Exploring the Potential of Immersive Virtual Reality for Social Science Research

Dissertation

to obtain the doctoral degree of Social Sciences (Dr. rer. soc.)

Faculty of Business, Economics and Social Sciences
University of Hohenheim, 2025

Institute of Communication Science, University of Hohenheim
Institute of Intelligent Interaction and Immersive Experiences,
Karlsruhe University of Applied Sciences

submitted by
Daniel Hepperle

born in
Kirchheim unter Teck, Germany

First Supervisor and primary reviewer

Prof. Dr. Jens Vogelgesang

Second Supervisor and second reviewer

Prof. Dr. Matthias Wölfel

Dean of Faculty Business, Economics and Social Sciences

Prof. Dr. Jörg Schiller

Chair of oral examination

Prof. Dr. Thomas Dimpfl

Day of defence

15.12.2025

Contents

Glossary

AI Artificial Intelligence.

AR Augmented Reality.

EDA Electrodermal Activity.

EEG Electroencephalography.

GIS Geographic Information System.

GPS Global Positioning System.

HCI Human-Computer Interaction.

HIT Human Intelligence Task.

HMD Head-Mounted Display.

IVET Immersive Virtual Environment.

IVR Immersive Virtual Reality.

LLM Large Language Model.

MR Mixed-Reality.

MTurk Amazon Mechanical Turk.

NLP Natural Language Processing.

XR Extended Reality.

Abstract

Immersive Virtual Reality (IVR) holds out the promise of laboratory-grade experimental control while preserving much of the richness of real world experience, yet several issues remain unresolved. The central theme of this dissertation is spanned around the idea of using IVR as a *tool* to help researchers conducting empirical studies in the domain of social sciences.

To address the question, the thesis incorporates five related studies. **Paper 1** introduces the main areas of concern in a typical research process and offers guidance where IVR toolkits might be a valuable addition. Based on those identified areas of concern, the paper suggests solutions such as automation workflows in order to reduce the human-error (i.e. using predefined scenes that already offer different basic standard methods in order to track all changes in the virtual world). **Paper 2** examines seven open-source IVR toolkits, demonstrating how to standardize modular scene setup, participant sensing, and data export. The analysis clarifies the features currently available in different toolkits and provides a basis for researchers to decide which features and toolkits offer the greatest benefits. We also discuss novel features such as AI-based analysis which is not present in most toolkits. Based on this we provide guidance for future IVR-based research software development. **Paper 3** offers a PRISMA-guided systematic scoping review of 56 publications, mapping the field of studies that compare either IVR with the real world or IVR with 2D screens. In short, the review finds that there are more similarities than differences between IVR and the real world. However, between IVR and 2D screens, more findings show differences between the two environments

than similarities. **Paper 4** provides an empirical test of transferability: the mere-exposure effect was successfully replicated in the original study setup (n = 70 m; 49 f) as well as within IVR (n = 39 m; 24 f). Overall, the studies demonstrate the efficacy and practicality of employing IVR to induce effects analogous to those observed in a real-world context in the case of the mere exposure effect. Finally, **Paper 5** introduces *asymmetric normalization,* a novel manipulation that decouples self-perception from how others see a participant in social IVR, thereby expanding the experimental design space with the possibility to reduce bias. This may concern various attributes such as size or age, as well as other visual or spatial characteristics. Pilot data from 40 participants shows that this technique reliably alters interpersonal-distance preferences, opening a new design space for social science research.

This dissertation advances research in the social sciences by showcasing the capabilities of IVR toolkits and illustrating how they can be integrated into established research processes. It further demonstrates that a cognitive-affective mechanism (mere exposure) also is present in IVR. Moreover, it introduces *asymmetric normalization* as a novel manipulation technique that expands the experimental design space beyond what is feasible in physical laboratories.

For research practice, the papers within the dissertation lower the barriers to entry for non-technical scholars, provide a decision matrix for selecting and extending IVR toolkits. Together, they shift IVR from a technological novelty to a mature, shareable, and cost-effective platform for conducting experiments in the social science domain.

Zusammenfassung

Immersive virtuelle Realität (IVR) verspricht Laborbedingungen mit höchster experimenteller Kontrolle, ohne dabei den Reichtum realweltlicher Erfahrung einzubüßen, doch sind noch einige entscheidende Fragen offen. Der rote Faden dieser Dissertation besteht darin, das Potenzial von IVR als *Werkzeug* für Forschende in den Sozialwissenschaften zu evaluieren.

Zur Beantwortung dieser Frage vereint die Arbeit fünf zusammenhängende Studien. **Paper 1** stellt einen automatisierten IVRWorkflow vor, der menschliche Fehler minimiert und *Areas of Concern* entlang eines herkömmlichen Ablaufs der wissenschaftlichen Prozesspipeline darstellt. Auf Basis dieser Problempunkte werden unter anderem Lösungen vorgeschlagen, wie etwa vordefinierte Szenen mit standardisierten Tracking-Methoden, um sämtliche Änderungen in der virtuellen Welt nachvollziehbar zu machen. **Paper 2** untersucht sieben IVR Research Toolkits und zeigt, wie modulare Szenenerstellung, Teilnehmererfassung und Datenexport standardisiert werden können. Die Arbeit stellt den Funktionsumfang verschiedener Toolkits dar, liefert Entscheidungshilfen für Forschende und skizziert fehlende Features wie bsp. KI-basierte Analysen, die als Orientierung für die künftig möglichen und notwendigen Entwicklungen dienen. **Paper 3** präsentiert ein PRISMA-geleitets systematic Scoping-Review über 56 Publikationen, die IVR entweder mit der realen Welt oder mit 2D Bildschirmen vergleichen. In Kürze zeigen sich zwischen IVR und Realität mehr Gemeinsamkeiten als Unterschiede; beim Vergleich IVR versus 2D-Screens hingegen überwiegen die Differenzen. **Paper 4** überprüft die Übertragbarkeit

empirisch: Der Mere-Exposure-Effekt wurde sowohl im ursprünglichen Setting (n = 70 m; 49 f) als auch in IVR (n = 39 m; 24 f) erfolgreich repliziert. **Paper 5** führt *Asymmetric Normalization* ein—eine neuartige Manipulation, die Selbstwahrnehmung und Fremdwahrnehmung eines Teilnehmenden in Social-IVR entkoppelt. Dies kann sich auf verschiedene Aspekte wie Größe oder Alter oder weitere visuelle oder räumliche Aspekte beziehen. Ein Pilotexperiment mit 40 Personen deutet darauf hin, dass sich die Wahrnehmung hinsichtlich interpersoneller Distanzen damit systematisch verändern lassen und so ein neues Untersuchungsfeld entsteht.

Diese Dissertation trägt zur sozialwissenschaftlichen Forschung bei, indem sie die Möglichkeiten von IVR-Toolkits aufzeigt und veranschaulicht, wie diese in etablierte Forschungsprozesse integriert werden können. Darüber hinaus wird gezeigt, dass ein kognitiv-affektiver Mechanismus (Mere-Exposure-Effekt) auch in IVR wirksam ist. Zudem wird mit der *asymmetric normalization* eine neuartige Manipulationstechnik eingeführt, die den experimentellen Gestaltungsraum über das hinaus erweitert, was in physischen Laborumgebungen möglich ist.

In der Forschungspraxis helfen die Arbeiten Einstiegshürden für Fachkolleg*innen zu senken und bieten eine Entscheidungsmatrix zur Auswahl und Erweiterung von IVR-Toolkits. Durch diese Beiträge entwickelt sich IVR insgesamt von einer technischen Neuheit zu einer teilbaren und möglicherweise kosteneffizienteren Plattform für sozialwissenschaftliche Experimente.

Acknowledgements

I would like to express my gratitude to all the people who have accompanied me through all the ups and downs over the years of finishing this dissertation.

First and foremost, I would like to thank my two supervisors, Jens Vogelgesang and Matthias Wölfel. Both Jens and Matthias have consistently encouraged me and supported me with advice and assistance at every step. This work would not have been possible without their support, helpful feedback, and inspiring discussions. A special thank-you goes to Matthias, who revealed how genuinely fun and endlessly interesting science can be, turning every experiment into an adventure—and whose steady support helped me tackle countless challenges along the way. I am proud to have built the iXperience Lab together with you, and I look forward to many more years of inspiring collaboration!

Thanks to my mentor, Tobias Dienlin, for his insightful guidance and enthusiastic tour of the Open-Science universe; he sharpened my thinking and enriched this work.

Furthermore, I would like to thank all my co-authors: Christian Felix Purps, Jonas Deuchler, Wladimir Hettmann, with whom I spent many days and nights (not only at the office) working and enjoying life. It has been a pleasure to collaborate with you, and I wish you great success with your all your future plans.

I would also like to extend my heartfelt thanks to Andi Sieß, with whom I published my first scientific paper. Since then, we have supported one another in our academic paths and every other important part of life.

I am especially grateful to Sarah Eberhard-Bölz, who first introduced me to the University of Hohenheim and later came to my rescue, collecting all the original signatures when I was stranded in traffic.

In addition, I would like to thank the entire team at the Institute for Intelligent Interaction and Immersive Experiences (IIX) at Karlsruhe University of Applied Sciences for the great discussions, exchanges and team spirit.

Another special thanks goes to Steffen Teichmann and his LTE router, without which some experiments probably would not have been possible.

My deepest gratitude is reserved for my family: Sandra, Regina, Rolf, Opa Paul, and Oma Rose. This work would not have been possible without your unwavering support. Your constant encouragement was my motivation, especially during the most challenging times, and you inspired me to finish this disseration. I am profoundly grateful to have you in my life and to know that you will be always there for me.

Thanks to each and every one of you!
Vielen Dank an jede und jeden Einzelnen von euch!

1 Introduction

Understanding human social behavior is a complex undertaking. It is inherently situated within rich, dynamic environments. For over a century, social scientists face a fundamental methodological challenge: the persistent trade-off between ecological validity—the extent to which research findings generalize to real-life settings—and experimental control—the ability to isolate variables and establish causal relationships [1]. The history of social science methods can thus be traced as a series of attempts to renegotiate this trade-off.

