz (BFSG).* 2024. URL: https://www.bundesfachstelle-barrierefreiheit.de/DE/Fachwissen/Produkte-und-Dienstleistungen/Barrierefreiheitsstaerkungsgesetz/barrierefreiheitsstaerkungsgesetz_node.html (visited on 11/08/2024).

[19] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. “Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots”. In: *International journal of social robotics* 1 (2009), pages 71–81.

[20] Russell M Bauer. “Physiologic measures of emotion”. In: *Journal of clinical neurophysiology* 15.5 (1998), pages 388–396.

[21] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the dangers of stochastic parrots: Can language models be too big?” In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.* 2021, pages 610–623.

[22] Robbert-Jan Beun, Eveliene de Vos, and Cilia Witteman. "Embodied conversational agents: effects on memory performance and anthropomorphisation". In: *International workshop on intelligent virtual agents*. Springer. 2003, pages 315–319.

[23] Timothy Bickmore, Daniel Mauer, Francisco Crespo, and Thomas Brown. "Negotiating Task Interruptions with Virtual Agents for Health Behavior Change". In: *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 3*. Aamas '08. Estoril, Portugal: International Foundation for Autonomous Agents and Multiagent Systems, 2008, pages 1241–1244. ISBN: 9780981738123.

[24] Timothy Bickmore, Amy Rubin, and Steven Simon. "Substance Use Screening Using Virtual Agents: Towards Automated Screening, Brief Intervention, and Referral to Treatment (SBIRT)". In: *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. New York, NY, USA: Association for Computing Machinery, 2020. ISBN: 9781450375863. URL: https://doi.org/10.1145/3383652.3423869.

[25] Laura-Bianca Bilius, Ovidiu Andrei Schipor, and Radu-Daniel Vatavu. "The Age-Reward Perspective: A Systematic Review of Reward Mechanisms in Serious Games for Older People". In: *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*. 2024, pages 168–181.

[26] Pauline Bimberg, Tim Weissker, and Alexander Kulik. "On the usage of the simulator sickness questionnaire for virtual reality research". In: *2020 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*. Ieee. 2020, pages 464–467.

[27] Mariéa José Blanca Mena, Rafael Alarcón Postigo, Jaume Arnau Gras, Roser Bono Cabré, Rebecca Bendayan, et al. "Non-normal data: Is ANOVA still a valid option?" In: *Psicothema* (2017).

[28] Jessica Blaynee, David Kreps, Maria Kutar, and Marie Griffiths. "Collaborative HCI and UX: longitudinal diary studies as a means of uncovering barriers to digital adoption". In: *Proceedings of the 30th International BCS Human Computer Interaction Conference 30*. 2016, pages 1–6. DOI: 10.14236/ewic/HCI2016.72.

[29] Charles F Bond and Linda J Titus. "Social facilitation: a meta-analysis of 241 studies." In: *Psychological bulletin* 94.2 (1983), page 265.

[30] Bruno Bouchard, Frédérick Imbeault, Abdenour Bouzouane, and Bob-Antoine J Menelas. "Developing serious games specifically adapted to people suffering from Alzheimer". In: *International Conference on Serious Games Development and Applications*. Springer. 2012, pages 243–254. DOI: 10.1007/978-3-642-33687-4_21.

[31] Maya Boustani, Stephanie Lunn, Ubbo Visser, and Christine Lisetti. "Development, Feasibility, Acceptability, and Utility of an Expressive Speech-Enabled Digital Health Agent to Deliver Online, Brief Motivational Interviewing for Alcohol Misuse: Descriptive Study." eng. In: *Journal of medical Internet research* 23 (9 2021), e25837.

[32] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology". In: *Qualitative research in psychology* 3.2 (2006), pages 77–101. DOI: 10.1191/1478088706qp063oa. URL: https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa.

[33] Virginia Braun and Victoria Clarke. "Reflecting on reflexive thematic analysis". In: *Qualitative Research in Sport, Exercise and Health* 11.4 (2019), pages 589–597. DOI: 10.1080/2159676x.2019.1628806. URL: https://doi.org/10.1080/2159676X.2019.1628806.

[34] Virginia Braun and Victoria Clarke. "Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches". In: *Counselling and Psychotherapy Research* 21.1 (2021), pages 37–47.

[35] Virginia Braun and Victoria Clarke. "One size fits all? What counts as quality practice in (reflexive) thematic analysis?" In: *Qualitative Research in Psychology* 18.3 (2021), pages 328–352. DOI: 10.1080/14780887.2020.1769238. URL: https://doi.org/10.1080/14780887.2020.1769238.

[36] Noel Brick, Tadhg MacIntyre, and Mark Campbell. "Attentional focus in endurance activity: new paradigms and future directions". In: *International Review of Sport and Exercise Psychology* 7.1 (2014), pages 106–134.

[37] Poppy Brown, Felicity Waite, Sinead Lambe, Julia Jones, Lucy Jenner, Rowan Diamond, and Daniel Freeman. "Automated Virtual Reality Cognitive Therapy (gameChange) in Inpatient Psychiatric Wards: Qualitative Study of Staff and Patient Views Using an Implementation Framework." eng. In: *JMIR formative research* 6 (4 2022), e34225.

[38] Renée FAG de Bruijn, Elisabeth MC Schrijvers, Karen A de Groot, Jacqueline CM Witteman, Albert Hofman, Oscar H Franco, Peter J Koudstaal, and Mohammad Arfan Ikram. "The association between physical activity and dementia in an elderly population: the Rotterdam Study". In: *European journal of epidemiology* 28.3 (2013), pages 277–283. DOI: 10.1007/s10654-013-9773-3.

[39] Diana Bruno and Sofia Schurmann Vignaga. "Addenbrooke's cognitive examination III in the diagnosis of dementia: a critical review". In: *Neuropsychiatric disease and treatment* 15 (2019), page 441.

[40] Bundesministerium für Bildung und Forschung. *EXGAVINE: Mit Virtueller Realität interaktiv gegen Demenz und Parkinson*. 2018. URL: https://www.interaktive-technologien.de/projekte/exgavine (visited on 11/10/2024).

[41] Marius Butz, Daniel Hepperle, and Matthias Wölfel. "Influence of visual appearance of agents on presence, attractiveness, and agency in virtual reality". In: *International Conference on ArtsIT, Interactivity and Game Creation*. Springer. 2021, pages 44–60.

[42] AJ Campbell and M Clare Robertson. "Otago exercise programme to prevent falls in older adults". In: *Wellington: ACC Thinksafe* 3 (2003).

[43] Pablo Campo-Prieto, José Cancela Carral, and Gustavo Rodríguez-Fuentes. "Immersive virtual reality as physical therapy in older adults: present or future (systematic review)". In: *Virtual Reality* (2021). DOI: 10.1007/s10055-020-00495-x.

[44] Dale Cantwell, Daire O Broin, Ross Palmer, and Greg Doyle. "Motivating elderly people to exercise using a social collaborative exergame with adaptive difficulty". In: *Proceedings of the 6th European Conference on Games Based Learning, Cork, Ireland*. 2012, pages 4–5.

[45] Wojtek J Chodzko-Zajko, David N Proctor, Maria A Fiatarone Singh, Christopher T Minson, Claudio R Nigg, George J Salem, and James S Skinner. "Exercise and physical activity for older adults". In: *Medicine & science in sports & exercise* 41.7 (2009), pages 1510–1530.

[46] Zubin Choudhary, Gerd Bruder, and Greg Welch. "Visual Hearing Aids: Artificial Visual Speech Stimuli for Audiovisual Speech Perception in Noise". In: *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*. 2023, pages 1–10.

[47] Sue VG Cobb, Sarah Nichols, Amanda Ramsey, and John R Wilson. "Virtual reality-induced symptoms and effects (VRISE)". In: *Presence: Teleoperators & Virtual Environments* 8.2 (1999), pages 169–186.

[48] Jacob Cohen. "A power primer." In: *Psychological bulletin* 112.1 (1992), page 155.

[49] Stanley Colcombe and Arthur F Kramer. "Fitness effects on the cognitive function of older adults: a meta-analytic study". In: *Psychological science* 14.2 (2003), pages 125–130. DOI: 10.1111/1467-9280.t01-1-01430.

[50] World Wide Web Consortium. *Introduction to Understanding WCAG 2.0*. 2023. URL: https://www.w3.org/TR/UNDERSTANDING-WCAG20/intro.html (visited on 11/08/2024).

[51] Aurora Constantin, Catherine Lai, Elaine Farrow, Beatrice Alex, Ruth Pel-Littel, Henk Herman Nap, and Johan Jeuring. ""Why is the Doctor a Man": Reactions of Older Adults to a Virtual Training Doctor". In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. Chi Ea '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pages 1–6. ISBN: 9781450359719. DOI: 10.1145/3290607.3312811. URL: https://doi.org/10.1145/3290607.3312811.

[52] Eric J. Cooks, Kyle A. Duke, Jordan M. Neil, Melissa J. Vilaro, Danyell Wilson-Howard, Francois Modave, Thomas J. George, Folakemi T. Odedina, Benjamin C. Lok, Peter Carek, Eric B. Laber, Marie Davidian, and Janice L. Krieger. "Telehealth and racial disparities in colorectal cancer screening: A pilot study of how virtual clinician characteristics influence screening intentions." eng. In: *Journal of clinical and translational science* 6 (1 2022), e48.

[53] Anna Cox, Paul Cairns, Pari Shah, and Michael Carroll. "Not doing but thinking: the role of challenge in the gaming experience". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2012, pages 79–88.

[54] Carolina Cruz-Neira, Daniel J Sandin, and Thomas A DeFanti. "Surround-screen projection-based virtual reality: the design and implementation of the CAVE". In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023, pages 51–58.

[55] *CUREO® ist die Zukunft der Therapie für sensomotorische, kognitive und perzeptive Therapie*. 2024. URL: `https://www.cureosity.com/de` (visited on 11/08/2024).

[56] *CUREosity und Partner starten Pilotprojekt für Virtual-Reality-Therapie*. 2024. URL: `https://www.rehacare.de/de/business/cureostiy-pilotprojekt-virtual-reality-therapie` (visited on 11/08/2024).

[57] Rachel G Curtis, Bethany Bartel, Ty Ferguson, Henry T Blake, Celine Northcott, Rosa Virgara, and Carol A Maher. "Improving User Experience of Virtual Health Assistants: Scoping Review". In: *Journal of medical Internet research* 23.12 (2021), e31737.

[58] Elaine Czech, Mina Shibasaki, Keitaro Tsuchiya, Roshan L Peiris, and Kouta Minamizawa. "Discovering Narratives: Multi-sensory Approach Towards Designing with People with Dementia". In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pages 1–8. DOI: `10.1145/3334480.3375209`.

[59] Ryan Dahl. *NodeJS*. 2009. URL: `https://nodejs.org/en/` (visited on 11/08/2024).

[60] Jiamin Dai and Karyn Moffatt. "Surfacing the voices of people with dementia: Strategies for effective inclusion of proxy stakeholders in qualitative research". In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. 2021, pages 1–13.

[61] Zhengyan Dai and Karl F. MacDorman. "The doctor's digital double: how warmth, competence, and animation promote adherence intention." eng. In: *PeerJ. Computer science* 4 (2018), e168.

[62] *Das DiGa Verzeichnis*. 2024. URL: `https://diga.bfarm.de/` (visited on 11/08/2024).

[63] Rick Davidson and Ron Henderson. "Electronic performance monitoring: A laboratory investigation of the influence of monitoring and difficulty on task performance, mood state, and self-reported stress levels". In: *Journal of Applied Social Psychology* 30.5 (2000), pages 906–920.

[64] Statista Research Department. *Distribution of opinions of adults on the preferred use of virtual reality (VR) for personal fitness training in Great Britain (GB) in 2019.* 2019. URL: `https://www.statista.com/statistics/1137335/distribution-of-opinions-on-the-preferred-use-of-vr-for-personal-fitness-training/` (visited on 07/13/2021).

[65] Statista Research Department. *Share of adults who watch more online exercise videos due to social distancing during the coronavirus pandemic in the United States as of March 26, 2020.* 2020. URL: `https://www.statista.com/statistics/1108538/online-fitness-video-usage-during-coronavirus-usa/` (visited on 11/08/2024).

[66] Statista Research Department. *Share of individuals worldwide accessing the internet via computers as of 2nd quarter 2024, by age group and gender.* 2024. URL: `https://www.statista.com/statistics/1380363/users-going-online-via-computers-by-age-and-gender/` (visited on 11/08/2024).

[67] Anna Dickinson, John Arnott, and Suzanne Prior. "Methods for human–computer interaction research with older people". In: *Behaviour & Information Technology* 26.4 (2007), pages 343–352.

[68] Angela DiDomenico and Maury A Nussbaum. "Interactive effects of physical and mental workload on subjective workload assessment". In: *International journal of industrial ergonomics* 38.11-12 (2008), pages 977–983.

[69] Mike van Diest, Jan Stegenga, Heinrich J Wörtche, Gijsbertus J Verkerke, Klaas Postema, and Claudine JC Lamoth. "Exergames for unsupervised balance training at home: a pilot study in healthy older adults". In: *Gait & posture* 44 (2016), pages 161–167.

[70] Ralf Doerner. "Perspectives for Using Smart Augmented Reality for the Future in Social Computing and Collaborative Assistance". In: *International Conference on Human-Computer Interaction*. Springer. 2022, pages 97–109.

[71] Benjamin Doré, Alex Gaudreault, Gauthier Everard, Johannes C Ayena, Ahmad Abboud, Nicolas Robitaille, and Charles Sebiyo Batcho. "Acceptability, feasibility, and effectiveness of immersive virtual technologies to promote exercise in older adults: A systematic review and meta-analysis". In: *Sensors* 23.5 (2023), page 2506.

[72] Marcia Cristina Nascimento Dourado, Tatiana Belfort, Alexandre Monteiro, Aline Tavares de Lucena, Isabel Barbeito Lacerda, Julia Gaigher, Maria Alice Tourinho Baptista, Michelle Brandt, Nathália Ramos Kimura, Natalie de Souza, et al. "COVID-19: challenges for dementia care and research". In: *Dementia & neuropsychologia* 14.4 (2020), pages 340–344.

[73] Dovetail Research Pty. Ltd. *Dovetail*. 2024. URL: https://dovetailapp.com/ (visited on 11/10/2024).