1.1 Ecological Validity and Experimental Control

In the late 19th century, Wilhem Wundt pioneered in extending experimental control within the first laboratory experiment in psychology in Leibzig, Germany [2]. Wundt's emphasis on introspection and the measurement of fundamental cognitive processes (e.g., reaction times) under strictly controlled conditions prioritizing experimental rigor over naturalistic observation. His intention was to disassemble conscious experience into its fundamental elements, operating under the belief that comprehending these foundational components was necessary for a comprehensive understanding. While this laboratory-centric approach was essential

for establishing psychology as an empirical science, it possessed inherent limitations. These limitations rendered findings less applicable to the complexities of everyday life and placed less emphasis on ecological validity in its modern sense.

At the same time, researchers looked for ways to capture aspects of real-world thoughts, feelings, and behaviours outside the limits of a laboratory. Questionnaires and surveys saw significant development and widespread adoption, particularly from the early-to-mid 20th century onwards, driven by advances in psychometrics and the demands of sociological research, market analysis, and opinion polls [3]. Walter Lippmann's 1922 work *Public Opinion* underscored urgency of this idea, arguing that individuals act not on objective reality but on socially constructed "pictures in their heads". Basically, mental models shaped by media, culture, and personal experience. To systematically map these subjective perceptions at scale, social scientists turned to tools that could quantify individual and collective attitudes. These tools allowed researchers to gather self-reported data from large samples about attitudes, beliefs, past behaviours, and experiences relevant to real-life contexts. Building on this, Louis L. Thurstone's equal-appearing-interval method standardized attitude measurement in the late 1920s, and Rensis Likert's summated-ratings scale soon offered an even more economical procedure for large-sample surveys [4], [5]. They offered a compromise, potentially increasing ecological relevance compared to purely abstract laboratory tasks. However, it relied heavily on participant memory, honesty, and self-awareness, lacking the direct manipulation and situational control of experiments.

1.2 Methodological Innovations from Field to Digital

Shortly thereafter, alternative perspectives emerged that implicitly challenged this approach and highlighted the importance of context and holistic perception (i.e. "*context is constitutive, not noise*"). Initiated by Max Wertheimer, Kurt Koffka, and Wolfgang Köhler Gestalt psychology emphasizes the role of perception and cognition in understanding the world. Gestalt theorists argued that to comprehend perception and cognition, one must consider the whole rather than merely analyzing individual components: "*the whole is other than the sum of its parts.*" [6].

From the late 1930s through the 1950s, social science shifted between immersive fieldwork and tightly controlled laboratory studies. Muzafer Sherif's Robbers Cave Experiment (1954) depicts the dynamics of intergroup conflict and reconciliation in a real-life summer camp setting [7], while laboratory-based studies such as Solomon Asch's Conformity Line-Judgement Series (1956) revealed the mechanics of social influence under sterile experimental control [8].

Eventually, inspired by this divide, Brunswik [1] formalized the *principle of representative design*, insisting that ecological validity is not an afterthought but a design requirement: Experimental situations, tasks, and stimuli must be sampled from everyday environments with the same statistical rigor researchers apply to sampling participants.

In the 1960s, Webb, Campbell, Schwartz, *et al.* [9] published their work *Unobtrusive Measures: Nonreactive Research in the Social Sciences*, which directly addressed the fundamental tension between ecological validity and experimental control. Their *non-reactive* research methods were designed

to overcome bias introduced by interacting with or observing the participants, nowadays known as the *Hawthorne effect*[1] [11]–[13]. By proposing techniques such as analyzing physical traces, archival records, and disguised observation, Webb, Campbell, Schwartz, *et al.* [9] offered a methodological idea to preserve natural behavior (enhancing ecological validity) while maintaining systematic measurement procedures (retaining experimental rigor).

Two decades later, technlogocial advances in the form of pocket-sized pagers combined with preprinted questionnaires made it possible to capture behavior and feelings in random samples throughout the day. The *Experience Sampling Method* introduced by Larson and Csikszentmihalyi [14] was implemented with pocket pagers that pinged participants at random moments and prompted them to complete a questionnaire. This new sampling method further narrowed the gap between ecological validity and experimental control [14]–[16].

The years from the 1990s through the 2010s brought further technological advances, with portable video recording, GPS logging, and smartphone-based survey tools enabling increasingly sophisticated mixed-method designs [17]. Using such tools, social scientists, particularly in psychology and behavioral science, have been able to capture participants' everyday experiences more directly and in closer alignment with real-time contexts. In 1994 Fahrenberg [18], an early advocate of ambulatory assessment, argued that while medicine was using advanced physiological monitoring systems such as blood pressure monitoring and 24-hour electrocardiogram recording, the social sciences had yet to fully embrace

[1] In a systematic review, McCambridge, Witton, and Elbourne [10] found that while most studies reported some evidence of an effect, significant bias was considered likely due to the complexity of the object of evaluation.

the potential of real-world, continuous data collection for behavioral and emotional processes [18]. A comprehensive review on the potential of smartphones[2] for (contextual) data-collection such as physical proximity, calendar entries or location data in the social science was done by Raento, Oulasvirta, and Eagle [19]. Miller [17] recaptures the first century of the 2000s smartphone usage in social science in 2012: Among other early smartphone-sensing applications are context-triggered surveys and experience sampling [20]; Global Positioning System (GPS)- and Geographic Information System (GIS)-based location tracking [21]; Bluetooth-connected biosensors [22]; acoustic sensing of ambient sounds to infer social context and behaviour [23]; and speaker- and emotion-recognition systems using microphone input [24]. These advances enhanced the external and ecological validity of studies, while innovative research designs leveraged the technology to retain key elements of experimental rigor and control [18], [25].

Smartphones paved the way for wearables devices such as fitness trackers and smartwatches equipped with a range of physiological sensors. These waearable devices can continiously measure physiological parameters, including heart rate, sleep, and Electrodermal Activity (EDA) as well as kinematic data such as posture, acceleration, and step counts. These new types of sensors and algorithms, have been combined using a sensmaking framework combining the wearers' context, physiological data as well as kinematic data in combination with machine learning for mental health studies [26].

[2] Although PDAs had already been used in earlier studies, Raento, Oulasvirta, and Eagle [19] emphasized that the key advantage of smartphones lay in their ease of customization through programmable applications.

Similarly, the rise of the internet and web-based infrastructure in the early 2000s [27] Amazon Mechanical Turk (MTurk) offered researchers a scalable pool of online participants [28]. Crowdsourcing tools such as MTurk and Prolific [29] have substantially accelerated data collection in the behavioral sciences on 2D screens connected to a PC or directly on smartphones and other devices. They offer access to more demographically diverse samples compared to the typical university student populations traditionally relied upon in laboratory studies [30].

However, these online platforms also pose challenges to data quality and control. For instance, a large proportion of MTurk workers have participated in classic behavioral paradigms like the Prisoner's Dilemma or the Ultimatum Game [31], which may reduce the internal validity of studies due to participant non-naïvety[3]. While strategies like filtering for less-experienced workers (e.g., those with under 100 prior Human Intelligence Tasks (HITs)[4]) can mitigate this, concerns persist [32]. More critically, issues with bots and fraudulent respondents have raised alarms about the integrity of online data collection. For example, Webb and Tangney [33] found that only 2.6% (N = 529) of their MTurk data originated from verified human participants. While other authors such as Keith and McKay [34] have responded to the claims made by Webb and Tangney [33] that their research might be "too anecdotal to be true", the issue of bots still exist.

Overall, having access to participants all over the world can increase generalizability of results, yet opens up a challenge regarding cultural

[3] Workers that have been previously been exposed to the experimental stimuli used in the current experiment are referred to as non-naïve [32]

[4] Amazon describes a HIT as "a single, self-contained, virtual task that a Worker can work on, submit an answer, and collect a reward for completing." See: `https://www.mturk.com/worker/help`

differences in perception, interpretation, and behavior, which may influence how participants respond to stimuli, understand instructions, or interpret survey questions.

These cross-cultural variations can introduce unintended biases or variability in the data, making it essential for researchers to carefully consider linguistic nuances, cultural norms, and contextual factors when designing studies for a global audience.

1.3 The Computational Turn and Immersive Technologies

The unprecedented scale and granularity of data generated through digital platforms, smartphones, and wearable devices created both new challenges and opportunities for social scientists. As datasets grew in size and complexity, traditional analytical approaches proved insufficient, leading to the emergence of computational social science—an interdisciplinary field leveraging computational methods to analyze large-scale social phenomena [35].

Machine learning approaches, in particular, offered novel solutions to long-standing methodological challenges. For instance, Natural Language Processing (NLP) techniques enabled researchers to systematically analyze vast quantities of text data from diverse cultural contexts, helping to

identify and adjust for the cross-cultural variations discussed earlier [36] [5]. Supervised classification algorithms allowed for the automated detection of patterns across large datasets while unsupervised learning techniques revealed structures and relationships that might otherwise remain hidden in complex datasets [37].

These computational approaches provided a crucial advantage: They could maintain the ecological validity of naturalistic data collection while introducing new forms of analytical control. For example, researchers could use reinforcement learning to model decision-making processes under varying environmental conditions [38]. This balance addressed Brunswik's longstanding concern about representative design while leveraging the power of computational methods to establish more robust causal inferences.

These computational methods have transformed social science by providing increasingly sophisticated analyses. Machine learning models operated in high-dimensional vector spaces, involving hundreds or thousands of dimensions to represent complex phenomena. Yet ironically, while researchers analyzed data in these rich multidimensional spaces, participants' experiences mostly remained constrained to 2D screens. This dimensional mismatch created new methodological opportunities: Computational social science could model behavior with high complexity,

[5] Grimmer and Stewart [36] highlighted significant pitfalls in automated text analysis at that time, including limitations in processing massive datasets efficiently and challenges in understanding semantic ambiguity. A classic example was the joke "Time flies like an arrow. Fruit flies like a banana." which was difficult for early NLP systems due to its syntactic ambiguity. Such limitations required substantial human supervision and domain knowledge. Modern Large Language Models (LLMs) have since overcome many of these challenges through transformer architectures and self-supervised learning on immense amount of data (something that has been mentioned as limitation in 2013), enabling more nuanced interpretation of textual variations.

but the interfaces through which humans interacted with these models remained fundamentally flat.

IVR technologies emerged as an evolution to resolve this dimensional conflict. By extending human-computer interaction from 2D screens to fully immersive 3D environments with six degrees of freedom (three translational and three rotational). This offeres potential in balancing ecological validity with experimental control—participants potentially providing stimuli in ways closer to natural human movement and perception, while researchers can maintain precise control over environmental parameters.

2 Immersive Virtual Reality

Throughout history, we have developed an evolving collection of tools and artifacts to represent and recreate aspects of reality in order to experience them. From early forms such as language, drawings, and sculpture to more recent inventions like photography, cinema, television, and sound recording, each medium has offered new ways of capturing and displaying the world around us [39].

IVR represents the most recent and technologically sophisticated attempt in this continuum of reality representation, offering potential for creating controlled yet naturalistic research environments. To take full advantage of IVR for social science research, it is important to understand the technological milestones that have shaped its development. By examining these milestones, we can understand how the overall technological progression has increasingly optimized the balance between creating convincing, naturalistic experiences and maintaining the methodological control necessary for rigorous scientific investigation.

2.1 Milestones

While immersive visual concepts such as the illusionistic frescoes of Pompeii have existed since antiquity, panoramic paintings of the 18th and 19th centuries are often seen as the conceptual precursors to modern immersive technologies because of their deliberate architectural and perceptual design to fully surround the viewer and create the sensation "being inside the picture" [40]. Unlike earlier artworks, which were typically confined to walls or ceilings, panoramic paintings occupied specially constructed buildings or rooms[1] and offered 360-degree continuous visual fields, often enhanced by lighting, sound, and perspective tricks to intensify immersion.

Although these installations represent early virtual reality, they achieved limited immersion due to their static nature and reliance solely on visual stimulation. The illusion was enhanced by carefully controlled lighting and strategic placement of three-dimensional objects in the foreground, creating spatial depth, yet viewers remained cognitively aware of observing a painting rather than experiencing presence in the depicted environment.

The development of mechanical and optical devices in the late 19th and early 20th centuries, such as Wheatstone's stereoscope [41] and later, cinematic technologies [39], marked a significant step toward IVR. These systems moved beyond static panoramas by introducing binocular vision and motion over time, simulating depth and narrative progression to

[1] The cylindrical, purpose-built halls that housed panoramas were commonly known as *rotundas*.

heighten the sense of realism. However, like their panoramic predecessors, these technologies still fell short of full immersion. They provided virtual reality in the sense of mediated representations of the real or imagined, but they lacked the interactive, multisensory, and responsive qualities that define IVR today.

The next steps towards IVR began to take shape in the mid-20th century with experimental systems such as the *Headsight Television System* by Comeau and Bryan [42], Morton Heilig's *Sensorama* [43], his *Telesphere Mask* [44], and Ivan Sutherland's *Sword of Damocles* [45].

While the *Sensorama* did not integrate motion tracking, it combined stereoscopic 3D visuals with auditory, haptic (vibration and airflow), and olfactory cues to create an immersive experience. The *Sensorama* exceeded the scope of simple virtual reality by engaging multiple sensory modalities simultaneously, moving significantly closer to IVR despite its pre-digital technology.

Heilig's *Telesphere Mask*, developed in 1960, represented the first Head-Mounted Display (HMD) prototype and another stept towards IVR [44]. One year later, the *Headsight Television System* was the first HMD to go beyond the prototypical stage and be fully fabricated [42], [46]. While providing stereoscopic television viewing with stereo sound, the device lacked head tracking and interactive capabilities, placing it in the virtual reality category without achieving full immersion. The *Telesphere Mask* demonstrated the importance of HMDs for creating intimate visual experiences but highlighted the additional technical requirements necessary for true immersion. However, both devices remained limited by its lack of interactivity or real-time response to user behavior.

Sutherland's new development was the first HMD used for displaying virtual content (mostly lines) over the real world while tracking the movements of the user's head, which by definition would be considered a technique that today would be referred to as Mixed-Reality (MR) or Augmented Reality (AR) Display. Ivan Sutherland's 1965 paper "The Ultimate Display" established the theoretical framework specifically for immersive computing, envisioning interactive systems that could simulate any conceivable environment with perfect sensory fidelity [47]:

> *The ultimate display would, of course, be a room within which the computer can control the existence of matter. A chair displayed in such a room would be good enough to sit in. With appropriate programming such a display could literally be the Wonderland into which Alice walked.*
>
> *— Ivan Sutherland*

This conceptualization moved beyond mere virtual reality toward complete sensory substitution and physical presence.

In 1987, Jaron Lanier introduced and popularized the term virtual reality, applying it to a range of systems without strict distinctions between immersive and non-immersive forms [48]. This broad usage might have contributed to an ambiguity in how the term would later be interpreted across disciplines, with virtual reality coming to encompass everything from HMDs environments to screen-based simulations and data visualizations [48].

The Oculus Rift Kickstarter campaign marked the renaissance specifically of IVR technology [46]. Palmer Luckey's innovation resided not in the

creation of virtual reality, a concept with a history dating back several decades, but rather in the technical feasibility and commercial viability of IVR. The Rift achieved immersion through high(er)-resolution OLED displays, low-latency head tracking, and wide(r) field of view optics that collectively created convincing spatial presence. Facebook's[2] $2.3 billion acquisition in 2014 represented investment specifically in IVR technology rather than virtual reality broadly [46].

With this series of inventions and incremental technological advances, we are already left with the following possibilities:

- The option to experience (real) space, as seen in panoramic paintings or the stereoscope
- Real-time graphics, such as early computer-generated imagery in Sutherland's system
- Real-time interaction, first achieved with Sutherland's so called *Sword of Damocles*, which introduced head tracking and view-dependent rendering

In his public lecture *Is There Any Real Virtue in Virtual Reality?* Frederick P. Brooks Jr. [49] defines virtual reality in terms of three essential features: *real time*, meaning that the viewpoint changes dynamically as the user's head moves; *real space*, referring to three-dimensional environments, whether concrete or abstract; and *real interaction*, as the user's ability to directly manipulate virtual objects.

[2] Facebook, Inc. was rebranded as Meta Platforms, Inc. in 2021.

Also, the Oculus Kickstarter campaing, helped HMDs gain popularity beyond military and academic research contexts because it solved most of the issues mentioned by Brooks [49]. Table 2.1 summarizes the evolution, beginning with the specifications of an early-1990s HMD, followed by the Oculus Rift DK1 from its Kickstarter campaign, and concluding with the current consumer-grade Meta Quest 3 HMD.

Table 2.1: Progress on Brooks' classic VR limitations from the early 1990s to Meta Quest 3.

Limitation	1990s HMD	Meta Quest 3
Poor resolution	360 × 230 px [50]	2 064 × 2 208 px [51]
Narrow field of view	75° Horizontal [50]	110° Horizontal [51]
Limited model complexity	≤ 7,000 polys [52]	1.3-1.8 M polys (Unity guideline) [53]
Poor registration with real world	Polhemus ISOTRAK (magnetic) 6-DoF [54]	Inside-out visual-inertial SLAM (4 IR cams + IMUs) plus IR structured-light/ToF depth; Depth API mesh, spatial anchors; 6-DoF [55], [56]
Bad ergonomics (mass)	935 g (with strap) [57]	515 g (with stock strap) [58]
Tethered ranging	Untethered; Wi-Fi 6E standalone [51]	
Tedious model-building	Hand-built low-poly databases [59]	Photogrammetry, Gaussian Splatting, large asset/character libraries [60]
Price	$ 6000	$ 499

2.2 Immersion and Presence

The before mentioned innovations marked a conceptual shift: from passive observation to embodied experience, where the user was no longer a viewer standing outside the scene but a participant within a simulated environment. The focus moved from visual illusion alone to the construction of presence—a core criterion of immersive virtual reality.

Slater and Wilbur [61] were among the first to distinguish presence from immersion in virtual environments. As Slater [62] defines it, immersion refers to the objective, technical properties of a VR system that deliver a vivid illusion of reality to the user's senses. Immersion serves as the foundation for presence: the subjective feeling of "being there" in the virtual environment [63]. Another interesting approach was done by Lombard and Ditton [64] who define presence as "perceptual illusion of non-mediation". Accordingly, the term captures the user's perception of engaging with the virtual environment or a virtual other as if the interaction occurs directly, without any perceptible technology in between.

Presence has emerged as one of the potentially decisive criteria for conducting meaningful studies within virtual reality environments, allowing researchers to elicit authentic responses in controlled settings. In regard to user studies, Wilson and Soranzo [65] put it as follows: *"Immersion is an objective description of the technical capabilities of the VR system that describes the level of detail with which a virtual environment can be rendered, while presence describes the user's psychological response to said environment."* [65]

There are two main distincions of presence: physcial presence and social presence. Slater describes presence (physical or spatial presence) as the subjective feeling of "being there" in a virtual environment, the illusion of physically being inside the virtual world. Social presence, often also called co-presence is the illusion of being with another person in the shared space. Short, Williams, and Christie [66] originally defined it as the degree to which a communication medium conveys the other person's presence. Interestingly, the concept of high social presence does not depend on photorealistic rendering or perfect physical accuracy. Slater, Pertaub, Barker, *et al.* [67] demonstrated that individuals experiencing

public-speaking anxiety demonstrated comparable physiological arousal and self-reported anxiety when addressing a low-fidelity virtual audience as when confronting a more realistic one. Consequently, even low graphical fidelity can evoke a convincing sense of "being" with others, provided the virtual audience exhibits socially believable behavior.

With these advances as a foundation, we can now focus on how IVR can be used as a research tool within the social sciences[3].

[3] While IVR can be valuable across various disciplines, this work primarily focuses on its applications within social sciences and related sub-disciplines.

3 Theoretical Background: Immersive Virtual Reality as a Research Tool

Pan and C. Hamilton [68] state several areas in which IVR has a potential to contribute in the research toolchain.

- **High experimental control** — IVR lets researchers vary exactly one factor at a time, even in complex social scenarios, which is hard to achieve with live actors.
- **Greater reproducibility** — once a VR scenario is built it can be shared and repeated across laboratories, enabling direct replications
- **Improved ecological validity** — participants can respond naturally in rich, interactive environments while their behaviour is logged unobtrusively (e.g. via motion capture).
- **Access to impossible or risky situations** IVR can safely recreate dangerous contexts (fires, violence, moral dilemmas) or physical transformations that are infeasible in real life.
- **Systematic manipulation of social variables** — virtual characters allow limitless combinations of traits (race, gender, etc.) without confounds such as attractiveness or height.