[74] Brittany F Drazich, Rachel McPherson, Emily F Gorman, Thomas Chan, Jejomar Teleb, Elizabeth Galik, and Barbara Resnick. "In too deep? A systematic literature review of fully-immersive virtual reality and cybersickness among older adults". In: *Journal of the American Geriatrics Society* 71.12 (2023), pages 3906–3915.

[75] Lucile Dupuy, Jean-Arthur Micoulaud-Franchi, and Pierre Philip. "Acceptance of virtual agents in a homecare context: Evaluation of excessive daytime sleepiness in apneic patients during interventions by continuous positive airway pressure (CPAP) providers." eng. In: *Journal of sleep research* 30 (2 2021), e13094.

[76] Lucile Dupuy, Etienne de Sevin, Jean-Arthur Micoulaud-Franchi, and Pierre Philip. "Factors Associated with Acceptance of a Virtual Companion Providing Screening and Advices for Sleep Problems during COVID-19 Crisis". In: *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. New York, NY, USA: Association for Computing Machinery, 2021, pages 48–51. ISBN: 9781450386197. URL: https://doi.org/10.1145/3472306.3478334.

[77] Ezequiel Duque, Guilherme Fonseca, Heitor Vieira, Gustavo Gontijo, and Lucila Ishitani. "A systematic literature review on user centered design and participatory design with older people". In: *Proceedings of the 18th Brazilian symposium on human factors in computing systems*. 2019, pages 1–11.

[78] Katherine Easton, Stephen Potter, Remi Bec, Matthew Bennion, Heidi Christensen, Cheryl Grindell, Bahman Mirheidari, Scott Weich, Luc de Witte, Daniel Wolstenholme, and Mark S. Hawley. "A Virtual Agent to Support Individuals Living With Physical and Mental Comorbidities: Co-Design and Acceptability Testing." eng. In: *Journal of medical Internet research* 21 (5 2019), e12996.

[79] Patrick Eggenberger, Vera Schumacher, Marius Angst, Nathan Theill, and Eling D de Bruin. "Does multicomponent physical exercise with simultaneous cognitive training boost cognitive performance in older adults? A 6-month randomized controlled trial with a 1-year follow-up". In: *Clinical interventions in aging* (2015), pages 1335–1349.

[80] Mahzar Eisapour, Shi Cao, and Jennifer Boger. "Game design for users with constraint: exergame for older adults with cognitive impairment". In: *Adjunct Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 2018, pages 128–130.

[81] Mahzar Eisapour, Shi Cao, Laura Domenicucci, and Jennifer Boger. "Participatory design of a virtual reality exercise for people with mild cognitive impairment". In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, pages 1–9. DOI: 10.1145/3170427.3174362.

[82] Mahzar Eisapour, Shi Cao, Laura Domenicucci, and Jennifer Boger. "Virtual reality exergames for people living with dementia based on exercise therapy best practices". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Volume 62. 1. SAGE Publications Sage CA: Los Angeles, CA. 2018, pages 528–532. DOI: 10.1177/1541931218621120.

[83] Russell A Epstein, Eva Zita Patai, Joshua B Julian, and Hugo J Spiers. "The cognitive map in humans: spatial navigation and beyond". In: *Nature neuroscience* 20.11 (2017), pages 1504–1513.

[84] Nazife Ertugrul, Lucie Kruse, and Frank Steinicke. *Make Me Run: The Effects of an Immersive Learning Experience on Physical Running Exercise*. GI VR / AR Workshop. 2024.

[85] Colin A. Espie, Simon D. Kyle, Chris Williams, Jason C. Ong, Neil J. Douglas, Peter Hames, and June S. L. Brown. "A randomized, placebo-controlled trial of online cognitive behavioral therapy for chronic insomnia disorder delivered via an automated media-rich web application." eng. In: *Sleep* 35 (6 2012), pages 769–81.

[86] Facebook Technologies, LLC. *Meta Quest*. 2022. URL: https://www.meta.com/ (visited on 11/08/2024).

[87] Gabi Fastner. *13 min. sanfte Senioren-Gymnastik im Stehen, ohne Hilfsmittel*. 2019. URL: https://youtu.be/IRZXXcjfgOs (visited on 09/15/2021).

[88] Gabi Fastner. *16 min. Senioren-Gymnastik am Morgen | ohne Geräte, im Stehen*. 2020. URL: https://youtu.be/BQnxFLkXAAI (visited on 09/15/2021).

[89] Pedro Guillermo Feijóo-Garcıéa, Mohan Zalake, Alexandre Gomes de Siqueira, Benjamin Lok, and Felix Hamza-Lup. "Effects of Virtual Humans' Gender and Spoken Accent on Users' Perceptions of Expertise in Mental Wellness Conversations". In: *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. New York, NY, USA: Association for Computing Machinery, 2021, pages 68–75. ISBN: 9781450386197. URL: https://doi.org/10.1145/3472306.3478367.

[90] David Felce and Jonathan Perry. "Quality of life: Its definition and measurement". In: *Research in developmental disabilities* 16.1 (1995), pages 51–74.

[91] Ylva Ferstl, Sean Thomas, Cédric Guiard, Cathy Ennis, and Rachel McDonnell. "Human or Robot? Investigating voice, appearance and gesture motion realism of conversational social agents". In: *Proceedings of the 21st ACM international conference on intelligent virtual agents*. 2021, pages 76–83.

[92] Paul M Fitts. "The information capacity of the human motor system in controlling the amplitude of movement." In: *Journal of experimental psychology* 47.6 (1954), page 381.

[93] Jeana Frost, Nora Boukris, and Peter Roelofsma. “We like to Move It Move It! Motivation and Parasocial Interaction”. In: *CHI ’12 Extended Abstracts on Human Factors in Computing Systems*. Chi Ea ’12. Austin, Texas, USA: Association for Computing Machinery, 2012, pages 2465–2470. ISBN: 9781450310161. DOI: 10.1145/2212776.2223820. URL: https://doi.org/10.1145/2212776.2223820.

[94] Patrick D Gajewski and Michael Falkenstein. “Physical activity and neurocognitive functioning in aging-a condensed updated review”. In: *European Review of Aging and Physical Activity* 13.1 (2016), pages 1–7. DOI: 10.1186/s11556-016-0161-3.

[95] Matthias Gamer, Jim Lemon, Maintainer Matthias Gamer, A Robinson, and W Kendall’s. “Package ‘irr’”. In: *Various coefficients of interrater reliability and agreement* 22 (2012), pages 1–32.

[96] Beat Games. *Beat Saber*. 2022. URL: https://beatsaber.com/ (visited on 11/08/2024).

[97] Rebeca I Garcia-Betances, Viveca Jiménez-Mixco, Maria T Arredondo, and Marıéa F Cabrera-Umpiérrez. “Using virtual reality for cognitive training of the elderly”. In: *American Journal of Alzheimer’s Disease & Other Dementias®* 30.1 (2015), pages 49–54. DOI: 10.1177/1533317514545866.

[98] Mauro de Gennaro, Eva G. Krumhuber, and Gale Lucas. “Effectiveness of an Empathic Chatbot in Combating Adverse Effects of Social Exclusion on Mood.” eng. In: *Frontiers in psychology* 10 (2019), page 3061.

[99] Kathrin Maria Gerling, Jonas Schild, and Maic Masuch. “Exergame design for elderly users: the case study of SilverBalance”. In: *Proceedings of the 7th International Conference on Advances in Computer Entertainment Technology*. 2010, pages 66–69.

[100] Kathrin Maria Gerling, Jonas Schild, and Maic Masuch. “Exergaming for elderly persons: analyzing player experience and performance”. In: *Mensch & Computer 2011: überMEDIEN| ÜBERmorgen* (2011).

[101] Bundesministerium für Gesundheit. *Diagnose Demenz: Krankheitsbild und Verlauf*. 2021. URL: https://www.bundesgesundheitsministerium.de/themen/pflege/online-ratgeber-demenz/krankheitsbild-und-verlauf (visited on 08/04/2024).

[102] Bundesministerium für Gesundheit. *Die Behandlung von Demenz*. 2024. URL: https://www.bundesgesundheitsministerium.de/themen/pflege/online-ratgeber-demenz/die-behandlung.html (visited on 11/08/2024).

[103] Aideen Gibson, Claire McCauley, Maurice Mulvenna, Assumpta Ryan, Liz Laird, Kevin Curran, Brendan Bunting, Finola Ferry, and Raymond Bond. “Assessing usability testing for people living with dementia”. In: *Proceedings of the 4th Workshop on ICTs for improving Patients Rehabilitation Research Techniques*. 2016, pages 25–31.

[104] John D. Gould and Clayton Lewis. “Designing for usability: key principles and what designers think”. In: *Commun. ACM* 28.3 (1985), pages 300–311. ISSN: 0001-0782. DOI: 10.1145/3166.3170. URL: https://doi.org/10.1145/3166.3170.

[105] Jonathan Gratch, Gale M. Lucas, Aisha Aisha King, and Louis-Philippe Morency. “It’s Only a Computer: The Impact of Human-Agent Interaction in Clinical Interviews”. In: *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*. Aamas ’14. Paris, France: International Foundation for Autonomous Agents and Multiagent Systems, 2014, pages 85–92. ISBN: 9781450327381.

[106] Manuel Guimarães, Rui Prada, Pedro A Santos, João Dias, Arnav Jhala, and Samuel Mascarenhas. “The impact of virtual reality in the social presence of a virtual agent”. In: *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 2020, pages 1–8.

[107] Andrea Gyarmati. *Situation of Those Working in Elderly Care in the EU and Hungary*. 2022.

[108] Lindsay Hahn, Stephen L. Rathbun, Michael D. Schmidt, Kyle Johnsen, James J. Annesi, and Sun Joo Grace Ahn. “Using Virtual Agents and Activity Monitors to Autonomously Track and Assess Self-Determined Physical Activity Among Young Children: A 6-Week Feasibility Field Study.” eng. In: *Cyberpsychology, behavior and social networking* 23 (7 2020), pages 471–478.

[109] Courtney D Hall, Alan L Smith, and Steven W Keele. “The impact of aerobic activity on cognitive function in older adults: A new synthesis based on the concept of executive control”. In: *European Journal of Cognitive Psychology* 13.1-2 (2001), pages 279–300.

[110] Kerry Hanna, Clarissa Giebel, Hilary Tetlow, Kym Ward, Justine Shenton, Jacqueline Cannon, Aravind Komuravelli, Anna Gaughan, Ruth Eley, Carol Rogers, et al. “Emotional and mental wellbeing following COVID-19 public health measures on people living with dementia and carers”. In: *Journal of Geriatric Psychiatry and Neurology* (2021), page 0891988721996816.

[111] Jonathon Derek Hart, Thammathip Piumsomboon, Gun A Lee, Ross T Smith, and Mark Billinghurst. “Manipulating avatars for enhanced communication in extended reality”. In: *2021 IEEE International Conference on Intelligent Reality (ICIR)*. Ieee. 2021, pages 9–16.

[112] Sandra G Hart. “NASA task load index (TLX)”. In: (1986).

[113] Sandra G Hart and Lowell E Staveland. “Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research”. In: *Advances in psychology*. Volume 52. Elsevier, 1988, pages 139–183.

[114] Judith Hartfill, Jenny Gabel, Lucie Kruse, Susanne Schmidt, Kevin Riebandt, Simone Kühn, and Frank Steinicke. “Analysis of Detection Thresholds for Hand Redirection during Mid-Air Interactions in Virtual Reality”. In: *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*. 2021, pages 1–10.

[115] Michael R Harwell, Elaine N Rubinstein, William S Hayes, and Corley C Olds. “Summarizing Monte Carlo results in methodological research: The one-and two-factor fixed effects ANOVA cases”. In: *Journal of educational statistics* 17.4 (1992), pages 315–339.

[116] Klaus Hauer, Michael Schwenk, Tania Zieschang, Marco Essig, Clemens Becker, and Peter Oster. “Physical training improves motor performance in people with dementia: a randomized controlled trial”. In: *Journal of the American Geriatrics Society* 60.1 (2012), pages 8–15. DOI: 10.1111/j.1532-5415.2011.03778.x.

[117] Yuan He, André Pereira, and Taras Kucherenko. “Evaluating data-driven co-speech gestures of embodied conversational agents through real-time interaction”. In: *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. 2022, pages 1–8.

[118] National Institutes of Health. *What Is Lewy Body Dementia? Causes, Symptoms, and Treatments*. 2021. URL: https://www.nia.nih.gov/health/lewy-body-dementia/what-lewy-body-dementia-causes-symptoms-and-treatments (visited on 11/08/2024).

[119] Fiona Heilemann, Gottfried Zimmermann, and Patrick Münster. “Accessibility guidelines for VR games-A comparison and synthesis of a comprehensive set”. In: *Frontiers in Virtual Reality* 2 (2021), page 697504.

[120] Hospital zum Heiligen Geist. *Spielfreude genießen beim digitalen Silver Gaming*. 2024. URL: https://www.hzhg.de/heilig-geist-am-alsterlauf/silver-gaming (visited on 11/08/2024).

[121] Sharp Helen, Preece Jenny, and Rogers Yvonne. *Interaction design: beyond human-computer interaction*. 2019.

[122] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. “Spleeter: a fast and efficient music source separation tool with pre-trained models”. In: *Journal of Open Source Software* 5.50 (2020). Deezer Research, page 2154. DOI: 10.21105/joss.02154. URL: https://doi.org/10.21105/joss.02154.

[123] Julia Hertel, Jenny Gabel, Lucie Kruse, Marcel Wollborn, and Frank Steinicke. “Co-Design of an Augmented Reality Maintenance Tool for Gas Pressure Regulation Stations”. In: *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. Ieee. 2022, pages 720–724.

[124] Patricia Heyn, Beatriz C Abreu, and Kenneth J Ottenbacher. “The effects of exercise training on elderly persons with cognitive impairment and dementia: a meta-analysis”. In: *Archives of physical medicine and rehabilitation* 85.10 (2004), pages 1694–1704. DOI: 10.1016/j.apmr.2004.03.019.

[125] Kristian Hildebrand, Frank Steinicke, Sebastian Rings, Christopher Kümmel, Philippe Wesberg, and Joachim Villwock. “Vitalabs-a multi-stage platform for the evaluation of virtual therapies”. In: *Proceedings of the IEEE VR Workshop on Applied VR for Enhanced Healthcare (AVEH)*. 2019.

[126] Nicole TM Hill, Loren Mowszowski, Sharon L Naismith, Verity L Chadwick, Michael Valenzuela, and Amit Lampit. “Computerized cognitive training in older adults with mild cognitive impairment or dementia: a systematic review and meta-analysis”. In: *American Journal of Psychiatry* 174.4 (2017), pages 329–340.