Even though there are many theoretical possibilities for using IVR, debate remains about how effectively these systems can be applied in practice. Kothgassner and Felnhofer [69] report that simulating social encounters remains particularly challenging. They argue that the main difficulty lies less in representational fidelity than in the lack of standardization across Immersive Virtual Environments (IVETs). Which leads them to conclude that there is an "*immanent need to develop gold-standard procedures*" to assure ecological validity in IVR research.

Despite efforts to improve the quality of studies as described above, significant challenges remain that require continued attention. Many of the innovations in this direction, such as preregistrations, stem from the Open Science movement. This practice aims to enhance transparency in the research process, thereby increasing the likelihood of successful replications. Moreover, it is becoming increasingly common for conference proceedings and journals to require authors to make their data sets and accompanying analyzes publicly available, thereby fostering greater reproducibility. Dienlin, Johannes, Bowman, *et al.* [70] suggested a 7-step open science agenda in the field of communication of which most points can be applied to many other research disciplines as well. The seven steps are as follows:

1. Publish materials, data, and code
2. Preregister studies and submit registered reports
3. Conduct replications
4. Collaborate

5. Foster open science skills

6. Implement Transparency and Openness Promotion Guidelines

7. Incentivize open science practices

In an article in Science magazine, the Open Science Collaboration presented its investigation into the reproducibility of studies in the social science domain. The team found that, whereas 97 % of the 100 original studies reported statistically significant results, only 36 % of the replication attempts yielded significant outcomes. The issue, according to the authors, is not that the original analyses were wrong but that the precise study designs cannot be fully reconstructed and are interpreted differently across cultural contexts [71]. This sereves as a theoretical grounding for **RQ 1** "In which parts of the research process can IVR be the most helpful and how do toolkits contribute to replicability and reproducibility?"

The replication shortfall highlighted by the Open Science Collaboration points to a pressing need for research settings in which stimuli and context can be reproduced with high fidelity [71]. IVR offers great potential to fill this gap. As of today, it offers the potential to elicit (good enough) stimuli while maintaining experimental control—an idea that has been gaining traction for nearly two decades. Already in 2005 Hunt [72] discussed the possibilities of applying IVR in psychology but came to the conclusion that it was almost impossible two reproduce the complexity of real-world situations. This changed in 2015, where Wilson and Soranzo [65] found that IVR now offers almost limitless possibilites for stimuli creation.

This promise of new opportunities, however, is accompanied by significant methodological challenges. Wilson and Soranzo [65] point out that

methodological hurdles extend well beyond standardization alone. In their work they highlight systematic perceptual discrepancies between virtual and real environments that matter whenever the goal is to replicate real-world perception precisely: differences in color, contrast, spatial scale, and motion. Finally, they note the practical constraints of IVR hardware: HMDs and projection systems vary widely in size, cost, and technical complexity, making some setups impractical for certain labs or participant groups. Together with the standardization gap identified by Kothgassner and Felnhofer [69], these factors underline the need for carefully defined "gold standard" protocols before IVR can fulfil its promise as a mainstream tool in psychological research.

Moreover, devising such gold-standard procedures also obliges researchers to benchmark IVR against two critical baselines. The first is the real world itself: only by comparing behaviour and perception in matched physical and virtual settings can we judge whether a given IVR system truly preserves—or intentionally departs from—real-world affordances and sensory cues. The second baseline is the conventional 2D laboratory paradigm, because a large share of existing psychological findings have been obtained on screens. These paradigms might not simply be "ported" to IVR without validating that any observed differences stem from the theoretical manipulation rather than from the change in display medium. Systematic cross-medium comparisons therefore help disentangle genuine psychological effects from artifacts introduced by depth cues, field-of-view, interaction metaphors, or device ergonomics [68]. In short, careful real-world-IVR and 2D-IVR comparisons are essential for accumulating a body of knowledge and for ensuring that results obtained in IVETs can be meaningfully related

to the real world. Based on this, **RQ 2** "What are the similiarities and differences between IVR, the real world and computer screens in studies from current literature?" is important and will be answered within the dissertation.

Carrying on based on these observations, we raise a central question: whether findings drawn in IVR can be meaningful in the real world. If perceptual or experiential discrepancies distort participants' behavior, results may not generalize as expected. At the same time, if immersive environments can replicate key effects reliably, they offer a powerful route toward enhancing ecological validity without sacrificing experimental control. This leads us to **RQ 3** "Are findings obtained in virtual reality transferable to real-world contexts?"

Regarding the aforementioned "gold standard", the need for an extended research pipeline is given also for creating such IVR research environments. For many experiements today that either use IVR

- *As a medium to be investigated for itself* that has its own rules to be understood and compared with other media [73], or
- *As a medium that provides a tool to overcome the need for real world experimentation* [73],

the complete environment, logic and data handling is created anew from scratch. This issue has been addressed by a larged number of toolkits and frameworks that offer to solve basic requirements needed in a large body of experiments [73]. These toolkits aim to provide pre-built features for handling hardware, avatars, data collection, etc., so that researchers can focus on experimental design rather than low-level coding [74].

Because the entire feature set is manitfested in the programming code itself (which often is released as open source, so experts also can contribute to it), such toolkits let researchers distribute an experiment to participants anywhere in the world with minimal friction. Each session can be recorded and revisited from unlimited viewpoints, while annotations may be added both manually by researchers and automatically by algorithms that synchronize real-time physiological data captured through dedicated hardware. These platforms have the potential to function like a "black-box flight recorder" for the experiment by preserving complete information about the experiment, including software versions, logs, stimulus assets, and time-stamped physiological traces. If a replication attempt yields different results, investigators can replay the entire procedure, adjust individual variables, and identify where the divergence occurs.

Additionally, the virtual environment can be manipulated in real time in response to live feedback from physiological or behavioral analysis systems. This capacity for dynamic adaptation introduces new opportunities for optimizing study designs and tailoring experimental conditions to participant behavior. However, it also raises methodological questions about how best to leverage such flexibility without introducing unintended biases or compromising experimental validity. To explore this potential and address these concerns, the following question arises: **RQ 4** "How can IVR factors be manipulated and varied to maximize knowledge gain, and how can potential biases be reduced?"

The open source approach lowers the barrier to reproduction on heterogeneous hardware further, as any lab can create an identical virtual environment or use the one that was shared. In combination with registered reports, IVR studies become auditable research objects that meet

broad reproducibility standards. Ultimately, this end-to-end transparency addresses the replication obstacle highlighted by the Open Science Collaboration [71] and closes the loop from design to dissemination.

While IVR cannot be regarded as a methodological remedy, it represents the most recent stage in the evolution of computer-mediated data-collection techniques and offers a credible pathway toward experiments that couple high ecological validity with rigorous experimental control. However, recognizing this potential requires addressing persistent challenges, including sampling bias, participants' varying levels of familiarity with immersive technologies, and the significant heterogeneity in current VR hardware. In 2002 Fahrenberg, Leonhart, and Foerster [25] wrote: *Acceptance for this new methodology seems to be greater among many patients and study participants than in the professional community of psychology*[1]. Leveraging this user acceptance, together with today's open science infrastructure, establishes VR not as a technological novelty, but as a practical means for narrowing the long-standing trade-off between ecological validity and experimental control.

[1] Original in German: Die Akzeptanz für diese neue Methodik scheint bei vielen Patienten und Untersuchungsteilnehmern größer zu sein als in der Fachwelt der Psychologie.

4 Thesis Overview

This dissertation explores the applicability and, more specifically, the ecological validity of IVR in social science experiments when translated from physical settings into IVR.

Within the dissertation five research papers were written in order to cover these aspects:

1. D. Hepperle, T. Dienlin, and M. Wölfel, "Reducing the human factor in virtual reality research to increase reproducibility and replicability," in *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2021

2. M. Wölfel, D. Hepperle, C. F. Purps, *et al.*, "Entering a new dimension in virtual reality research: An overview of existing toolkits, their features and challenges," in *2021 International Conference on Cyberworlds (CW)*, 2021

3. D. Hepperle and M. Wölfel, "Similarities and differences between immersive virtual reality, real world, and computer screens: A systematic scoping review in human behavior studies," *Multimodal Technologies and Interaction*, vol. 7, no. 6, 2023

4. D. Hepperle and M. Wölfel, "Exploring ecological validity: A comparative study of the mere exposure effect on screens and in immersive virtual reality," in *Advances in Visual Computing - ISVC 2024, LNCS 15047*, G. Bebis, V. Patel, J. Gu, *et al.*, Eds., Springer Nature Switzerland AG, 2025

5. J. Deuchler, D. Hepperle, and M. Wölfel, "Asymmetric normalization in social virtual reality studies," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2022

Paper Overview

The papers are presented in a methodological progression: replicability-oriented integration of IVR into the research pipeline (Paper 1), platform evaluation (Paper 2), a cross-reality scoping review (Paper 3), a preregistered replication study (Paper 4), and an investigation of the concept of asymmetric normalization (Paper 5). This order offers readers the conceptual framework necessary to understand how each subsequent study addresses its respective research question. In Table 4.1 the exact mapping of RQs and Papers is shown.

Table 4.1: Research questions addressed by each paper

	Paper 1	Paper 2	Paper 3	Paper 4	Paper 5
RQ 1	●	●			
RQ 2			●		
RQ 3				●	
RQ 4					●

The first objective is to assess how IVR can help tackle pressing methodological challenges such as the replication crisis and increase ecologicial

validity. By examining IVR as an alternative research platform, we determine how this technology might enrich the research landscape by reducing human-induced errors. By integrating it into a common scientific workflow in **Paper 1**, we apply open science standards into the IVR research process with the help of IVR toolkits in order to help to resolve methodologcial weeknesses. Specifically, we (1) analyze how human error might affect reproducibility and (2) introduce the use of a toolkit designed to streamline the entire IVR experimental pipeline. The toolkit provides guided study setup, automated logging, and standardized export of both raw and processed data.