[127] James Hodge, Madeline Balaam, Sandra Hastings, and Kellie Morrissey. “Exploring the design of tailored virtual reality experiences for people with dementia”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, pages 1–13. DOI: 10.1145/3173574.3174088.

[128] James Hodge, Kyle Montague, Sandra Hastings, and Kellie Morrissey. “Exploring Media Capture of Meaningful Experiences to Support Families Living with Dementia”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pages 1–14. DOI: 10.1145/3290605.3300653.

[129] Adrian Hoffmann, Birk Diedenhofen, Bruno Verschuere, and Jochen Musch. “A strong validation of the crosswise model using experimentally-induced cheating behavior”. In: *Experimental Psychology* (2015).

[130] Michael Hogan. “Physical and cognitive activity and exercise for older adults: a review”. In: *The International Journal of Aging and Human Development* 60.2 (2005), pages 95–126.

[131] Kenneth Holmqvist. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011.

[132] Anne Hösch. “Simulator Sickness in Fahrsimulationsumgebungen-drei Studien zu Human Factors”. PhD thesis. Technische Universität Ilmenau, 2018.

[133] Marjan Hosseini, Roanne Thomas, Lara Pilutti, Pascal Fallavollita, and Jeffrey W Jutai. “Acceptance of physical activity virtual reality games by residents of long-term care facilities: a scoping review”. In: *Disability and Rehabilitation: Assistive Technology* (2023), pages 1–9.

[134] Maarten Houben, Rens Brankaert, Saskia Bakker, Gail Kenning, Inge Bongers, and Berry Eggen. “The role of everyday sounds in advanced dementia care”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pages 1–14. DOI: 10.1145/3313831.3376577.

[135] Sharpley Hsieh, Samantha Schubert, Christopher Hoon, Eneida Mioshi, and John R Hodges. “Validation of the Addenbrooke’s Cognitive Examination III in frontotemporal dementia and Alzheimer’s disease”. In: *Dementia and geriatric cognitive disorders* 36.3-4 (2013), pages 242–250.

[136] Shangying Hua, Shuangci Jin, and Shengyi Jiang. “The limitations and ethical considerations of chatgpt”. In: *Data intelligence* 6.1 (2024), pages 201–239.

[137] Han-Chung Huang, May-Kuen Wong, Ju Lu, Wei-Fan Huang, and Ching-I Teng. “Can using exergames improve physical fitness? A 12-week randomized controlled trial”. In: *Computers in Human Behavior* 70 (2017), pages 310–316.

[138] Kuo-Ting Huang. “Exergaming executive functions: An immersive virtual reality-based cognitive training for adults aged 50 and older”. In: *Cyberpsychology, Behavior, and Social Networking* 23.3 (2020), pages 143–149. DOI: 10.1089/cyber.2019.0269.

[139] Tobias Huber, Silvan Mertes, Stanislava Rangelova, Simon Flutura, and Elisabeth André. “Dynamic difficulty adjustment in virtual reality exergames through experience-driven procedural content generation”. In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2021, pages 1–8.

[140] *Informationen für DiGa Nutzende*. 2024. URL: https://diga.bfarm.de/de/diga-nutzende (visited on 11/08/2024).

[141] Insta360. *Insta360 Pro*. 2017. URL: https://www.insta360.com/product/insta360-pro (visited on 04/08/2022).

[142] *Invirto*. 2024. URL: https://invirto.de/ (visited on 11/08/2024).

[143] Eduardo Islas-Cota, J Octavio Gutierrez-Garcia, Christian O Acosta, and Luis-Felipe Rodrıéguez. “A systematic review of intelligent assistants”. In: *Future Generation Computer Systems* 128 (2022), pages 45–62.

[144] Andrew F Jarosz and Jennifer Wiley. “What are the odds? A practical guide to computing and reporting Bayes factors”. In: *The Journal of Problem Solving* 7.1 (2014), page 2.

[145] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. “Survey of hallucination in natural language generation”. In: *ACM Computing Surveys* 55.12 (2023), pages 1–38.

[146] Cameron Jones and Benjamin Bergen. “Does GPT-4 pass the Turing test?” In: *arXiv preprint arXiv:2310.20216* (2023).

[147] Onno Kampman, Farhad Bin Siddique, Yang Yang, and Pascale Fung. “Adapting a virtual agent to user personality”. In: *Advanced Social Interaction with Agents*. Springer, 2019, pages 111–118.

[148] Sukran Karaosmanoglu, Sebastian Cmentowski, Lennart E Nacke, and Frank Steinicke. “Born to Run, Programmed to Play: Mapping the Extended Reality Exergames Landscape”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2024, pages 1–28.

[149] Sukran Karaosmanoglu, Lucie Kruse, Sebastian Rings, and Frank Steinicke. “Canoe VR: An Immersive Exergame to Support Cognitive and Physical Exercises of Older Adults”. In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. Chi Ea ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022, pages 1–7. ISBN: 9781450391566. DOI: 10.1145/3491101.3519736. URL: https://doi.org/10.1145/3491101.3519736.

[150] Sukran Karaosmanoglu, Sebastian Rings, Lucie Kruse, Christian Stein, and Frank Steinicke. "Lessons Learned from a Human-Centered Design of an Immersive Exergame for People with Dementia". In: *PACM on Human-Computer Interaction* 5.Chi Play (2021), (accepted). DOI: 10.1145/3474679. URL: https://doi.org/10.1145/3474679.

[151] Sukran Karaosmanoglu, Katja Rogers, Dennis Wolf, Enrico Rukzio, Frank Steinicke, and Lennart E. Nacke. "Feels like Team Spirit: Biometric and Strategic Interdependence in Asymmetric Multiplayer VR Games". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Chi '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445492. URL: https://doi.org/10.1145/3411764.3445492.

[152] Robert E Kass and Adrian E Raftery. "Bayes factors". In: *Journal of the american statistical association* 90.430 (1995), pages 773–795.

[153] Harleen Kaur, Tushar Singh, Yogesh Kumar Arya, and Shalini Mittal. "Physical fitness and exercise during the COVID-19 pandemic: a qualitative enquiry". In: *Frontiers in psychology* 11 (2020), page 2943.

[154] Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness". In: *The international journal of aviation psychology* 3.3 (1993), pages 203–220.

[155] K. Kim, L. Boelling, S. Haesler, J. Bailenson, G. Bruder, and G.F. Welch. "Does a Digital Assistant Need a Body? The Influence of Visual Embodiment and Social Behavior on the Perception of Intelligent Virtual Agents in AR". In: *Proceedings of the IEEE Symposium on Mixed and Augmented Reality (ISMAR)*. 2018, pages 105–114.

[156] Kangsoo Kim, Nahal Norouzi, Tiffany Losekamp, Gerd Bruder, Mindi Anderson, and Gregory Welch. "Effects of Patient Care Assistant Embodiment and Computer Mediation on User Experience". In: *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. 2019, pages 17–177. DOI: 10.1109/aivr46125.2019.00013.

[157] Kangsoo Kim, Nahal Norouzi, Tiffany Losekamp, Gerd Bruder, Mindi Anderson, and Gregory Welch. "Effects of Patient Care Assistant Embodiment and Computer Mediation on User Experience". In: *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. 2019, pages 17–177. DOI: 10.1109/aivr46125.2019.00013.

[158] Oksoo Kim, Yanghee Pang, and Jung-Hee Kim. "The effectiveness of virtual reality for people with mild cognitive impairment or dementia: a meta-analysis". In: *BMC psychiatry* 19.1 (2019), page 219. DOI: 10.1186/s12888-019-2180-x.

[159] Everlyne Kimani, Timothy Bickmore, Ha Trinh, and Paola Pedrelli. "You'll be Great: Virtual Agent-based Cognitive Restructuring to Reduce Public Speaking Anxiety". In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2019, pages 641–647. DOI: 10.1109/acii.2019.8925438.

[160] Jacqueline Kindell, Karen Sage, John Keady, and Ray Wilkinson. "Adapting to conversation with semantic dementia: using enactment as a compensatory strategy in everyday social interaction". In: *International Journal of Language & Communication Disorders* 48.5 (2013), pages 497–507. DOI: 10.1111/1460-6984.12023.

[161] Randy Klaassen, Kim C. M. Bul, Rieks Op den Akker, Gert Jan van der Burg, Pamela M. Kato, and Pierpaolo Di Bitonto. "Design and Evaluation of a Pervasive Coaching and Gamification Platform for Young Diabetes Patients." eng. In: *Sensors (Basel, Switzerland)* 18 (2 2018).

[162] Masatomo Kobayashi, Akihiro Kosugi, Hironobu Takagi, Miyuki Nemoto, Kiyotaka Nemoto, Tetsuaki Arai, and Yasunori Yamada. "Effects of age-related cognitive decline on elderly user interactions with voice-based dialogue systems". In: *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part IV 17*. Springer. 2019, pages 53–74.

[163] Leon Korkmaz, Susanne Schmidt, Lucie Kruse, and Frank Steinicke. "I Talk - You Write. Exploring Asymmetric Text- and Voice-Based Communication Between Humans and Virtual Agents". In: *Proceedings of the 12th International Conference on Human-Agent Interaction*. HAI '24. Swansea, United Kingdom: Association for Computing Machinery, 2024, pages 25–33. ISBN: 9798400711787. DOI: 10.1145/3687272.3688306. URL: https://doi.org/10.1145/3687272.3688306.

[164] Krzysztof Krejtz, Andrew T Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. "Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze". In: *PloS one* 13.9 (2018), e0203629.

[165] Lucie Kruse, Julia Hertel, Fariba Mostajeran, Susanne Schmidt, and Frank Steinicke. "Would You Go to a Virtual Doctor? A Systematic Literature Review on User Preferences for Embodied Virtual Agents in Healthcare". In: *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. Ieee. 2023, pages 672–682.

[166] Lucie Kruse, Sungchul Jung, Richard Chen Li, and Robert W Lindeman. "On the Use of Jumping Gestures for Immersive Teleportation in VR." In: *Icat-egve*. 2020, pages 113–120.

[167] Lucie Kruse, Sukran Karaosmanoglu, Sebastian Rings, Benedikt Ellinger, Daniel Apken, Thandiwe Feziwe Mangana, and Frank Steinicke. "A Long-Term User Study of an Immersive Exergame for Older Adults with Mild Dementia during the COVID-19 Pandemic". In: *ICAT-EGVE 2021 - International Conference*

on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments. Edited by Jason Orlosky, Dirk Reiners, and Benjamin Weyers. The Eurographics Association, 2021, pages 9–18. ISBN: 978-3-03868-142-7. DOI: 10.2312/egve.20211322.

[168] Lucie Kruse, Sukran Karaosmanoglu, Sebastian Rings, Benedikt Ellinger, and Frank Steinicke. "Enabling Immersive Exercise Activities for Older Adults: A Comparison of Virtual Reality Exergames and Traditional Video Exercises". In: *Societies* 11.4 (2021), page 134.

[169] Lucie Kruse, Sukran Karaosmanoglu, Sebastian Rings, and Frank Steinicke. "Evaluating Difficulty Adjustments in a VR Exergame for Younger and Older Adults: Transferabilities and Differences". In: *Symposium on Spatial User Interaction*. 2022.

[170] Lucie Kruse, Lea Knaack, and Frank Steinicke. "Back to (Virtual) Reality: Preferences and Effects of Entry and Exit Transitions to Virtual Experiences for Older Adults". In: *CM Symposium on Spatial User Interaction (SUI '24)*. 2024. DOI: 10.1145/3677386.3688881.

[171] Lucie Kruse, Eike Langbehn, and Frank Steinicke. "I can see on my feet while walking: Sensitivity to translation gains with visible feet". In: *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. Ieee. 2018, pages 305–312.

[172] Lucie Kruse, Eike Langbehn, and Frank Steinicke. "Where are you? Influence of Redirected Walking on Audio-Visual Position Estimation of Co-Located Users". In: *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. Ieee. 2021, pages 550–551.

[173] Lucie Kruse, Fariba Mostajeran, and Frank Steinicke. "High Levels of Visibility of Virtual Agents Increase the Social Presence of Users". In: *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. Ieee. 2023, pages 843–844.

[174] Lucie Kruse, Fariba Mostajeran, and Frank Steinicke. "The Influence of Virtual Agent Visibility in Virtual Reality Cognitive Training". In: *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*. 2023, pages 1–9.

[175] Lucie Kruse, Sebastian Rings, and Frank Steinicke. "My Focus was on the Game: Investigating the Use of AI Assistants in a Virtual Reality Exergame for Older Adults". In: *Extended Abstracts of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI EA '25. New York, NY, USA: Association for Computing Machinery, 2025. DOI: 10.1145/3706599.3719808. URL: https://doi.org/10.1145/3706599.3719808.

[176] Lucie Kruse, Catharina Rudschies, Sebastian Rings, and Frank Steinicke. "Considering Psychotherapy in the Metaverse". In: *EMRN A Metaverse for the Good*. 2024, pages 119–121.

[177] Lucie Kruse, Joel Wittig, Sebastian Finnern, Melvin Gundlach, Niclas Iserlohe, Oscar Ariza, and Frank Steinicke. "Blended Collaboration: Communication and Cooperation Between Two Users Across the Reality-Virtuality Continuum". In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pages 1–8.

[178] Alexandra M Kueider, Jeanine M Parisi, Alden L Gross, and George W Rebok. "Computerized cognitive training with older adults: a systematic review". In: *PloS one* 7.7 (2012), e40588.

[179] Dilshani Kumarapeli, Sungchul Jung, and Robert W Lindeman. "Privacy threats of behaviour identity detection in vr". In: *Frontiers in Virtual Reality* 5 (2024), page 1197547.

[180] Odders Labs. *OhShape*. 2021. URL: `https://ohshapevr.com/` (visited on 11/08/2024).

[181] Daniël Lakens. "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs". In: *Frontiers in psychology* 4 (2013), page 863.

[182] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. "Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues". In: *2019 International Conference on Multimodal Interaction*. 2019, pages 226–234.

[183] Francis Langlois, Thien Tuong Minh Vu, Kathleen Chassé, Gilles Dupuis, Marie-Jeanne Kergoat, and Louis Bherer. "Benefits of physical exercise training on cognition and quality of life in frail older adults". In: *The Journals of Gerontology: Series B* 68.3 (2013), pages 400–404.

[184] Bettina Laugwitz, Theo Held, and Martin Schrepp. "Construction and evaluation of a user experience questionnaire". In: *Symposium of the Austrian HCI and Usability Engineering Group*. Springer. 2008, pages 63–76.