With the rise of specialized IVR toolkits, **Paper 2** undertakes a comprehensive evaluation of eight representative frameworks: UXF, VREX, EVE, Toggle Toolkit, Cognitive3D's Scene Explorer, Vizard, CyberSession, and our in-house Virtual Reality Scientific Toolkit (VRSTK). Each toolkit is evaluated against four operational dimensions: (1.) setup and control, (2.) participant sensing, (3.) representation, and (4.) data handling. We create a feature matrix that benchmarks each toolkit to identify and compare their features.

Building on the toolkit-centric comparison of **Paper 2**, **Paper 3** shifts the lens from *how* researchers build IVR studies to *what works good* in IVR. In order to address this, **Paper 3** presents a preregistered systematic scoping review that screens 1083 records and analyzes 56 paper involving a comparison of results between head-mounted-display VR (HMD-VR), 2D screen-based set-ups, and matched real-world scenarios. We map the results onto three theoretically motivated categories—*perception*, *interaction*, and *sensing & reconstruction of reality*. Notably, HMD-VR exhibits a higher proportion of similarity to the real world than to screens in

both perception and interaction subdomains, whereas the sensing-and-reconstruction category still shows more differences than similarities when VR is compared with physical reality. These findings supply an essential evidential backdrop to the toolkit and avatar discussions: they indicate that, once technical confounds such are met, immersive VR can approximate real-world behaviour more closely than conventional 2D paradigms, yet they also highlight the domains where methodological innovation remains necessary.

Where **Paper 3** offered a first evidence that IVR often approximates real-world behaviour more closely than 2D screens [77], the next step is to test this conclusion against a single, highly robust phenomenon. In **Paper 4** we therefore conduct a two-tiered replication of the classic mere exposure effect.

1. **Direct replication on a 2D monitor.** The authors first reproduce the original study from Mrkva and Boven [80] by using the same stimuli, exposure schedule, and recruitment. This re-establishes the linear increase in liking with exposure.

2. **Replication in HMD-VR.** The identical materials and timings are then ported into a head-mounted-display environment built within our own virtual reality environment (VR Campus) using tools from the above-mentioned VRSTK while every non-display parameter (sample, instructions, questionnaire, analysis script) remains unchanged.

By chaining the original screen experiment, a preregistered direct replication, and a VR replication **Paper 4** aims to deliver evidence that immersive

VR can sustain a well-established psychological effect. It is important to state that the effect also was chosen because there are no technical challenges that might introduce unwanted artifacts such as the technology-specific confounds identified in Paper 3. The study thus completes the narrative arc that began with, the updated scientifc process using toolkits (Paper 1), assesing toolkit capabilities (Paper 2), comparison between IVR and the real world and between IVR and 2D Screens (Paper 3), showing the potential of conducting experiements within an IVR environment.

In addition to testing whether IVR is a viable medium for social science research, for **Paper 5** we introduce and elaborate on a novel procedure called *asymmetric normalization* that is unique to and for most parts, only feasible within immersive virtual environments. Asymmetric normalization standardizes the experimental scene to an agreed reference frame (e.g., the average adult eye-height or other factors on appearence or translation such as position or movement) while still letting each participant perceive the environment from an individually tailored vantage point. Consequently, Participant A and Participant B may not see precisely the same visual layout at any given moment. Instead, each view is dynamically rescaled, shifted, or occluded so that critical stimuli occupy equivalent egocentric across observers.

5 Reducing the Human Factor in Virtual Reality Research to Increase Reproducibility and Replicability

Publication Note: This chapter is based on the following published work. The content of the chapter is identical to the published article, with only the formatting and numbering being adapted for this dissertation.

Hepperle, D., Dienlin T. and Wölfel, M. (2021)
IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct), pp. 100-105
DOI: 10.1109/ISMAR-Adjunct54149.2021.00030

Abstract

The replication crisis is real, and awareness of its existence is growing across disciplines. We argue that research in human-computer interaction (HCI), and especially virtual reality (VR), is vulnerable to similar challenges due to many shared methodologies, theories, and incentive structures. For this reason, in this work, we transfer established solutions from other fields to address the lack of replicability and reproducibility in HCI and VR. We focus on reducing errors resulting from the so-called *human factor* and adapt established solutions to the specific needs of VR research. In addition, we present a toolkit to support the setup, execution, and evaluation of VR research. Some of the features aim to reduce human

errors and thus improve replicability and reproducibility. Finally, the identified chances are applied to a typical scientific process in VR.

Keywords Human-centered computing – HCI theory-concepts and models – HCI design and evaluation methods—user studies

5.1 Introduction

The necessity for increased qualitative rigor in research (especially in regards to reproducibility and replicability) has spread around manifold disciplines ranging from (social) psychology [1] over neuroscience [2] to communication [3] and the social sciences in general [4]. Even though the disciplines differ, they share a common set of methods (e.g., questionnaires), theories (e.g., technology acceptance model), and incentives (e.g., focus on publication quantity). Hence, we can assume that the problems originally causing the so-called *replication crisis*—such as a lack of robustness and trustworthiness [1]—are likely to exist in disciplines such as empirical sciences and human-computer interaction (HCI) as well. Another reason could be that researchers from empirical computer science disciplines are not sensitive to these issues. It seems that HCI and VR research lacks statistical power analyses and coherent approaches to increase reproducibility and replicability [5], [6].

Fortunately, however, disciplines such as computer sciences or HCI are per se qualified for systematic, structured procedures especially regarding approaches related to the open source community. For example, using version control software [7] such as GIT is a common practice in this area, and also for research purposes, it offers increased transparency and traceability. Hence, we argue that there are essential possibilities to increase research quality in terms of transparency and objectiveness in HCI by adapting established practices from neighboring disciplines, while also incorporating strengths typical of and unique to HCI. Many of the approaches that show high potential to reduce biases, heuristics, and other errors were already introduced by the open science community.

Building on top of that, also in empirical computer sciences we see much potential for dedicated software-based approaches such as frameworks and toolkits that offer a sophisticated pipeline for research setup, test, and evaluation, flexible enough to be customized while powerful enough to provide the most common functions and methods to reduce human-induced errors.

In the following, we discuss these possibilities through the introduction and use of toolkits particularly designed for conducting empirical VR research. A good overview of such toolkits can be found in [8] which also briefly describes the *Virtual Reality Scientific Toolkit* (VRSTK) introduced by the authors just recently. The main idea behind such toolkits is to support researchers to be more efficient by reducing the amount of redundant work that occurs on most VR research projects such as implementing questionnaires or data-logging. But, even more importantly, some of those toolkits are available open source which enables the community to constantly reevaluate and improve them, as well as to discuss improvements directly with the developers. When open source, most of the projects including their documentation are hosted in open repositories including a visible development history.

With this in mind, on the one hand, we see the possibility of using (predefined) computer instructions to create, modify, and capture all aspects of these virtual (research) worlds. On the other hand, the differences between the virtual and the real world give rise to various challenges—not only technical ones. Here we discuss the relationship between the traditional human factors that are crucial for reproducibility and replicability in research and the specific characteristics of VR that can support this. The main points discussed are:

Implementing Standards: Frameworks or toolkits in computer science are an established way of reducing redundant work by providing a code base that offers a proven set of methods. This is discussed in two directions. First, providing an open source code base so that dedicated scientists are able to get a full overview of the code and also to participate in the development process. Second, providing scientifically proven solutions to common problems, such as best practice examples on how to implement questionnaires in VR to reduce the break in presence [9] or how to prepare recorded data for further analysis in statistical software such as R or similar.

Process Documentation: One of the reasons why research fails to replicate is missing or unclear documentation [1]. Broken down to its very essence, VR is a set of data and code in which all treatises can be logged and processed for further analysis.

Optimizing Generalizability: In a traditional research environment, participants and investigators have predefined characteristics, such as size, ethnicity, or age, which cannot be altered. However, in VR, participants and investigators are represented by avatars that can freely be modified, liberating from given constraints and thus allowing for a setup that can be tailored to the individual user or experiment.

Sharing Test Environments: Since most of the parts in the VR research process are digital, the complete test environment can be saved and shared via online repositories such as github and gitlab[1]. "Barriers to effective data sharing and preservation are deeply rooted in

[1] An overview of how to use version control software in order to increase reproducibility is given here by "theturingway" project [10]

the practices and culture of the research process as well as the researchers themselves. New mandates for data management plans from the National Science Foundation (NSF) and other federal agencies and world-wide attention to the need to share and preserve data could lead to changes." [11]

Technical Limitations: Although VR research equipment is generally less expensive than traditional laboratory facilities, researchers still need to invest in appropriate hardware, software, and content. Hardware is also generally developed in rapid iterations. This must be taken into account in terms of reproducibility and replicability.

Taking these points into account, our goal in this paper is to provide interested researchers with an approach to improve their research process. Specifically, we aim to reduce errors resulting from human factors. We believe this can be achieved without much additional work, since the toolkits themselves have the goal of minimizing workload and the processes we present can be seen more as a by-product when using them.

5.2 Theoretical Foundation

"Everyone agrees that scientific studies should be reproducible and replicable. The problem is almost no one agrees upon what those terms mean." This vivid statement by Patil et al. [12] exemplifies the problem with the terminology between different research disciplines as well as within the same area of research. Widely accepted definitions of *replicablity* and *reproducibility* are given by Asendorpf [13] who have their origin in the field

of psychology and are also applied in communication [3] and empirical computer sciences:

Reproducibility "means that 'Researcher B' [. . .] obtains exactly the same results (e.g. statistics and parameter estimates) that were originally reported by 'Researcher A' [. . .] from A's data when following the same methodology."

Replicability "means that the finding can be obtained with other random samples drawn from a multidimensional space that captures the most important facets of the research design."

Another critical aspect in regards to the *human error* is *generalizablity*, which is defined as a finding that "[. . .] does not depend on an originally unmeasured variable that has a systematic effect. [. . .] *Generalizability* requires replicability but extends the conditions to which the effect applies." [3], [13]

With these definitions, we see two main points emerging which are prone to *human error*. First, in case of *reproducibility*, to achieve the exact same results as 'Researcher A', it is necessary for 'Researcher B' to:

- be able to gain the datasets used by 'Researcher A', completely prepared and well documented for further use.
- be able to use and work with data provided by 'Researcher A'. This requires a comprehensive report about methodology as well as a clear understanding of the research question.