[185] Bettina Laugwitz, Martin Schrepp, and Theo Held. "Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten." In: *MuC*. 2006, pages 125–134.

[186] Joseph J LaViola Jr. "A discussion of cybersickness in virtual environments". In: *ACM Sigchi Bulletin* 32.1 (2000), pages 47–56.

[187] Nora Lehmann, Lucie Kruse, and Frank Steinicke. "Assistance in Virtual Reality Exergames: Preference for Species of Agents in Relation to Personality of Users". In: *Proceedings of Mensch und Computer 2023*. 2023, pages 422–426.

[188] Cynthia LeRouge, Kathryn Dickhut, Christine Lisetti, Savitha Sangameswaran, and Toree Malasanos. "Engaging adolescents in a computer-based weight management program: avatars and virtual coaches could help." eng. In: *Journal of the American Medical Informatics Association : JAMIA* 23 (1 2016), pages 19–28.

[189] Alessandro Liberati, Douglas G Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C Gøtzsche, John PA Ioannidis, Mike Clarke, Philip J Devereaux, Jos Kleijnen, and David Moher. "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration". In: *Journal of clinical epidemiology* 62.10 (2009), e1–e34.

[190] Damian Lilla, Marc Herrlich, Rainer Malaka, and Dennis Krannich. "The influence of music on player performance in exergames for parkinson's patients". In: *International Conference on Entertainment Computing*. Springer. 2012, pages 433–436.

[191] Na Liu and Ruifeng Yu. "Determining effects of virtually and physically present co-actor in evoking social facilitation". In: *Human Factors and Ergonomics in Manufacturing & Service Industries* 28.5 (2018), pages 260–267.

[192] Teresa Liu-Ambrose, Jennifer C Davis, John R Best, Larry Dian, Kenneth Madden, Wendy Cook, Chun Liang Hsu, and Karim M Khan. "Effect of a home-based exercise program on subsequent falls among community-dwelling high-risk older adults after a fall: a randomized clinical trial". In: *Jama* 321.21 (2019), pages 2092–2100.

[193] Teresa Liu-Ambrose, Meghan G Donaldson, Yasmin Ahamed, Peter Graf, Wendy L Cook, Jacqueline Close, Stephen R Lord, and Karim M Khan. "Otago home-based strength and balance retraining improves executive functioning in older fallers: a randomized controlled trial". In: *Journal of the American Geriatrics Society* 56.10 (2008), pages 1821–1830.

[194] Matthew C Lohman, George W Rebok, Adam P Spira, Jeanine M Parisi, Alden L Gross, and Alexandra M Kueider. "Depressive symptoms and memory performance among older adults: results from the ACTIVE memory training intervention". In: *Journal of aging and health* 25.8_suppl (2013), 209s–229s.

[195] Matthew Lombard, Theresa B Ditton, and Lisa Weinstein. "Measuring presence: the temple presence inventory". In: *Proceedings of the 12th annual international workshop on presence*. 2009, pages 1–15.

[196] Rosemarijn Looije, Anna van der Zalm, Mark A. Neerincx, and Robbert-Jan Beun. "Help, I need some body the effect of embodiment on playful learning". In: *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. 2012, pages 718–724. DOI: 10.1109/roman.2012.6343836.

[197] Zhipeng Lu, Wenjin Wang, Wei Yan, Chung Lin Kew, Jinsil Hwaryoung Seo, Marcia Ory, et al. "The application of fully immersive virtual reality on reminiscence interventions for older adults: Scoping review". In: *JMIR Serious Games* 11.1 (2023), e45539.

[198] Martin H Luerssen and Tim Hawke. "Virtual Agents as a Service: Applications in Healthcare". In: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 2018, pages 107–112.

[199] David D Luxton. “Recommendations for the ethical use and design of artificial intelligent care providers”. In: *Artificial intelligence in medicine* 62.1 (2014), pages 1–10.

[200] Bingxin Ma, Jin Yang, Frances Kam Yuet Wong, Arkers Kwan Ching Wong, Tingting Ma, Jianan Meng, Yue Zhao, Yaogang Wang, and Qi Lu. “Artificial intelligence in elderly healthcare: A scoping review”. In: *Ageing Research Reviews* 83 (2023), page 101808.

[201] Birthe Macdonald and Gizem Hülür. “Well-being and loneliness in Swiss older adults during the COVID-19 pandemic: The role of social relationships”. In: *The Gerontologist* 61.2 (2021), pages 240–250.

[202] Seri Maeng, Jin Pyo Hong, Won-Hyoung Kim, Hyeyoung Kim, Seo-Eun Cho, Jae Myeong Kang, Kyoung-Sae Na, Seok-Hee Oh, Jung Woon Park, Jae Nam Bae, et al. “Effects of Virtual Reality-Based Cognitive Training in the Elderly with and without Mild Cognitive Impairment”. In: *Psychiatry investigation* 18.7 (2021), page 619.

[203] Guido Makransky, Philip Wismer, and Richard E Mayer. “A gender matching effect in learning with pedagogical agents in an immersive virtual reality science simulation”. In: *Journal of Computer Assisted Learning* 35.3 (2019), pages 349–358.

[204] Anabela Correia Martins, Cláudia Santos, Catarina Silva, Daniela Baltazar, Juliana Moreira, and Nuno Tavares. “Does modified Otago Exercise Program improves balance in older people? A systematic review”. In: *Preventive medicine reports* 11 (2018), pages 231–239.

[205] Mara Mather and Laura L Carstensen. “Aging and motivated cognition: The positivity effect in attention and memory”. In: *Trends in cognitive sciences* 9.10 (2005), pages 496–502.

[206] Pavagada Mathuranath, Peter Nestor, German Berrios, Wojtek Rakowicz, and John Hodges. “A brief cognitive test battery to differentiate Alzheimer’s disease and frontotemporal dementia”. In: *Neurology* 55.11 (2000), pages 1613–1620.

[207] Deborah Dillon McDonald, Timothy Gifford, and Stephen Walsh. “Effect of a virtual pain coach on older adults’ pain communication: a pilot study.” eng. In: *Pain management nursing : official journal of the American Society of Pain Management Nurses* 12 (1 2011), pages 50–6.

[208] Joan M McDowd and Fergus IM Craik. “Effects of aging and task difficulty on divided attention performance.” In: *Journal of experimental psychology: human perception and performance* 14.2 (1988), page 267.

[209] Andre McGrail, Alex Best, and Stintah. Boat Attack. https://github.com/Unity-Technologies/BoatAttack. 2020. (Visited on 01/06/2022).

[210] Mary L McHugh. “Interrater reliability: the kappa statistic”. In: *Biochemia medica* 22.3 (2012), pages 276–282.

[211] Sumit Mehra, Bart Visser, Nazli Cila, Jantine van den Helder, Raoul HH Engelbert, Peter JM Weijs, and Ben JA Kröse. "Supporting older adults in exercising with a tablet: a usability study". In: *JMIR human factors* 6.1 (2019), e11598.

[212] Sanjana Mendu, Mehdi Boukhechba, Janna R. Gordon, Debajyoti Datta, Edwin Molina, Gloria Arroyo, Sara K. Proctor, Kristen J. Wells, and Laura E. Barnes. "Design of a Culturally-Informed Virtual Human for Educating Hispanic Women about Cervical Cancer." eng. In: *International Conference on Pervasive Computing Technologies for Healthcare : [proceedings].* 2018 (2018), pages 360–366.

[213] Hui Meng, Yang Xu, Jiali Dai, Yang Zhang, Baogeng Liu, and Haibo Yang. "Analyze the psychological impact of COVID-19 among the elderly population in China and make corresponding suggestions". In: *Psychiatry research* 289 (2020), page 112983.

[214] Sarah Mennicken, Oliver Sack, and Martina Ziefle. "People and a virtual doctor's visit: Learning about multiple acceptance aspects of a telemedical scenario". In: *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*. 2011, pages 577–584.

[215] Meta. *Ray-Ban Meta-Smartglasses*. 2024. URL: `https://www.meta.com/de/en/smart-glasses/` (visited on 11/10/2024).

[216] Microsoft. *Kinect for Windows*. 2022. URL: `https://learn.microsoft.com/en-us/windows/apps/design/devices/kinect-for-windows` (visited on 11/08/2024).

[217] Kimberly J Miller, Brooke S Adair, Alan J Pearce, Catherine M Said, Elizabeth Ozanne, and Meg M Morris. "Effectiveness and feasibility of virtual reality and gaming system use at home by older adults for enabling physical activity to improve health-related domains: a systematic review". In: *Age and ageing* 43.2 (2014), pages 188–195. DOI: `10.1093/ageing/aft194`.

[218] Mark Roman Miller, Hanseul Jun, Fernanda Herrera, Jacob Yu Villa, Greg Welch, and Jeremy N Bailenson. "Social interaction in augmented reality". In: *PloS one* 14.5 (2019), e0216290.

[219] Madison Milne-Ives, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando, Edward Meinert, et al. "The effectiveness of artificial intelligence conversational agents in health care: systematic review". In: *Journal of medical Internet research* 22.10 (2020).

[220] Alexander Miloff, Per Carlbring, William Hamilton, Gerhard Andersson, Lena Reuterskiöld, and Philip Lindner. "Measuring Alliance Toward Embodied Virtual Therapists in the Era of Automated Treatments With the Virtual Therapist Alliance Scale (VTAS): Development and Psychometric Evaluation." eng. In: *Journal of medical Internet research* 22 (3 2020), e16660.

[221] Xiongkuo Min, Guangtao Zhai, Ke Gu, Jing Liu, Shiqi Wang, Xinfeng Zhang, and Xiaokang Yang. "Visual attention analysis and prediction on human faces". In: *Information Sciences* 420 (2017), pages 417–430.

[222] Jang Ho Moon, Eunice Kim, Sejung Marina Choi, and Yongjun Sung. “Keep the social in social media: The role of social interaction in avatar-based virtual shopping”. In: *Journal of Interactive Advertising* 13.1 (2013), pages 14–26.

[223] Fariba Mostajeran, Melik Berk Balci, Frank Steinicke, Simone Kühn, and Jürgen Gallinat. “The effects of virtual audience size on social anxiety during public speaking”. In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. Ieee. 2020, pages 303–312.

[224] Fariba Mostajeran, Nadia Burke, Nazife Ertugrul, Kilian Hildebrandt, Joshua Matov, Noémie Tapie, Wilhelm Gottlieb Zittel, Pia Reisewitz, and Frank Steinicke. “Anthropomorphism of Virtual Agents and Human Cognitive Performance in Augmented Reality”. In: *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. Ieee. 2022, pages 329–332.

[225] Fariba Mostajeran, Nikolaos Katzakis, Oscar Ariza, Jann Philipp Freiwald, and Frank Steinicke. “Welcoming a Holographic Virtual Coach for Balance Training at Home: Two Focus Groups with Older Adults”. In: *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. Ieee. 2019, pages 1465–1470. DOI: 10.1109/vr.2019.8797813.

[226] Fariba Mostajeran, Nikolaos Katzakis, Oscar Ariza, Jann Philipp Freiwald, and Frank Steinicke. “Welcoming a holographic virtual coach for balance training at home: two focus groups with older adults”. In: *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. Ieee. 2019, pages 1465–1470.

[227] Fariba Mostajeran, Frank Steinicke, Oscar Javier Ariza Nunez, Dimitrios Gatsios, and Dimitrios Fotiadis. “Augmented reality for older adults: Exploring acceptability of virtual coaches for home-based balance training in an aging population”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pages 1–12.

[228] Mozilla. *What is Accessibility?* 2024. URL: https://developer.mozilla.org/en-US/docs/Learn/Accessibility/What_is_accessibility (visited on 11/08/2024).

[229] Maurice D Mulvenna, Chris D Nugent, Ferial Moelaert, David Craig, Rose-Marie Dröes, and Johan E Bengtsson. “Supporting people with dementia using pervasive healthcare technologies”. In: *Supporting people with dementia using pervasive health technologies*. Springer, 2010, pages 3–14.

[230] John Muñoz, Samira Mehrabi, Yirou Li, Aysha Basharat, Laura E Middleton, Shi Cao, Michael Barnett-Cowan, and Jennifer Boger. “Immersive Virtual Reality Exergames for Persons Living With Dementia: User-Centered Design Study as a Multistakeholder Team During the COVID-19 Pandemic”. In: *JMIR Serious Games* 10.1 (2022), e29987. ISSN: 2291-9279. DOI: 10.2196/29987.

[231] Daniel L Murman. “The impact of age on cognition”. In: *Seminars in hearing*. Volume 36. 03. Thieme Medical Publishers. 2015, pages 111–121. DOI: 10.1055/s-0035-1555115.

[232] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. "The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment". In: *Journal of the American Geriatrics Society* 53.4 (2005), pages 695–699.

[233] National Institute for Health and Care Excellence. *Falls in older people: assessing risk and prevention.* Nice, 2013.

[234] Roberto Navigli, Simone Conia, and Björn Ross. "Biases in large language models: origins, inventory, and discussion". In: *ACM Journal of Data and Information Quality* 15.2 (2023), pages 1–21.

[235] Jakyung Ngeemasara Thapa Hyejin Park, Hansol Yang, Haeun Kim, Minwoo Son, Jihyeon Jang, Hyuntae Lee, and Park. "The effect of virtual reality (VR)-based intervention program on brain and cognition in older adults with mild cognitive impairment (MCI)". In: *Alzheimer's & Dementia* 16 (2020). DOI: `https://doi.org/10.1002/alz.042835`.

[236] Max M North and Jason Rives. "Virtual reality therapy in aid of senior citizens' psychological disorders". In: *Studies in Health Technology and Informatics* (2003), pages 245–247.

[237] Kristine Nowak and Frank Biocca. "The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments". In: *Presence Teleoperators & Virtual Environments* (2003). DOI: `10.1162/105474603322761289`.

[238] Teresa K. O Leary, Elizabeth Stowell, Everlyne Kimani, Dhaval Parmar, Stefan Olafsson, Jessica Hoffman, Andrea G. Parker, Michael K. Paasche-Orlow, and Timothy Bickmore. "Community-Based Cultural Tailoring of Virtual Agents". In: *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. New York, NY, USA: Association for Computing Machinery, 2020. ISBN: 9781450375863. URL: `https://doi.org/10.1145/3383652.3423875`.

[239] Lukas Oberfrank, Lucie Kruse, and Frank Steinicke. "Sphere Saber: A Virtual Reality Exergame to Study Age-Related Differences in Selective Visual Attention and Information Processing". In: *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*. 2023, pages 1–10.

[240] Catherine S Oh, Jeremy N Bailenson, and Gregory F Welch. "A systematic review of social presence: Definition, antecedents, and implications". In: *Frontiers in Robotics and AI* 5 (2018), page 114.

[241] Yoonsin Oh and Stephen Yang. "Defining exergames & exergaming". In: *Proceedings of meaningful play* 2010 (2010), pages 21–23.

[242] Stefan Olafsson, Byron Wallace, and Timothy Bickmore. "Towards a Computational Framework for Automating Substance Use Counseling with Virtual Agents". In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. Aamas '20. Auckland, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, 2020, pages 966–974. ISBN: 9781450375184.

[243] OpenAI. *Gpt-4*. Technical report. 2023.

[244] OpenAI. *OpenAI Gpt4o*. 2024. URL: https://openai.com/index/gpt-4/ (visited on 11/08/2024).

[245] OpenJS Foundation. *NodeJS Express*. 2010. URL: http://expressjs.com/ (visited on 11/08/2024).

[246] World Health Organization. *Dementia*. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/dementia (visited on 11/08/2024).

[247] Sung Park and Richard Catrambone. "Social facilitation effects of virtual humans". In: *Human factors* 49.6 (2007), pages 1054–1060.

[248] Susan Persky, Kimberly A. Kaphingst, Vincent C. Jr Allen, and Ibrahim Senay. "Effects of patient-provider race concordance and smoking status on lung cancer risk perception accuracy among African-Americans." eng. In: *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine* 45 (3 2013), pages 308–17.

[249] Gustav Bøg Petersen, Aske Mottelson, and Guido Makransky. "Pedagogical agents in educational vr: An in the wild study". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pages 1–12.

[250] Manuel Pezzera and N Alberto Borghese. "Dynamic difficulty adjustment in exergames for rehabilitation: a mixed approach". In: *2020 IEEE 8th International Conference on Serious Games and Applications for Health (SeGAH)*. Ieee. 2020, pages 1–7.

[251] Pierre Philip, Lucile Dupuy, Marc Auriacombe, Fushia Serre, Etienne de Sevin, Alain Sauteraud, and Jean-Arthur Micoulaud-Franchi. "Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients." eng. In: *NPJ digital medicine* 3 (2020), page 2.

[252] Pierre Philip, Jean-Arthur Micoulaud-Franchi, Patricia Sagaspe, Etienne De Sevin, Jérôme Olive, Stéphanie Bioulac, and Alain Sauteraud. "Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders." eng. In: *Scientific reports* 7 (2017), page 42656.

[253] Giuseppe Pichierri, Peter Wolf, Kurt Murer, and Eling D de Bruin. "Cognitive and cognitive-motor interventions affecting physical functioning: a systematic review". In: *BMC geriatrics* 11 (2011), pages 1–19.

[254] Vivek Pillai, Jason R. Anderson, Philip Cheng, Luisa Bazan, Sophie Bostock, Colin A. Espie, Thomas Roth, and Christopher L. Drake. "The Anxiolytic Effects of Cognitive Behavior Therapy for Insomnia: Preliminary Results from a Web-delivered Protocol." eng. In: *Journal of sleep medicine and disorders* 2 (2 2015).

[255] Diane Podsiadlo and Sandra Richardson. "The timed "Up & Go": a test of basic functional mobility for frail elderly persons". In: *Journal of the American geriatrics Society* 39.2 (1991), pages 142–148.

[256] Polar Electro. *Polar OH1 Optical Heart Rate Sensor*. 2021. URL: https://www.polar.com/products/accessories/oh1-optical-heart-rate-sensor (visited on 07/19/2021).

[257] Jay Portnoy, Morgan Waller, and Tania Elliott. "Telemedicine in the era of COVID-19". In: *The Journal of Allergy and Clinical Immunology: In Practice* 8.5 (2020), pages 1489–1491.

[258] Prefrontal Cortex. Boat Attack. https://github.com/prefrontalcortex/BoatAttack/tree/boat-attack-vr. 2020. (Visited on 01/13/2022).

[259] Amanda Purington, Jessie G Taft, Shruti Sannon, Natalya N Bazarova, and Samuel Hardman Taylor. "" Alexa is my new BFF" Social Roles, User Satisfaction, and Personification of the Amazon Echo". In: *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 2017, pages 2853–2859.

[260] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), page 9.

[261] Eqbal Radwan, Afnan Radwan, and Walaa Radwan. "Challenges facing older adults during the COVID-19 outbreak". In: *European Journal of Environment and Public Health* 5.1 (2020), em0059.

[262] Adrian E Raftery. "Bayesian model selection in social research". In: *Sociological methodology* (1995), pages 111–163.

[263] Ashwin Rajkumar, Fabio Vulpi, Satish Reddy Bethi, Preeti Raghavan, and Vikram Kapila. "Usability study of wearable inertial sensors for exergames (WISE) for movement assessment and exercise." eng. In: *mHealth* 7 (2021), page 4.

[264] Hedieh Ranjbartabar, Deborah Richards, Ayse Aysin Bilgin, and Cat Kutay. "First Impressions Count! The Role of the Human's Emotional State on Rapport Established with an Empathic versus Neutral Virtual Therapist". In: *IEEE Transactions on Affective Computing* 12.3 (2021), pages 788–800. ISSN: 1949-3045. DOI: 10.1109/taffc.2019.2899305.

[265] Megan Rose Readman, Megan Polden, Melissa Chloe Gibbs, Lettie Wareing, and Trevor J Crawford. "The potential of naturalistic eye movement tasks in the diagnosis of Alzheimer's disease: a review". In: *Brain sciences* 11.11 (2021), page 1503.

[266] Jens Reinhardt, Luca Hillen, and Katrin Wolf. "Embedding conversational agents into ar: Invisible or with a realistic human body?" In: *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction*. 2020, pages 299–310.

[267] Deborah Richards and Patrina Caldwell. “An empathie virtual medical specialist: It’s not what you say but how you say it”. In: *2017 23rd International Conference on Virtual System & Multimedia (VSMM)*. 2017, pages 1–8. DOI: 10.1109/vsmm.2017.8346285.

[268] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. “Improving the accuracy of medical diagnosis with causal machine learning”. In: *Nature communications* 11.1 (2020), pages 1–9.

[269] Raoul Rickenberg and Byron Reeves. “The Effects of Animated Characters on Anxiety, Task Performance, and Evaluations of User Interfaces”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Chi ’00. The Hague, The Netherlands: Association for Computing Machinery, 2000, pages 49–56. ISBN: 1581132166. DOI: 10.1145/332040.332406. URL: https://doi.org/10.1145/332040.332406.

[270] Sebastian Rings, Sukran Karaosmanoglu, Lucie Kruse, Daniel Apken, Tobias Picker, and Frank Steinicke. “Using Exergames to Train Patients with Dementia to Accomplish Daily Routines”. In: *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play*. 2020, pages 345–349.

[271] Sebastian Rings, Lucie Kruse, Marvin Grabowski, Sebastian Rings, and Frank Steinicke. “Meeting Intelligent Agents: How Virtually Anything is Possible with an Intelligent Virtual Agent at the Table”. In: *EMRN A Metaverse for the Good*. 2024, pages 32–34.

[272] Diego Rivera-Gutierrez, Rick Ferdig, Jian Li, and Benjamin Lok. “Getting the Point Across: Exploring the Effects of Dynamic Virtual Humans in an Interactive Museum Exhibit on User Perceptions”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.4 (2014), pages 636–643. ISSN: 1941-0506. DOI: 10.1109/tvcg.2014.26.

[273] Albert Rizzo, Gale Lucas, Jonathan Gratch, Giota Stratou, Louis-Philippe Morency, Kenneth Chavez, Russ Shilling, and Stefan Scherer. “Automatic Behavior Analysis During a Clinical Interview with a Virtual Human.” eng. In: *Studies in health technology and informatics* 220 (2016), pages 316–22.

[274] Clare Robertson, Archibald Campbell, Melinda Gardner, and Nancy Devlin. “Preventing injuries in older people by preventing falls: A meta-analysis of individual-level data”. In: *Journal of the American geriatrics society* 50.5 (2002), pages 905–911.

[275] Yves Rolland, Fabien Pillard, Adrian Klapouszczak, Emma Reynish, David Thomas, Sandrine Andrieu, Daniel Rivière, and Bruno Vellas. “Exercise program for nursing home residents with Alzheimer’s disease: A 1-year randomized, controlled trial”. In: *Journal of the American Geriatrics Society* 55.2 (2007), pages 158–165. DOI: 10.1111/j.1532-5415.2007.01035.x.

[276] Vienna Rose, Inga Stewart, Keith G Jenkins, Luma Tabbaa, Chee Siang Ang, and Maria Matsangidou. “Bringing the outside in: The feasibility of virtual reality with people with dementia in an inpatient psychiatric care setting”. In: *Dementia* 20.1 (2021), pages 106–129.

[277] Catharina Rudschies, Sebastian Rings, Lucie Kruse, Gesche Schauenburg, Hamed Marmarshahi, and Christian-Norbert Zimmer. “Psychotherapy with the Help of ChatGPT? Current Technical and Ethical Boundaries of Intelligent Virtual Agents”. In: *Mensch und Computer 2023 – Workshopband* (2023). DOI: 10.18420/muc2023-mci-ws06-367.

[278] Catharina Rudschies and Ingrid Schneider. “Ethical, legal, and social implications (ELSI) of virtual agents and virtual reality in healthcare”. In: *Social Science & Medicine* (2023), page 116483.

[279] Pninit Russo-Netzer and Hadassah Littman-Ovadia. ““Something to live for”: Experiences, resources, and personal strengths in late adulthood”. In: *Frontiers in psychology* 10 (2019), page 2452.

[280] Richard M Ryan and Edward L Deci. *Intrinsic Motivation Inventory.* 1982. URL: http://selfdeterminationtheory.org/intrinsic-motivation-inventory.

[281] Richard M Ryan, C Scott Rigby, and Andrew Przybylski. “The motivational pull of video games: A self-determination theory approach”. In: *Motivation and emotion* 30.4 (2006), pages 344–360. DOI: 10.1007/s11031-006-9051-8. URL: https://doi.org/10.1007/s11031-006-9051-8.

[282] Richard M. Ryan and Edward L. Deci. “Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions”. In: *Contemporary Educational Psychology* 25.1 (2000), pages 54–67. ISSN: 0361-476x. DOI: https://doi.org/10.1006/ceps.1999.1020. URL: https://www.sciencedirect.com/science/article/pii/S0361476X99910202.

[283] Alan Rychert, Marıéa Luján Ganuza, and Matias Nicolás Selzer. “Integrating GPT as an Assistant for Low-Cost Virtual Reality Escape-Room Games”. In: *IEEE Computer Graphics and Applications* (2024).

[284] Zelai Sáenz-de-Urturi, Begoña Garcia Zapirain, and Amaia Méndez Zorrilla. “Kinect-based virtual game for motor and cognitive rehabilitation: a pilot study for older adults”. In: *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare.* 2014, pages 262–265.

[285] *Safety and Responsibility.* 2024. URL: https://openai.com/safety/ (visited on 11/08/2024).

[286] Sara Santini, Vera Stara, Flavia Galassi, Alessandra Merizzi, Cornelia Schneider, Sabine Schwammer, Elske Stolte, and Johannes Kropf. “User Requirements Analysis of an Embodied Conversational Agent for Coaching Older Adults to Choose Active and Healthy Ageing Behaviors during the Transition to Retirement: A Cross-National User Centered Design Study.” eng. In: *International journal of environmental research and public health* 18 (18 2021).

[287] Dimitrios Saredakis, Ancret Szpak, Brandon Birckhead, Hannah AD Keage, Albert Rizzo, and Tobias Loetscher. "Factors associated with virtual reality sickness in head-mounted displays: a systematic review and meta-analysis". In: *Frontiers in human neuroscience* 14 (2020), page 96.

[288] Jeff Sauro and James R Lewis. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann, 2016.

[289] Jenny Schlichtiger, Julius Steffen, Bruno C Huber, and Stefan Brunner. "Physical activity during COVID-19 lockdown in older adults". In: *The Journal of sports medicine and physical fitness* 61.1 (2021). DOI: 10.23736/s0022-4707.20.11726-2.

[290] Thereza Schmelter, Lucie Kruse, Sukran Karaosmanoglu, Sebastian Rings, Frank Steinicke, and Kristian Hildebrand. "Towards More Inclusive and Accessible Virtual Reality: Conducting Large-scale Studies in the Wild". In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pages 1–5.

[291] Emanuel Schmider, Matthias Ziegler, Erik Danay, Luzi Beyer, and Markus Bühner. "Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption." In: *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 6.4 (2010), page 147.

[292] Susanne Schmidt, Oscar Ariza, and Frank Steinicke. "Intelligent blended agents: Reality–virtuality interaction with artificially intelligent embodied virtual humans". In: *Multimodal Technologies and Interaction* 4.4 (2020), page 85.

[293] Susanne Schmidt, Gerd Bruder, and Frank Steinicke. "Effects of virtual agent and object representation on experiencing exhibited artifacts". In: *Computers & Graphics* 83 (2019), pages 1–10.

[294] Susanne Schmidt, Tim Rolff, Henrik Voigt, Micha Offe, and Frank Steinicke. "Natural Expression of a Machine Learning Model's Uncertainty Through Verbal and Non-Verbal Behavior of Intelligent Virtual Agents". In: *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 2024, pages 1–15.

[295] Brenda Schraepen, Hanne Huygelier, Raymond Van Ee, Vero Vanden Abeele, and Céline Gillebert. "Attitudes towards and Usability of Head-mounted Virtual Reality in Older Adults". In: *CHI PLAY, Date: 2017/10/15-2017/10/18, Location: Amsterdam, The Netherlands*. 2017.

[296] Martin Schrepp. *User Experience Questionnaire*. 2018. URL: https://www.ueq-online.org/ (visited on 11/08/2024).

[297] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. "Design and evaluation of a short version of the user experience questionnaire (UEQ-S)". In: *International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108.* (2017).

[298] Ameneh Shamekhi and Timothy Bickmore. “Breathe Deep: A Breath-Sensitive Interactive Meditation Coach”. In: *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*. Pervasive-Health ’18. New York, NY, USA: Association for Computing Machinery, 2018, pages 108–117. ISBN: 9781450364508. DOI: 10.1145/3240925.3240940. URL: https://doi.org/10.1145/3240925.3240940.