While there already exist several approaches on how to provide data beside the basic journal submission such as provided by the open science

foundation[2], it is still time consuming. Especially in regards to the *publish or perish* concept, it means additional work that not everyone is given enough time for providing their (clean) datasets.

For *replicability*, comprehensibility of the experimental procedure is of upmost importance. If possible, the best way to achieve this is to be present when the original research was carried out. However, in many cases this is not possible and sometimes it's helpful if the authors do not know each other. As a result, we believe it is necessary to create documentation that goes beyond the standard scope of a journal. This extended way of documenting the work should enable researchers to understand the different, sometimes vague, conclusions made in order to formulate hypothesis and reporting results.

Generalizability requires similar standards such as *replicability*. However, it adds even more need for documentation since it is rooted in the conflict between verbal expressions or qualitative theoretical claims and quantitative measure of it (i.e., construct validity [14]). Yarkoni [15] for example emphasizes the fact that authors should "disclose [. . .] non-trivial effects (e.g. effects on stimuli, experimenter, research site, culture etc.) to the best of authors' abilities."

5.3 The Human Error

A sole observation without any form of transcription or recording will lead to information loss. Even with transcription, factors such as social cues, tonality, and non-verbal information get lost. This is not a new fact

[2] `https://osf.io/`

but has been observed already in the 80s, where almost 90% of social science investigations were done using qualitative interviews [16]. Since then (if not earlier), researchers suggest ubiquitous recording (both video and audio) for better comprehension[17]. " 'human error' is not just about humans, it is about how features of people's tools and tasks and working environment systematically influence human performance." [18] This can be interpreted in two directions. On the one hand, complicated system design increase errors. On the other hand, tools help reduce human-induced errors. Therefore, an interface easy to understand is necessary to reach as many people as possible, while also reducing the above mentioned kind of errors and challenges.

Even though there is a shift from qualitative approaches to quantitative research methods such as questionnaires—aiming to reduce subjective interpretation by making it quantifiable for example via Likert-scale ratings—these approaches are still prone to human error, for example with regard to inconsistencies in test-statistics [19] and refusal of data sharing [11]. This is especially concerning the high error rates and overall low transparency in the field of VR [5]. In addition, only two out of 61 works provided supplementary material. Also, the rather low number of participants in the respective studies (median = 25) implies that studies tend to be underpowered. To illustrate, chances of detecting a small to medium ($d = 0.2; d = 0.3$) effect with the given median participant number is between 17% to 28% [5]. Inter-rater reliability (IRR)—which basically is a quantization of the mismatch between the different coders—often is not fully reported and the analysis of such is misinterpreted so that other researchers cannot address how the IRR affects the conducted evaluation [20].

As pointed out earlier, several of these issues based on human error can be overcome by using structured approaches such as those offered by computer programs. The recent development of various toolkits to support the research process, in general, have the potential to counteract these sources of error in various ways. In the following sections, we will discuss these possibilities based on typical research steps in VR research.

5.4 Toolkits for Virtual Reality Research

Recently, a number of toolkits were introduced that aim to support research in VR [8]. Most of those toolkits share the goal to reduce redundancies by supporting recurring research tasks such as data logging and data export. While this is also one of the goals for the VRSTK, there exist several other functions that help to increase the research process in general and thus are introduced here briefly.

The VRSTK originated as a by-product of our ongoing research in the Intelligent Interaction & Immersive Experience Lab at the Karlsruhe University of Applied Science by reducing redundancies in the context of data logging and data export. Over time, while learning more about the specific needs and chances of VR research, the set of functions implemented increased, and thus, we decided to take the project to another level by rethinking the whole process of (easy) implementation, standardization, and collaboration.

As of now, the toolkit comes with several predefined scenes that provide easy access to technologies such as

- basic interaction forms,
- questionnaire implementation,
- eye & gaze tracking, or
- pose detection.

Everything happening within the virtual test environment can be recorded in real-time for a later replay, augmented analysis, and export to dedicated statistic software such as R. The toolkit is implemented within the game engine Unity[3]. It is provided as open-source software on github[4].

Currently, next to a multiplayer setup for remote testing, the capturing of brain activities via EEG is implemented. Another important aspect of the VRSTK is the investigation and representation of humans which seems to be particularly challenging in VR in contrast to other media; see e.g., the uncanny valley effect [21]. For a full feature overview of the VRSTK see [8].

5.5 Current Computational and Statistical Approaches to Reduce the Human Factor

"I would hazard a guess that if we examined the combined effects of the numerous sources of bias that operate in various sociological and psychological studies we would find that they account for considerably

[3] https://unity3d.com/

[4] https://github.com/ixperience-lab/VRSTK/

more of the variance in the dependent variables of interest than do the major independent variables." [22].

As stated before, also VR research lacks approaches to ensure replicability and reproducibility, which lowers trust in and also the applicability of those results [23]. However, in writing this manuscript for a dedicated workshop on replicability in extended reality, even though it is the first of its kind, we see an increased awareness, similar to what has happened in other disciplines before. Therefore, in what follows we look at valuable practices introduced in other disciplines that help to increase overall research quality such as the open science approach. We then adapt and extend them to the specific needs of VR research.

Dienlin et al. [3] suggested a 7-step open science agenda in the field of communication of which most points can be applied to many other research disciplines as well. The seven steps are as follows:

1. publish materials, data, and code
2. preregister studies and submit registered reports
3. conduct replications
4. collaborate
5. foster open science skills
6. implement transparency and openness promotion guidelines
7. incentivize open science practices

Below, we discuss how we address the points 1, 3 and 4 with our toolkit.

Next to proposals on how to create more trust and reliability in research work regarding open science, there are calls to improve the way how results are reported. For example, Dragicevic [24] promotes the use of confidence intervals instead of p-values for the field of computer science, aiming to provide a better interpretation of the results.

In addition to these specific approaches, web-based tools such as "parsif.al"[5] or "asreview.nl"[6] emerge to support the standardization approach for systematic review. The "statcheck" web interface[7] offers automated support in detecting errors in statistical reporting by simply uploading the PDF file.

In what follows we apply those points to VR research, adding points where we see potential for such toolkits to offer a valuable contribution.

[5] https://parsif.al

[6] https://asreview.nl/

[7] http://statcheck.io/

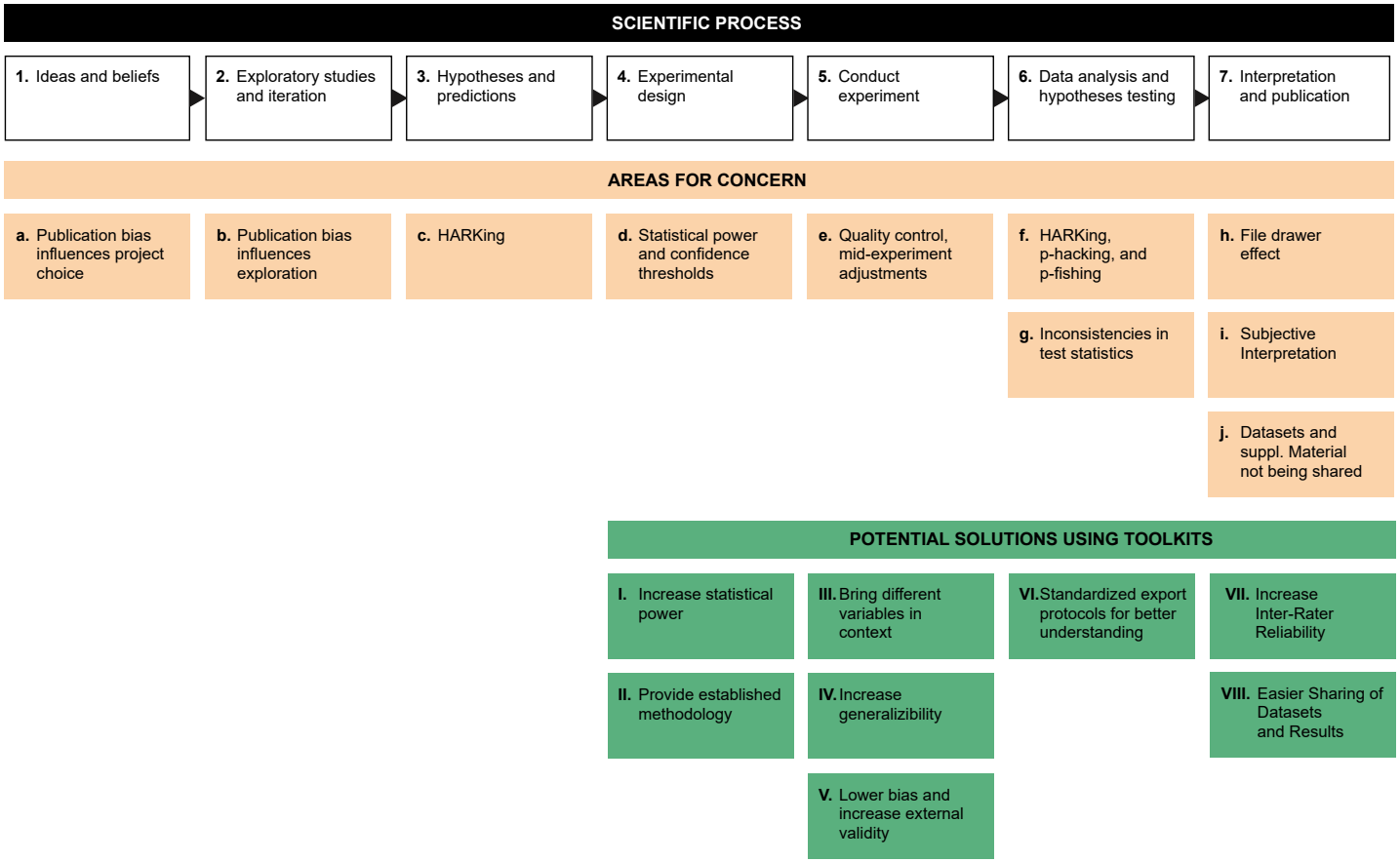

Figure 5.1: Stages of a typical experimental process (top) including typical areas of concern (mid) and potential solutions using toolkits (bottom). Originally introduced by [25] and adapted from [6].