[299] Ameneh Shamekhi, Ha Trinh, Timothy W. Bickmore, Tamara R. DeAngelis, Theresa Ellis, Bethlyn V. Houlihan, and Nancy K. Latham. “A Virtual Self-Care Coach for Individuals with Spinal Cord Injury”. In: *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. Assets ’16. Reno, Nevada, USA: Association for Computing Machinery, 2016, pages 327–328. ISBN: 9781450341240. DOI: 10.1145/2982142.2982199. URL: https://doi.org/10.1145/2982142.2982199.

[300] Larissa Shamseer, David Moher, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, and Lesley A Stewart. “Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation”. In: *Bmj* 349 (2015).

[301] Lindsay Alexander Shaw, Burkhard Claus Wünsche, Christof Lutteroth, Stefan Marks, and Rodolphe Callies. “Challenges in virtual reality exergame design”. In: (2015).

[302] Zhonghao Shi, Ellen Landrum, Amy O’Connell, Mina Kian, Leticia Pinto-Alva, Kaleen Shrestha, Xiaoyuan Zhu, and Maja J Matarić. “How Can Large Language Models Enable Better Socially Assistive Human-Robot Interaction: A Brief Survey”. In: *Proceedings of the AAAI Symposium Series*. Volume 3. 1. 2024, pages 401–404.

[303] Kazuhiro Shidara, Hiroki Tanaka, Hiroyoshi Adachi, Daisuke Kanayama, Yukako Sakagami, Takashi Kudo, and Satoshi Nakamura. “Analysis of Mood Changes and Facial Expressions during Cognitive Behavior Therapy through a Virtual Agent”. In: *Companion Publication of the 2020 International Conference on Multimodal Interaction*. ICMI ’20 Companion. Virtual Event, Netherlands: Association for Computing Machinery, 2020, pages 477–481. ISBN: 9781450380027. DOI: 10.1145/3395035.3425223. URL: https://doi.org/10.1145/3395035.3425223.

[304] Kazuhiro Shidara, Hiroki Tanaka, Hiroyoshi Adachi, Daisuke Kanayama, Yukako Sakagami, Takashi Kudo, and Satoshi Nakamura. “Relationship between Mood Improvement and Questioning to Evaluate Automatic Thoughts in Cognitive Restructuring with a Virtual Agent”. In: *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. 2021, pages 1–5. DOI: 10.1109/aciiw52867.2021.9666312.

[305] Greg J Siegle, Naho Ichikawa, and Stuart Steinhauer. “Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses”. In: *Psychophysiology* 45.5 (2008), pages 679–687.

[306] Emily M Simek, Lucy McPhate, and Terry P Haines. “Adherence to and efficacy of home exercise programs to prevent falls: a systematic review and meta-analysis of the impact of exercise program characteristics”. In: *Preventive medicine* 55.4 (2012), pages 262–275.

[307] Emily M Simek, Lucy McPhate, Keith D Hill, Caroline F Finch, Lesley Day, and Terry P Haines. “What are the characteristics of home exercise programs that older adults prefer?: A cross-sectional study”. In: *American Journal of Physical Medicine & Rehabilitation* 94.7 (2015), pages 508–521.

[308] Panote Siriaraya and Chee Siang Ang. “Recreating living experiences from past memories through virtual worlds for people with dementia”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2014, pages 3977–3986. DOI: 10.1145/2556288.2557035.

[309] Mel Slater. “Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1535 (2009), pages 3549–3557.

[310] Mel Slater. “Immersion and the illusion of presence in virtual reality”. In: *British journal of psychology* 109.3 (2018), pages 431–433.

[311] Mel Slater, Beau Lotto, Maria Marta Arnold, and Mariéa Victoria Sánchez-Vives. “How we experience immersive virtual environments: the concept of presence and its measurement”. In: *Anuario de Psicologiéa, 2009, vol. 40, p. 193-210* (2009).

[312] Mel Slater and Martin Usoh. “Body centred interaction in immersive virtual environments”. In: *Artificial life and virtual reality* 1.1994 (1994), pages 125–148.

[313] Mel Slater, Martin Usoh, and Anthony Steed. “Depth of presence in virtual environments”. In: *Presence: Teleoperators & Virtual Environments* 3.2 (1994), pages 130–144.

[314] Jan D Smeddinck, Sandra Siegel, and Marc Herrlich. “Adaptive difficulty in exergames for Parkinson’s disease patients.” In: *Graphics Interface*. 2013, pages 141–148.

[315] Sarah Smith, Donna Lamping, Sube Banerjee, Rowan Harwood, Beth Foley, Paul Smith, Joanna Cook, Joanna Murray, Martin Prince, Enid Levin, et al. “Development of a new measure of health-related quality of life for people with dementia: DEMQOL”. In: *Psychological medicine* 37.5 (2007), pages 737–746.

[316] Sarah Smith, Donna Lamping, Sube Banerjee, Rowan Harwood, Beth Foley, Peter Smith, Joanna Cook, Joanna Murray, Martin Prince, Enid Levin, Anthony Mann, and Martin Knapp. “Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology.” In: *Health Technology Assessment (Winchester, England)* 9.10 (2005), pages 1–iv.

[317] Sphery AG. *Exercube*. 2024. URL: https://exercube.ch/ (visited on 11/10/2024).

[318] Micol Spitale, Silvia Silleresi, Giulia Cosentino, Francesca Panzeri, and Franca Garzotto. “”Whom Would You like to Talk with?”: Exploring Conversational Agents for Children’s Linguistic Assessment”. In: *Proceedings of the Interaction Design and Children Conference*. Idc ’20. London, United Kingdom: Association for Computing Machinery, 2020, pages 262–272. ISBN: 9781450379816. DOI: 10.1145/3392063.3394421. URL: https://doi.org/10.1145/3392063.3394421.

[319] Oskar Stamm and Susan Vorwerg. “Virtual Reality Sickness Evaluation in Exergames for Older Hypertensive Patients: A Comparative Study of Training Methods in a Virtual Environment”. In: *International Conference on Human-Computer Interaction*. Springer. 2021, pages 134–146.

[320] International Organization for Standardization. *ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems*. International Organization for Standardization. Geneva, Switzerland. Geneva, Switzerland, 2019. URL: https://www.iso.org/standard/77520.html.

[321] Kay M Stanney, Robert S Kennedy, and Julie M Drexler. “Cybersickness is not simulator sickness”. In: *Proceedings of the Human Factors and Ergonomics Society annual meeting*. Volume 41. 2. SAGE Publications Sage CA: Los Angeles, CA. 1997, pages 1138–1142.

[322] Carolin Straßmann, Nicole C. Krämer, Hendrik Buschmeier, and Stefan Kopp. “Age-Related Differences in the Evaluation of a Virtual Health Agent’s Appearance and Embodiment in a Health-Related Interaction: Experimental Lab Study.” eng. In: *Journal of medical Internet research* 22 (4 2020), e13726.

[323] Jinglin Sun, Yu Liu, Hao Wu, Peiguang Jing, and Yong Ji. “A novel deep learning approach for diagnosing Alzheimer’s disease based on eye-tracking data”. In: *Frontiers in Human Neuroscience* 16 (2022), page 972773.

[324] Jamie L Tait, Rachel L Duckham, Catherine M Milte, Luana C Main, and Robin M Daly. “Influence of sequential vs. simultaneous dual-task exercise training on cognitive function in older adults”. In: *Frontiers in aging neuroscience* 9 (2017), page 368. DOI: 10.3389/fnagi.2017.00368.

[325] Ada Chung Ying Tam, Amanda Wan Yee Chan, Daphne Sze Ki Cheung, Lily Yuen Wah Ho, Angel Shuk Kwan Tang, Martin Christensen, Mimi Mun Yee Tse, and Rick Yiu Cho Kwan. “The effects of interventions to enhance cognitive and physical functions in older people with cognitive frailty: A systematic review and meta-analysis”. In: *European Review of Aging and Physical Activity* 19.1 (2022), page 19.

[326] Hiroki Tanaka and Satoshi Nakamura. “Virtual Agent Design for Social Skills Training Considering Autistic Traits.” eng. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* 2021 (2021), pages 4953–4956.

[327] Unity Technologies. *Unity Real-Time Development Platform*. 2024. URL: https://unity.com/ (visited on 11/08/2024).

[328] Tuong Thai. "The Influence Of Exergaming On Heart Rate, Perceived Exertion, Motivation To Exercise, And Time Spent Exercising". In: (2019).

[329] The JASP Team. *JASP: A Fresh Way to Do Statistics.* 2018. URL: https://jasp-stats.org/ (visited on 11/10/2024).

[330] Myrthe L. Tielman, Mark A. Neerincx, Rafael Bidarra, Ben Kybartas, and Willem-Paul Brinkman. "A Therapy System for Post-Traumatic Stress Disorder Using a Virtual Agent and Virtual Storytelling to Reconstruct Traumatic Memories." eng. In: *Journal of medical systems* 41 (8 2017), page 125.

[331] Myrthe L. Tielman, Mark A. Neerincx, and Willem-Paul Brinkman. "Design and Evaluation of Personalized Motivational Messages by a Virtual Agent that Assists in Post-Traumatic Stress Disorder Therapy." eng. In: *Journal of medical Internet research* 21 (3 2019), e9240.

[332] Myrthe L. Tielman, Mark A. Neerincx, Marieke van Meggelen, Ingmar Franken, and Willem-Paul Brinkman. "How should a virtual agent present psychoeducation? Influence of verbal and textual presentation on adherence." eng. In: *Technology and health care : official journal of the European Society for Engineering and Medicine* 25 (6 2017), pages 1081–1096.

[333] Myrthe L. Tielman, Mark A. Neerincx, Claudia Pagliari, Albert Rizzo, and Willem-Paul Brinkman. "Considering patient safety in autonomous e-mental health systems - detecting risk situations and referring patients back to human care." eng. In: *BMC medical informatics and decision making* 19 (1 2019), page 47.

[334] Berkan Torpil, Sedef Şahin, Serkan Pekçetin, and Mine Uyanık. "The Effectiveness of a virtual reality-based intervention on cognitive functions in older adults with mild cognitive impairment: a single-blind, RANDOMIZED CONTROLLED TRIAL". In: *Games for Health Journal* 10.2 (2021), pages 109–114.

[335] National Health Service UK. *Sitting Exercises.* 2021. URL: https://www.nhs.uk/live-well/exercise/sitting-exercises/ (visited on 06/30/2021).

[336] David Unbehaun, Konstantin Aal, Daryoush Daniel Vaziri, Peter David Tolmie, Rainer Wieching, David Randall, and Volker Wulf. "Social Technology Appropriation in Dementia: Investigating the Role of Caregivers in Engaging People with Dementia with a Videogame-based Training System". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 2020, pages 1–15. DOI: 10.1145/3313831.3376648.

[337] David Unbehaun, Konstantin Aal, Daryoush Daniel Vaziri, Rainer Wieching, Peter Tolmie, and Volker Wulf. "Facilitating Collaboration and Social Experiences with Videogames in Dementia: Results and Implications from a Participatory Design Study". In: *Proceedings of the ACM on Human-Computer Interaction* 2.Cscw (2018), pages 1–23. DOI: 10.1145/3274444. URL: https://doi.org/10.1145/3274444.

[338] David Unbehaun, Sebastian Taugerbeck, Konstantin Aal, Daryoush Daniel Vaziri, Jasmin Lehmann, Peter Tolmie, Rainer Wieching, and Volker Wulf. "Notes of memories: Fostering social interaction, activity and reminiscence through an interactive music exergame developed for people with dementia and their caregivers". In: *Human–Computer Interaction* (2020), pages 1–34.

[339] David Unbehaun, Daryoush Daniel Vaziri, Konstantin Aal, Rainer Wieching, Peter Tolmie, and Volker Wulf. "Exploring the Potential of Exergames to affect the Social and Daily Life of People with Dementia and their Caregivers". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Chi '18. Montreal QC, Canada: Association for Computing Machinery, 2018, pages 1–15. ISBN: 9781450356206. DOI: 10.1145/3173574.3173636. URL: https://doi.org/10.1145/3173574.3173636.

[340] Martin Usoh, Ernest Catena, Sima Arman, and Mel Slater. "Using presence questionnaires in reality". In: *Presence* 9.5 (2000), pages 497–503.

[341] Stephen Uzor and Lynne Baillie. "Investigating the long-term use of exergames in the home with elderly fallers". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2014, pages 2813–2822.

[342] Karen M Van Leeuwen, Miriam S Van Loon, Fenna A Van Nes, Judith E Bosmans, Henrica CW De Vet, Johannes CF Ket, Guy AM Widdershoven, and Raymond WJG Ostelo. "What does quality of life mean to older adults? A thematic synthesis". In: *PloS one* 14.3 (2019), e0213263.

[343] Ashish Vaswani. "Attention is all you need". In: *Advances in Neural Information Processing Systems* (2017).

[344] Anthony Venning, Madeleine Ce Herd, Tassia K. Oswald, Sabran Razmi, Fiona Glover, Tim Hawke, Victoria Quartermain, and Paula Redpath. "Exploring the acceptability of a digital mental health platform incorporating a virtual coach: The good, the bad, and the opportunities." eng. In: *Health informatics journal* 27 (1 2021), page 1460458221994873.

[345] C Lee Ventola. "Virtual reality in pharmacy: opportunities for clinical, research, and educational applications". In: *Pharmacy and Therapeutics* 44.5 (2019), page 267.

[346] Sonja Vestergaard, Christian Kronborg, and Lis Puggaard. "Home-based video exercise intervention for community-dwelling frail older women: a randomized controlled trial". In: *Aging clinical and experimental research* 20.5 (2008), pages 479–486.

[347] Melissa J Vilaro, Danyell S Wilson-Howard, Mohan S Zalake, Fatemeh Tavassoli, Benjamin C Lok, François P Modave, Thomas J George, Folakemi Odedina, Peter J Carek, and Janice L Krieger. "Key changes to improve social presence of a virtual health assistant promoting colorectal cancer screening informed by a technology acceptance model". In: *BMC Medical Informatics and Decision Making* 21.1 (2021), pages 1–9.

[348] Melissa J. Vilaro, Danyell S. Wilson-Howard, Lauren N. Griffin, Fatemeh Tavassoli, Mohan S. Zalake, Benjamin C. Lok, Francois P. Modave, Thomas J. George, Peter J. Carek, and Janice L. Krieger. "Tailoring virtual human-delivered interventions: A digital intervention promoting colorectal cancer screening for Black women." eng. In: *Psycho-oncology* 29 (12 2020), pages 2048–2056.

[349] Melissa J. Vilaro, Danyell S. Wilson-Howard, Mohan S. Zalake, Fatemeh Tavassoli, Benjamin C. Lok, François P. Modave, Thomas J. George, Folakemi Odedina, Peter J. Carek, and Janice L. Krieger. "Key changes to improve social presence of a virtual health assistant promoting colorectal cancer screening informed by a technology acceptance model." eng. In: *BMC medical informatics and decision making* 21 (1 2021), page 196.

[350] Alessandro Vinciarelli, Paraskevi Chatziioannou, and Anna Esposito. "When the words are not everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls". In: *Frontiers in ICT* 2 (2015), page 4.

[351] Matias Volonte, Andrew T Duchowski, and Sabarish V Babu. "Effects of a virtual human appearance fidelity continuum on visual attention in virtual reality". In: *Proceedings of the 19th ACM international conference on intelligent virtual agents*. 2019, pages 141–147.

[352] Peter Vorderer, Werner Wirth, Feliz Ribeiro Gouveia, Frank Biocca, Timo Saari, Lutz Jäncke, Saskia Böcking, Holger Schramm, Andre Gysbers, Tilo Hartmann, et al. "MEC spatial presence questionnaire". In: *Retrieved Sept* 18 (2004), page 2015. DOI: `10.13140/rg.2.2.26232.42249`.

[353] Andreas Voss. "Motivierte Wahrnehmung: selektive Aufmerksamkeit und entlastende Umdeutungen bei der Aufnahme valenter Informationen". PhD thesis. Trier, Univ., Diss., 2004, 2004.

[354] Sarel van Vuuren and Leora R. Cherney. "A Virtual Therapist for Speech and Language Therapy." eng. In: *International Conference On Intelligent virtual agents* 8637 (2014), pages 438–448.

[355] Sandra Wachter and Brent Mittelstadt. "A right to reasonable inferences: rethinking data protection law in the age of big data and AI". In: *Colum. Bus. L. Rev.* (2019), page 494.

[356] Alexander Walker and Peter Schumacher. "Co-designing an Arthritis Nurse-Call Device, with Lead Aged-Care Users". In: *Advances in Ergonomics in Design: Proceedings of the AHFE 2017 International Conference on Ergonomics in Design, July 17- 21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*. Springer. 2018, pages 274–285.

[357] Janet H Walker, Lee Sproull, and R Subramani. "Using a human face in an interface". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 1994, pages 85–91.

[358] Isaac Wang, Jesse Smith, and Jaime Ruiz. "Exploring virtual agents for augmented reality". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pages 1–12.

[359] Sarah Webber, Steven Baker, and Jenny Waycott. “Virtual visits: reminiscence in residential aged care with digital mapping technologies”. In: *Australasian Journal on Ageing* 40.3 (2021), pages 293–300.

[360] Florian Weidner, Gerd Boettcher, Stephanie Arevalo Arboleda, Chenyao Diao, Luljeta Sinani, Christian Kunert, Christoph Gerhardt, Wolfgang Broll, and Alexander Raake. “A Systematic Review on the Visualization of Avatars and Agents in AR & VR displayed using Head-Mounted Displays”. In: *IEEE Transactions on Visualization and Computer Graphics* (2023).

[361] Clare Wenger, Jaber Gubrium, and James Holstein. “Interviewing older people”. In: *Inside interviewing: New lenses, new concerns* (2003), pages 111–130.

[362] Matthias Wilde, Katrin Bätz, Anastassiya Kovaleva, and Detlef Urhahne. “Überprüfung einer Kurzskala intrinsischer Motivation (KIM)”. In: *Zeitschrift für Didaktik der Naturwissenschaften* 15 (2009).

[363] Sherry L Willis, Sharon L Tennstedt, Michael Marsiske, Karlene Ball, Jeffrey Elias, Kathy Mann Koepke, John N Morris, George W Rebok, Frederick W Unverzagt, Anne M Stoddard, et al. “Long-term effects of cognitive training on everyday functional outcomes in older adults”. In: *Jama* 296.23 (2006), pages 2805–2814.

[364] Danyell Wilson-Howard, Melissa J. Vilaro, Jordan M. Neil, Eric J. Cooks, Lauren N. Griffin, Taylor T. Ashley, Fatemeh Tavassoli, Mohan S. Zalake, Benjamin C. Lok, Folakemi T. Odedina, Francois Modave, Peter J. Carek, Thomas J. George, and Janice L. Krieger. “Development of a Credible Virtual Clinician Promoting Colorectal Cancer Screening via Telehealth Apps for and by Black Men: Qualitative Study.” eng. In: *JMIR formative research* 5 (12 2021), e28709.

[365] Nahathai Wongpakaran, Tinakon Wongpakaran, and Robert Van Reekum. “Discrepancies in Cornell Scale for Depression in Dementia (CSDD) items between residents and caregivers, and the CSDD’s factor structure”. In: *Clinical interventions in aging* 8 (2013), page 641.

[366] World Health Organization. *Building foundations for eHealth: progress of Member States: report of the WHO Global Observatory for eHealth*. World Health Organization, 2006.

[367] World Health Organization. *Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth*. World Health Organization, 2017.

[368] World Health Organization. *Classification of digital health interventions v1.0: a shared language to describe the uses of digital technology for health*. Technical report. World Health Organization, 2018.

[369] World Health Organization. *World health statistics 2019: monitoring health for the SDGs, sustainable development goals*. Technical report. 2019.

[370] Wenge Xu, Hai-Ning Liang, Nilufar Baghaei, Xiaoyue Ma, Kangyou Yu, Xuanru Meng, Shaoyue Wen, et al. "Effects of an Immersive Virtual Reality Exergame on University Students' Anxiety, Depression, and Perceived Stress: Pilot Feasibility and Usability Study". In: *JMIR Serious Games* 9.4 (2021), e29330.

[371] Wenge Xu, Hai-Ning Liang, Qiuyu He, Xiang Li, Kangyou Yu, and Yuzheng Chen. "Results and Guidelines From a Repeated-Measures Design Experiment Comparing Standing and Seated Full-Body Gesture-Based Immersive Virtual Reality Exergames: Within-Subjects Evaluation". In: *JMIR serious games* 8.3 (2020), e17972.

[372] Wenge Xu, Hai-Ning Liang, Kangyou Yu, and Nilufar Baghaei. "Effect of Gameplay Uncertainty, Display Type, and Age on Virtual Reality Exergames". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pages 1–14.

[373] Wenge Xu, Hai-Ning Liang, Zeying Zhang, and Nilufar Baghaei. "Studying the effect of display type and viewing perspective on user experience in virtual reality exergames". In: *Games for health journal* 9.6 (2020), pages 405–414.

[374] Ying Xu, Chi Zhang, Baobao Pan, Qing Yuan, and Xu Zhang. "A portable and efficient dementia screening tool using eye tracking machine learning and virtual reality". In: *NPJ Digital Medicine* 7.1 (2024), page 219.

[375] Hao Yang, Ruihan Wang, Changyu Wang, hui guo hui, Hanlin Cai, Fengying Zhang, Jialin Liu, and Siru Liu. "GPT-4 and Neurologists in Screening for Mild Cognitive Impairment in the Elderly: A Comparative Analysis Study". In: *medRxiv* (2023), pages 2023–12.

[376] yes2next. *Warm up for Seniors*. 2020. URL: `https://youtu.be/9khVr_NtBxI` (visited on 06/30/2021).

[377] Betty Yin, Samuel Bailey, Emma Hu, Milinda Jayarekera, Alex Shaw, and Burkhard C Wünsche. "Tour de Tune 2-Auditory-Game-Motor Synchronisation with Music Tempo in an Immersive Virtual Reality Exergame". In: *2021 Australasian Computer Science Week Multiconference*. 2021, pages 1–10.

[378] Gentle Yoga and Wellness. *Standing Yoga for Seniors - Yoga for Seniors - Standing Yoga - Senior Yoga*. 2021. URL: `https://youtu.be/y5DbDOZl-Kg` (visited on 06/30/2021).

[379] Ha-na Yoo, EunJung Chung, and Byoung-Hee Lee. "The effects of augmented reality-based Otago exercise on balance, gait, and falls efficacy of elderly women". In: *Journal of physical therapy science* 25.7 (2013), pages 797–801.

[380] Boram Yoon, Hyung-il Kim, Gun A Lee, Mark Billinghurst, and Woontack Woo. "The effect of avatar appearance on social presence in an augmented reality remote collaboration". In: *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. Ieee. 2019, pages 547–556.

[381] Atsuya Yoshida, Yasuharu Hagita, Kiyoyuki Yamazaki, and Takami Yamaguchi. "Which do you feel comfortable, interview by a real doctor or by a virtual doctor? A comparative study of responses to inquiries with various psychological intensities, for the development of the Hyper Hospital". In: *Proceedings of 1993 2nd IEEE International Workshop on Robot and Human Communication*. 1993, pages 370–374. DOI: 10.1109/roman.1993.367690.

[382] Ludmiła Zając-Lamparska, Monika Wiłkość-Dębczyńska, Adam Wojciechowski, Marta Podhorecka, Anna Polak-Szabela, Łukasz Warchoł, Kornelia Kędziora-Kornatowska, Aleksander Araszkiewicz, and Paweł Izdebski. "Effects of virtual reality-based cognitive training in older adults living without and with mild dementia: a pretest–posttest design pilot study". In: *BMC research notes* 12.1 (2019), pages 1–8.

[383] Robert B Zajonc. "Social facilitation". In: *Science* 149.3681 (1965), pages 269–274.

[384] Mohan Zalake, Fatemeh Tavassoli, Kyle Duke, Thomas George, Francois Modave, Jordan Neil, Janice Krieger, and Benjamin Lok. "Internet-based tailored virtual human health intervention to promote colorectal cancer screening: design guidelines from two user studies". In: *Journal on Multimodal User Interfaces* 15.2 (2021), pages 147–162.

[385] Mohan Zalake, Fatemeh Tavassoli, Lauren Griffin, Janice Krieger, and Benjamin Lok. "Internet-Based Tailored Virtual Human Health Intervention to Promote Colorectal Cancer Screening: Design Guidelines from Two User Studies". In: *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. Iva '19. Paris, France: Association for Computing Machinery, 2019, pages 73–80. ISBN: 9781450366724. DOI: 10.1145/3308532.3329471. URL: https://doi.org/10.1145/3308532.3329471.

[386] Catherine Amine Zanbaka, Amy Catherine Ulinski, Paula Goolkasian, and Larry F Hodges. "Social responses to virtual humans: implications for future interface design". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2007, pages 1561–1570.

[387] Nan Zeng, Zachary Pope, Jung Eun Lee, and Zan Gao. "Virtual reality exercise for anxiety and depression: A preliminary review of current research in an emerging field". In: *Journal of clinical medicine* 7.3 (2018), page 42.

[388] Hao Zhang, Qiong Wu, Chunyan Miao, Zhiqi Shen, and Cyril Leung. "Towards age-friendly exergame design: The role of familiarity". In: *Proceedings of the annual symposium on computer-human interaction in play*. 2019, pages 45–57.

[389] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. "Opt: Open pre-trained transformer language models". In: *arXiv preprint arXiv:2205.01068* (2022).

[390] Yinan Zhao, Hui Feng, Xinyin Wu, Yan Du, Xiufen Yang, Mingyue Hu, Hongting Ning, Lulu Liao, Huijing Chen, Yishan Zhao, et al. "Effectiveness of exergaming in improving cognitive and physical function in people with mild cognitive impairment or dementia: systematic review". In: *JMIR serious games* 8.2 (2020), e16841.

[391] Shuo Zhou, Timothy Bickmore, Amy Rubin, Catherine Yeksigian, Molly Sawdy, and Steven R Simon. "User gaze behavior while discussing substance use with a virtual agent". In: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 2018, pages 353–354.

[392] Mohammad Zohaib. "Dynamic difficulty adjustment (DDA) in computer games: A review". In: *Advances in Human-Computer Interaction* 2018 (2018).

Part V

Appendix

Appendix A

Supplementary Material: Chapter 5

This chapter includes the questions of the semi-structured interviews for the paper: *A Long-Term User Study of an Immersive Exergame for Older Adults with Mild Dementia.*

A.1 HRQOL and DEMQOL

Measures	Sessions	Control Group					Test Group					Mean		SD	
		CG_1	CG_2	CG_3	CG_4	CG_5	TG_1	TG_2	TG_3	CG_4	TG_5	CG	TG	CG	TG
HRQoL	*Session 1*	0	1	1	1	1	1	1	2	2	3	0.8	1.8	0.45	0.84
	Session 2	0	-	2	2	-	1	-	1	1	3	1.33	1.5	1.15	1
	session 3	0	-	2	0	2	-	1	3	2	2	1	2	1.15	0.82
DEMQoL-Proxy	*Session 1*	92	90	90	107	117	93	102	100	88	118	99.2	95.4	12.24	10.92
	Session 2	88	91	97	107	-	90	-	78	81	99	95.75	82.75	8.38	8.73
	Session 3	106	94	100	88	83	-	104	89	116	115	94.2	101	9.18	12.03

Table A.1: Descriptive values from the HRQOL and DEMQOL-Proxy questionnaire.

A.2 Game-Related Interviews

In this section, we will enumerate the questions that were asked in the semi-structured interviews directly after the last game session.

A.2.1 Participants

1. How would you describe your game experience? / game experience
2. How was your attachment to the game? / game engagement
3. How much fun or boredom did you have while playing? / enjoyment
   (a) If you got bored, what could have been changed to counteract that?
   (b) Would more / different videos have helped? Or were the game mechanics themselves boring?
4. How did you feel when you were in the places you saw? / presence

5. How would you describe your visits to the touristic locations? / presence
6. How do you feel about the camera in your hand? / interaction with game mechanics
7. Did you feel different after playing than right before? Did you feel happier, or less happy?
8. What are your thoughts on the game in general?
   (a) What did you like about the game?
   (b) What did you not like?
9. Would you participate in a fitness game study again? / repeatability
   (a) If yes, why?
   (b) If no, why not?
10. Would you recommend this game to others? / generalization
    (a) If yes, to whom and why?
    (b) If no, why not?
11. What would make the game better? / game mechanics

A.2.2 Caregiver

1. How would you describe the mental (e.g emotional) well-being of the participant before the study? / psychological
2. How would you describe the mental (e.g emotional) well-being of the participant after the study? / psychological
3. How would you describe the cognitive (e.g memory, complex actions, making plans) well-being of the participant before the study?/ cognitive
4. How would you describe the cognitive (e.g memory, complex actions, making plans) well-being of the participant after the study? / cognitive
5. How would you describe the physical (e.g mobility, pain, performance) well-being of the participant before the study? / physical
6. How would you describe the physical (e.g mobility, pain, performance) well-being of the participant after the study? / physical
7. How was the social life of the participant before the study? Was s/he often alone? / social

8. How was the social life of the participant after the study? Was s/he often alone? / social
9. How would you describe the quality of life (e.g., feelings, memory, everyday activities) of the participant before the study? / HRQOL
10. How would you describe the quality of life (e.g., feelings, memory, everyday activities) of the participant after the study? / HRQOL
11. Was the participant's mood different immediately before and immediately after a gaming session? Were they in a more positive or negative mood?
12. Did the participant talk about the study?
    (a) Did the participant talk about it right before or after the study? Or in between?