5.6 Steps to Improve Virtual Reality Study Procedures

Most of the individual steps within a VR research process are similar to those from related disciplines, although some differ slightly. Looking at Fig. 5.1, which shows a common research process, we identify *step 4, 5, 6, and 7* for which mentioned toolkits are especially helpful. Given that in quantitative research objectivity is the upper limit for reliability, one important goal is to eliminate subjective influence wherever possible.

In the following, we discuss how mentioned problems concerning objectivity can be overcome or lowered by functions as offered by the use of toolkits. We will discuss the options based on 4 steps about the features currently provided—or planned to be implemented—in the VRSTK while they might also apply to other toolkits.

Step 4 – Experimental Design in which statistical power is a concern can be addressed by the VRSTK such that it facilitates sharing the whole research project with other researchers to increase the number as well as diversity of participants. Also, the option to apply a (remote) multiplayer setup, so that participants can join from either their living room with their technical setup or from a lab nearby, can increase participation and thereby statistical power (see Point I. in Fig. 5.1). Since it is always a critical factor for reproducible research to have enough participants, easy access to participation can increase participant number and therefore power.

Another helpful aspect of formal toolkits is that they provide established methods critical for VR research (see Point II. in Fig. 5.1). This includes

in-VR-questionnaires to keep up presence as well as other approaches to cope with fatigue, motion sickness, and other unwanted effects. For example, the ETRI (Korea's Electronics and Telecommunications Research Institute) has successfully trained an artificial intelligence that helps identify potential issues regarding motion sickness in VR applications[8].

Notably, within the VR environment, everything can be shared across the world. However, this excludes aspects from reality in which the experiment takes place. Here aspects such as the general onboarding process (e.g. how to familiarize people with a headset), cultural background (e.g. preferences of the color temperature of the light), presumptions about the experiment, and other aspects such as outside temperature might lead to a potential bias—even though the virtual setup is identical. Hence, there still is a need for clear documentation besides sharing the sole codebase.

Step 5 – Conduct Experiment is where the VRSTK offers the greatest benefit. With the implementation of pre-defined procedures, (technical) generalizability can be increased. This is achieved by the following aspects of the VRSTK:

Sensing Participants A major feature of VR research is its ability to comprehend every single variable within the VR environment, creating a unique common context (see Point III. in Fig. 5.1). For example, it can be recorded in real-time where a participant looked at and for how long in a very exact manner using simple ray-casting. In

[8] `https://disruptivetechasean.com/big_news/technology-to-reduce-motion-sickness-in-vr-applications-developed-using-ai/`

addition, this can be annotated for example with simultaneous transcription of what currently is said in the experiment. Even small differences in the implementation of the data processing of different sensor types can lead to a bias in results. With a predefined implementation in combination with comprehensive documentation, we hope for increased standardization and higher generalizability. Using a different microphone for example can lead to a different frequency pattern and therefore the algorithms' interpretation of the speakers' mood might differ (see Point IV. in Fig. 5.1).

Representation in VR One of the most critical aspects of VR research is the representation of either the participant or the operator, which greatly influences the participant's perception of all interpersonal interactions within a virtual environment. Representation can be divided into several parts. On the one hand, we have the outer appearance of the avatar which is defined by variables such as gender, height, ethnicity but also clothing. Also depending on character style, the perception of those characters might be prone to the uncanny valley effect, which tends to be even more pronounced within VR environments than on 2D monitors for example [8]. Without technical support by systems—such as Microsoft's Azure Kinect or HTC Vive Trackers combined with inverse kinematics—the transmission of the real participants' movement of extremities in VR is not given. Therefore, the VRSTK offers a predefined scene that implements these transmission possibilities. Since these are rather complicated procedures, a common research or implementation standard would be helpful to foster generalizability, and possibly

increase external validity. Notably, one problem we see in this regard is varying technical setups. If one person for example lacks one of the technologies mentioned above, tracking is not possible and therefore the participant can either not take part or will create a large bias. Another aspect of representation within VR is the ability to alter the look of the person in real-time to the respective needs and to lower bias based on different ethnicity, height, age, or gender for example (see Point V. in Fig. 5.1). This can be taken even further in multi-user applications in which one participant can see other participants in an average height for the respective ethnicity while their perspective fits their real height and vice versa to foster external validity. Also, movement can be adjusted in the same way to gain additional information on for example proxemics. Here, in each version of the virtual setup in the multi-user application, the movement (translation in the horizontal direction) of one participant is not transmitted to the second participant and therefore does not influence the personal proxemics' preference.

Step 6 – Data Analysis and Hypotheses Testing is supported by the VRSTK in parts. An export of pre-selected data can be prepared for post-processing in statistical software such as R. Also custom scripts are offered for R to plot common data such as eye-tracking heat maps.

Since research methods are manifold and it often differs from what the respective researcher aspires to implement based on their different needs, we did not implement further scripts about specific data analysis. The provided standardized export protocol though is essential for researchers to better comprehend the data.

Likewise, other researchers can more easily replicate or reproduce the findings (see Point VI. in Fig. 5.1). Features such as API support to common questionnaire platforms such as soscisurvey[9] are planned but not implemented yet.

Step 7 – Interpretation and publication Interpretation of results usually is prone to subjective influence. A way to overcome this issue is to work in groups of several persons to check if everyone would come to the same conclusion and measure the IRR for variances within the ratings (see Point VII. in Fig. 5.1). Here, features such as the annotated replay function come in handy because each researcher can individually take a look at the complete research process, allowing computer-measured support in form of annotation such as current heart rate.

Furthermore, the provision of supplementary material and data in the course of the publication means effort and expense which leads to the fact that this often does not happen. This timely effort can be reduced by toolkits due to the standardized exports and general sharing options (see Point VIII. in Fig. 5.1).

5.7 Conclusion and Outlook

In this paper, we showed that many of the findings identified and formulated in the context of the replication crisis likely apply also to research in VR. On the one hand, we see several indicators that show research in

[9] https://soscisurvey.de

the HCI is similarly prone to questionable research practices that eventually lead to the lack of reproducibility, replicability, and trust in (this) research. On the other hand, venues such as the "workshop on replication in extended reality" and other approaches address, discuss and—most importantly—spread the word about the issues and raise overall awareness.

Due to the basic systematics underlying computer programs, and therefore also VR systems, we argue that the field of HCI is particularly suitable for integrating parts of the processes in a typical experimental setup. Furthermore, we apply these findings to existing research toolkits, which were originally developed to help non-specialist or untrained users set up VR experiments and at the same time save redundant work. Our discussion showed that there is potential to further systematize the scientific discourse with given technology (in the sense of programs), to make the process and the insights more transparent and thus more open. Although the propagated precautions and the associated technology are still in their infancy, we recognize the potential to increase trust in research results in the field of VR. As part of the discussion, the VRSTK is presented. The VRSTK is made available as open-source and other scholars are explicitly invited to contribute by suggesting improvements or adding features.

Since some points are predestined to increase reproducibility and replicability, in the foreseeable future we recommend developing a roadmap that takes into account additional features necessary to further implement processes and solutions as discussed in the course of the open science initiative.

As of today, the interface of the tool is not yet sufficiently developed to guarantee error-free use for untrained users which, as discusses above can also lead to human-induced error in a typical research process. Also, it has not been conclusively clarified to what extent results can also be projected onto the real world. Approaches such as open hardware (e.g., the connection and design of open-source hardware components), which can be used to capture sensations from the real world and feed them to the virtual test setup could help improve this process.

Here, too, we see a need for standardization, since the sensors used are noisy and error-prone. The VRSTK currently offers a serial interface for connecting such sensors, but this is no more sophisticated than reading out a simple data stream and documenting it.

Raising awareness in our opinion is a crucial first step for others to become part of projects such as “theturingway” or “open science foundation” to eventually create a critical mass to carry on the development of software tools such as the VRSTK or similar. While several of the open science suggestions imply a substantial rethinking of how research is conducted, the proposed toolkits are easy to implement and offer also benefits in case of time-reduction for creating research procedures in VR. In conclusion, next to the timely benefits mentioned above, implementing dedicated toolkits into the research process helps make research more robust, and credible. Since there has been increased attention in related disciplines to such challenges, there exist various solutions to counteract the problem.

Notably, data-sharing comes with concerns especially concerning privacy standards of the participants as well as of the cooperating funding partners from industry and other copyright reasons. Fortunately, there are

already standards that might apply to these issue such as the DSGVO introduced by the German government in the 'Bundesdatenschutzgesetz'[10]. The principles of 'Datensparsamkeit'[11] mentioned in this context can also be applied to an open research process. Fundamental to this is the idea that only data which actually will be used later is saved and that the purpose of use must be presented in a simple and comprehensible manner. Thomson et al. [26] for example propose the use of proxy data to make data available for secondary use.

In conclusion, we hope the VRSTK helps readers increase their research quality by implementing standardized procedures that decrease errors resulting from human factors, which together should help reproduce, replicate and after all improve research in VR. On top of that, ultimately, the question arises whether—or in how far—VR can be a valuable tool to conduct research that originally would have been conducted in the real in order to reduce human error and to improve replicability and reproducibility.

[10] Translates to Federal Data Protection Act

[11] Translates to data minimization

References

[1] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, 2015. DOI: 10.1126/science.aac4716.

[2] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, *et al.*, "Power failure: Why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, vol. 14, no. 5, pp. 365–376, Apr. 2013. DOI: 10.1038/nrn3475.

[3] T. Dienlin, N. Johannes, N. D. Bowman, *et al.*, "An agenda for open science in communication," *Journal of Communication*, vol. 71, no. 1, pp. 1–26, Feb. 2020.

[4] C. F. Camerer, A. Dreber, F. Holzmeister, *et al.*, "Evaluating the replicability of social science experiments in nature and science between 2010 and 2015," *Nature Human Behaviour*, vol. 2, no. 9, pp. 637–644, Aug. 2018.

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

[6] A. Cockburn, P. Dragicevic, L. Besançon, and C. Gutwin, "Threats of a replication crisis in empirical computer science," *Communications of the ACM*, vol. 63, no. 8, pp. 70–79, Jul. 2020. DOI: 10.1145/3360311.

[7] C. Rodríguez-Bustos and J. Aponte, "How distributed version control systems impact open source software projects," in *2012 9th IEEE*

Working Conference on Mining Software Repositories (MSR), IEEE, 2012, pp. 36–39.