A.2.3 Healthcare Professional

For each participant:

1. Did the participant have prior experience with VR (VR games, another study, etc...)?
2. How would you describe your game experience? / game experience
3. How would you describe the game experience of the participant? / game experience
4. How would you describe the game engagement of the participant? / game engagement
5. How would you describe the interaction (e.g controlling, like to use the camera) of the participants with the photo camera? / interaction with game mechanics
6. How would you describe the enjoyment of the participant during game-play? / enjoyment
7. How would you describe the participant's presence during the game? / presence
8. Have you observed differences in the participant's mood immediately before and immediately after a gaming session?

Only once for the healthcare professional:

1. What are your thoughts on the game in general?
2. What are your thoughts on teamwork with the participant?

3. Would you recommend this game for people with dementia and their caregivers or relatives?
    (a) If yes, why?
    (b) If no, why not?
4. What would make the game better? / game mechanics

A.3 General Interview (post-game)

The following questions were asked to the caregivers of both the test group and the control group in the week after the last game session.

1. How would you describe the mental (e.g emotional) well-being of the participant before the study? / psychological
2. How would you describe the mental (e.g emotional) well-being of the participant after the study? / psychological
3. How would you describe the cognitive (e.g memory, complex actions, making plans) well-being of the participant before the study?/ cognitive
4. How would you describe the cognitive (e.g memory, complex actions, making plans) well-being of the participant after the study? / cognitive
5. How would you describe the physical (e.g mobility, pain, performance) well-being of the participant before the study? / physical
6. How would you describe the physical (e.g mobility, pain, performance) well-being of the participant after the study? / physical
7. How was the social life of the participant before the study? Was s/he often alone? / social
8. How was the social life of the participant after the study? Was s/he often alone? / social
9. How would you describe the quality of life (e.g., feelings, memory, everyday activities) of the participant before the study? / HRQOL
10. How would you describe the quality of life (e.g., feelings, memory, everyday activities of life) of the participant after the study? / HRQOL
11. Did the participant talk about the study?
    (a) Did the participant talk about it right before or after the study? Or in between?

Appendix B

Supplementary Material: Chapter 6

In detail, the exercises and number of repetitions in the video condition were the following, taken from [168]:

- side arm raises (5×)
- neck lateral flexion stretch (5×)
- neck rotation to the sides (5×)
- neck flexion/extension forward and backward (5×)
- shoulder circles to the front (7×)
- shoulder circles to the front, including elbows (7×)
- shoulder circles to the back (7×)
- shoulder circles to the back including elbows (7×)
- arm circles to the front (6×)
- arm circles to the back (6×)
- torso rotations (7×)
- punches to the upper sides (5×)
- overhead punches (7×)
- side arm raises (5×)

B.1 Correct Bubble Percentage in the Follow-Up Study

Figure B.1 shows the number of correctly collected bubbles for each distance between them. It was fit with a loess-function.

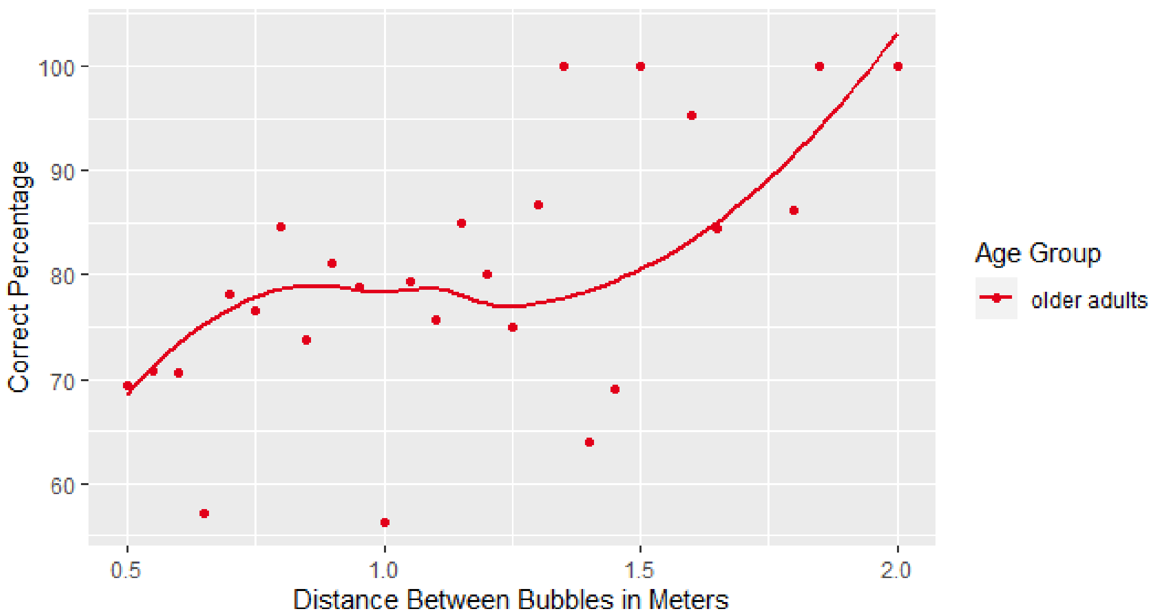


Figure B.1: The percentage of correctly collected bubbles for each spawning time between bubbles, measured during the confirmation study. Fit with a Loess function.

Appendix C

Supplementary Material: Chapter 11

This section shows supplementary information about Chapter 11.

Easy	**Medium**	**Hard**
BAUM	SCHILD	VORRAT
BRUT	TASCHE	UMLAUF
BURG	MORGEN	REPTIL
CHEF	GRUPPE	ANKAUF
CLUB	BAGGER	WINKEL
DAMM	TEUFEL	EGOIST
ECKE	ABWEHR	UNRUHE
EXIL	FLUCHT	SONATE
HAUT	SOMMER	IRRTUM
HEMD	VERBOT	ADVENT
HUND	STRICH	HAUFEN
KAUF	OBJEKT	CHORAL
KEIM	VENTIL	UNHEIL
LAMM	SCHNEE	BENGEL
NAHT	SCHLAG	ROHBAU
NAPF	SYMBOL	KLETTE
NETZ	RITTER	LIZENZ
QUIZ	UMFANG	FREUDE
TAXI	EXPORT	OKTAVE
TURM	ZUCKER	EMPORE

Table C.1: All words used for the anagram solving task, divided by difficulty.

Appendix D

Supplementary Material: Chapter 12

This section presents the supplementary material for Chapter 12.

D.1 GPT Prompts

This section lists the exact prompts given to GPT during the game. These have been translated from German to English for this thesis, using ChatGPT.

D.1.1 General Information

The following prompt was provided to the IVA at the start of the game. It contains all information on the game, the game flow as well as the game rules. We found it beneficial to provide this information right from the start in order to empower the IVA to answer questions and be aware of the game flow before reaching certain events:

You are a parrot named Ara, controlled by artificial intelligence, in a virtual reality game programmed by the Universität Hamburg. Your task is to assist older adults by providing them with tips on interactions and the game rules. Please address me formally and use polite language.

Importantly, the text will be converted to speech, so avoid complex punctuation like asterisks, semicolons, or similar. Also, try to keep explanations fluent and as brief as possible. Use simple words that older adults can easily understand. Speak in German and refrain from using foreign words, English terms, or slang. Do not ask after every sentence if you can assist further.

I will explain to you the rules and setting of the game. I must go through the steps in the given sequence. Explain one step at a time to me and do not jump ahead to future steps. If I ask you questions in between, answer them in an easily understandable manner.

The game takes place on a tropical island. I am wearing a virtual reality headset, the Meta Quest. With it, I perceive the world around me in stereoscopic 3D. I have a controller in each hand, which are small handy technical devices with various buttons. They have trigger buttons. These are located at the back of the controller and can be reached with the index finger.

Step 1: The game begins with a calibration of my seating position. For this, I have to look forward and press the two trigger buttons on the controllers for three seconds.

Step 2: Next, my range of motion is measured with four balloons. These balloons are located above and beside me. If they can't be found, I should look up a bit. The balloons are very sluggish. It's enough to briefly touch them. When I touch them, they move further away from me. If I'm satisfied with the distance, I can simply lower my hands briefly. The blue balloons must be touched with the left hand and the orange balloons with the right hand.

Step 3: Then the real game starts. The game involves paddling down a river in a virtual canoe. Soap bubbles appear in front of me, which can be popped. There are three levels in total, each lasting one minute. There are different rules in the levels. In the first level, all soap bubbles should be destroyed. In the second and third levels, various mental tasks are added.

The game is intended to keep older adults like me fit in advanced age. Targeted movements and mental tasks strengthen my abilities. The game should not be too strenuous. It is not a test, but should be fun. It is played while sitting. The game automatically becomes more difficult if many correct bubbles are popped and automatically becomes easier if I make some mistakes.

The rule of the first level is to touch all soap bubbles. The rule in the second level is to pop all red soap bubbles with the right hand and all yellow soap bubbles with the left hand. The rule in the last level is to alternate between popping a bubble with a letter above it and then a bubble with a number above it.

D.1.2 Situational Prompts

At certain points in the game, additional prompts were given to the agent, independent of the verbal input of the user. These prompts informed the agent about the current state of the game and gave further instructions on how to proceed. Prompts were given after transitioning from passthrough to the immersive environment, after a successful calibration of the sitting position, after calibration of the ROM, once a level was started, and once it ended. Further, the agent made comments if the user played well, or if they made mistakes. The specific prompts are the following:

Introduction: *Introduce yourself with your name Ara and say that you are an artificial parrot. Then ask if I have any questions about the game.*

After calibration: *Finished with the calibration of the seating position. Now explain to me that my range of motion will be measured. For this, I need to push four balloons as far away from me as possible.*

After ROM calibration: *Finished with the measurement of the range of movement. Now there is a start button on the right side. As soon as it is pressed, the canoe will slowly start moving. Point out to me that I should immediately inform you if I start feeling seasick. If I do, it helps to close my eyes and quickly take off the headset.*

After pressing the start button to play a new level: *Explain to me the rule for level number [1,2,3].*

If the user was doing well: *Give me a compliment because I'm playing so well.*

If the user is making mistakes: *Politely point out to me that I'm still making some mistakes and explain the rule of the level again.*

At the end of a level: *The level is over. Ask me if I want to play again or if I want to end the game. The corresponding buttons are on the right side.*

At the end of the game: "*The game is now over. Congratulate me and tell me that I collected [number] correct soap bubbles. Point out to me that I can take off the virtual reality headset now, or I can continue talking with you if I want.*

D.2 Semi-Structured Interview Questions

The following questions were used as a basis for the semi-structured interviews:

1. How did you find entering the virtual world? Did you like the passthrough mode?
2. Were you able to hear everything clearly?
3. Was there anything you didn't like? Do you have any wishes?
4. How did you find it to play with the parrot as a helper?
5. Did you miss having a human contact person?
6. How intelligent did you find the parrot? Did it make mistakes?
7. What else would you have wished for from the parrot?
8. What behaviors should it show more of? What should it stop doing?
9. Have you used AI before, such as GPT? What do you think about it?
10. Can you imagine using AI in real life as well? If so, for what and how?

Appendix E

Usage of Generative AI

OpenAI's ChatGPT was used for correcting grammar and improving the readability of my writing. The assistance of ChatGPT helped to enhance the overall quality of this thesis, while all original research and analysis remains my own. I used UHHGPT with the models *GPT 4 omni* and *GPT 4 omni mini*. Additionally, ChatGPT was used to translate the Abstract of this thesis from English to German, and to translate the prompts used for the study presented in Chapter 12 from German to English. A list of typical sample prompts I used can be found below.

- [**Grammar Correction**] Is it A or B in this sentence?
  Sample Input: Is it adopted or adapted in this: "It has been widely adopted for virtual reality research"?
  Sample Output: In this context, "adopted" is correct. The phrase means that something has been widely accepted or utilized for virtual reality research. "Adapted" would imply that it has been modified or adjusted for a specific purpose, which does not fit the intended meaning here.

- [**Grammar Correction**] Correct this.
  Sample Input: Correct this: "It was no longer save"
  Sample Output: "It was no longer safe."

- [**Correction and Rephrasing**] Is the following sentence correct?
  Sample Input: Is the following sentence correct? Is there a better word for "high"? "It shows a high success in engaging the older adults to exercise."
  Sample Output: The sentence has a small grammatical error; it should be "It shows a high level of success in engaging older adults to exercise." For a better word than "high," you could use "significant," "notable," or "considerable." Here's a revised version of the sentence with one of those options: "It shows a significant level of success in engaging older adults to exercise."

- [**Rephrasing**] Make this sound better.
  Sample Input: Make this sound better: "the Social Presence Survey (SPS), which consists of five questions, which are answered on a Likert-scale"
  Sample Output: "The Social Presence Survey (SPS) comprises five questions, each answered using a Likert scale."

- [**Rephrasing**] Give me a better word for X in this sentence.
  Sample Input: Give me a better word for "deteriorated" in this sentence: Technical

limitations and the inability to perform well in crisis situations deteriorated the view of participants.
Sample Output: Technical limitations and the inability to perform well in crisis situations undermined the view of participants.

- [**Translating**] Translate this to English/German.
  Sample Input: Translate this to English: Also dem Papagei mache ich gar keine Vorwürfe.
  Sample Output: "I don't blame the parrot at all."

Ich bin damit einverstanden, dass meine Arbeit in den Bestand der Bibliothek eingestellt wird.

Hamburg, 24.03.2025

Ort, Datum

Unterschrift

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate. Beispiele für die Nutzung von gKI sind in Appendix E angegeben.

Hamburg, 24.03.2025

Ort, Datum

Unterschrift