[8] M. Wölfel, D. Hepperle, C. Purps, J. Deuchler, and W. Hettmann, "Entering a new dimension in virtual reality research: An overview of existing toolkits, their features, and challenges," in *Proceedings of Cyberworlds*, 2021.

[9] S. Putze, D. Alexandrovsky, F. Putze, S. Höffner, J. D. Smeddinck, and R. Malaka, "Breaking the experience: Effects of questionnaires in VR user studies," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ACM, Apr. 2020. DOI: 10.1145/3313831.3376144.

[10] T. T. W. Community, B. Arnold, L. Bowler, *et al.*, *The turing way: A handbook for reproducible data science*, 2019. DOI: 10.5281/ZENODO.3233853.

[11] C. Tenopir, S. Allard, K. Douglass, *et al.*, "Data sharing by scientists: Practices and perceptions," *PLoS ONE*, vol. 6, no. 6, C. Neylon, Ed., e21101, Jun. 2011. DOI: 10.1371/journal.pone.0021101.

[12] P. Patil, R. D. Peng, and J. T. Leek, "A statistical definition for reproducibility and replicability," *BioRxiv*, 2016.

[13] J. B. Asendorpf, M. Conner, F. D. Fruyt, *et al.*, "Recommendations for increasing replicability in psychology," *European Journal of Personality*, vol. 27, no. 2, pp. 108–119, Mar. 2013. DOI: 10.1002/per.1919.

[14] L. J. Cronbach and P. E. Meehl, "Construct validity in psychological tests.," *Psychological Bulletin*, vol. 52, no. 4, pp. 281–302, Jul. 1955. DOI: 10.1037/h0040957.

[15] T. Yarkoni, "The generalizability crisis," *Behavioral and Brain Sciences*, pp. 1–37, Dec. 2020. DOI: 10.1017/s0140525x20001685.

[16] M. Brenner, "Problems in collecting social data: A review for the information researcher," *Social Science Information Studies*, vol. 1, no. 3, pp. 139–151, 1981. DOI: https://doi.org/10.1016/0143-6236(81)90029-6.

[17] C. Briggs, *Learning how to ask: a sociolinguistic appraisal of the role of the interview in social science research*. Cambridge Cambridgeshire New York: Cambridge University Press, 1986.

[18] S. Dekker, *The field guide to understanding 'human error'*. CRC press, 2014.

[19] M. B. Nuijten, C. H. J. Hartgerink, M. A. L. M. van Assen, S. Epskamp, and J. M. Wicherts, "The prevalence of statistical reporting errors in psychology (1985–2013)," *Behavior Research Methods*, vol. 48, no. 4, pp. 1205–1226, Oct. 2015. DOI: 10.3758/s13428-015-0664-2.

[20] K. A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Tutorials in Quantitative Methods for Psychology*, vol. 8, no. 1, pp. 23–34, Feb. 2012. DOI: 10.20982/tqmp.08.1.p023.

[21] D. Hepperle, C. F. Purps, J. Deuchler, and M. Wölfel, "Aspects of visual avatar appearance: Self-representation, display type, and uncanny valley," *The Visual Computer*, Jun. 2021. DOI: 10.1007/s00371-021-02151-0.

[22] B. Phillips and J. C. Gazet, *Abandoning method*, 1973.

[23] T. Wingen, J. B. Berkessel, and B. Englich, "No replication, no trust? how low replicability influences trust in psychology," *Social Psychological and Personality Science*, vol. 11, no. 4, pp. 454–463, Oct. 2019. DOI: 10.1177/1948550619877412.

[24] P. Dragicevic, "Fair statistical communication in hci," in *Modern Statistical Methods for HCI*, J. Robertson and M. Kaptein, Eds. Cham: Springer International Publishing, 2016, pp. 291–330. DOI: 10.1007/978-3-319-26633-6_13.

[25] O. E. Gundersen and S. Kjensmo, "State of the art: Reproducibility in artificial intelligence," in *AAAI*, 2018.

[26] D. Thomson, L. Bzdel, K. Golden-Biddle, T. Reay, and C. A. Estabrooks, "Central questions of anonymization: A case study of secondary use of qualitative data," *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, vol. 6, no. 1, Jan. 2005. DOI: 10.17169/fqs-6.1.511.

6 Entering a new Dimension in Virtual Reality Research: An Overview of Existing Toolkits, their Features and Challenges

Publication Note: This chapter is based on the following published work. The content of the chapter is identical to the published article, with only the formatting and numbering being adapted for this dissertation.

Wölfel, M., Hepperle, D. Purps, C. F., Deuchler J., and Hettmann, W. (2021)
2021 International Conference on Cyberworlds (CW), pp. 180-187,
DOI: 10.1109/CW52790.2021.00038

Abstract

Virtual reality becomes a medium to be explored for itself, to study human factors and human behavior within these worlds, and to infer possible behavior in the real world. Among many advantages, building test routines in virtual environments remains a challenge due to the lack of established procedures and toolkits. To encourage research in this direction and lower the barrier to entry, it is necessary to simplify the process of setting up a research environment in virtual reality by providing appropriate toolkits. This paper discusses what challenges need to be overcome, what features might be relevant, and compares available toolkits.

Keywords Research Tools – Virtual Reality – Human Factors – Human Behavior

6.1 Introduction

Putting on a *virtual reality* (VR) headset the outside world disappears and lets the user dives into an alternative reality. Meanwhile, this technology is not only widely adopted by gamers, treatments (in particular phobias) and for virtual trainings but is getting adopted by researchers in the field of human factors, human behavior, and the like. VR is interesting for them in, at least, two ways:

- As *a medium to be investigated for itself,* that has its own rules to be understood and compared with other media. This offers new potential for designers and developers of VR content to optimize their applications' experience, narration, or content.

- As *a medium that provides a tool to overcome the need for real world experimentation*, as it provides some advantages over real-world experiments including, lower expenses, higher repeatability, or reducing the risk in potentially dangerous situations. The hope here is, that the findings of human behavior in VR can be applied to real-world situations.

However, virtual environments and testing routines are yet difficult to establish as most of the times it has to be re-invented or re-implemented for each experiment even though similar data might be measured. To do so, most researchers creating their test environment use one of the two

engines Unity by Unity Technologies or Unreal by Epic Games. However, these engines do not provide tools to investigate human behavior as their preliminary goal is to support interactive 3d content development. To lower the entrance barrier and to foster the use of VR in the field of human factors and human behavior requires features such as the recording, replay, and analysis of sensor data aligned with the virtual environment to be integrated within the development platform.

To overcome this drawback and provide tools to facilitate the entry into human factors and human behavior research using VR, we see four main challenges that need to be addressed:

- To *provide a general framework* that offers a sophisticated pipeline between setup, test and evaluation in a way flexible enough to be customized but powerful enough to provide most common functions.

- To *make implementation easy*, so that non experts can make use of basic functionality and set up own experiments without the need to consult experts.

- To *implement common research methods*. For example, one opportunity in using a VR scientific toolkit, is to make results more *generalizable* by establishing standards in tracking, visualization as well as in recording and providing specialized test-setups with evaluated test-scenes.

- To *share the complete environment* with other researchers so that they have the possibility to repeat the experiments.

Manufacturers of real time 3D render engines, such as Unity include native toolkits for analytics[1]. However, these tools are primarily focused on the technological development of interactive 3D applications rather than scientific applications, have constraints considering data quality, customization capabilities and lack open-source code. Therefore, to support research in human factors and human behavior within virtual worlds additional solutions are needed.

Until recently, there were not many software or toolkits available to support research in this direction. Therefore, in 2017, we started the development of our toolkit dubbed *Virtual Reality Scientific Toolkit* (VRSTK). Meanwhile, additional toolkits have been introduced to the research landscape offering various capabilities to support research in this direction. Since most of these toolkits have been developed within or for different research disciplines, their focus is on different functions and there is no single toolkit that provides all the necessary functions for investigating immersive virtual worlds accessed via a *head-mounted display* (HMD).

6.1.1 Introduction of Existing Toolkits

While many toolkits have 'VR' and 'toolkit' in their name, the focus of many is not on research within our scope but aims to integrate VR technology and interactive 3D visualization into technical and scientific applications. As laid out in the introduction we are interested in toolkits offering to support research on human factors. Therefore, we will only present and discuss toolkits in this narrower focus next. A detailed comparisment of

[1] https://www.https://unity.com/de/features/analytics/

the different toolkits and frameworks, including our VRSTK, is presented in Section 6.6.

The **Unity Experiment Framework** (UXF)[2] is intended to study human behavior in VR [1]. The open-source framework offers a unified and simplified procedure to set up experiments and makes produced measures of subsequent data analysis straightforward. The UXF however does not implement solutions for specific data acquisition and processing (motion tracking, object usage, bio-signals, etc.).

VREX[3] focuses on experimental psychology and neuroscience [2]. Although VREX provides good generic help in setting up VR experiments, including various study protocols about perception, attention, cognition, and memory, it lacks the same, more specific features as UXF (e.g. sensing interfaces).

The **Toggle Toolkit** [3] allows allocation of triggers (e.g. collision, distance, key, etc.) and toggles (e.g. user teleportation or light-, object or audio manipulation) to VR objects and logging of the generated data for later analysis. The solution remains however a generic one that does not implement specific features such as body- or face tracking interfaces or operator embodiment.

The **Experiments in Virtual Environments** (EVE) provides inherent interaction, logging, and evaluation scripts, and comes with utilities for physiological measurements, questionnaires, and movement tracking [4].

---

[2] http://immersivecognition.com/unity-experiment-framework/

[3] http://vrex.mozello.com/

However, EVE has limited potential to conduct social experiments, lacking features such as tracking facial expressions, posture, or audio signals, and an implementation of a social environment.

Cognitive3D's **Scene Explorer**[4] comes with a large feature set for customers and academic research purposes in mixed reality. Their system is capable to obtain data through versatile tracking capabilities, such as eye-tracking, user position, EEG, galvanic skin response, heart rate, user interactions, and to perform analysis. Nevertheless, the system is missing out on opportunities for more detailed participant analysis that can be derived from more specific data, such as full-body and facial features.

VIZARD[5] by WorldWiz offers unparalleled hardware connectivity and options for data collection and analysis. It supports motion and body tracking devices, such as VIVE Tracking, Vicon or OptiTrack, and eye-tracking hardware, such as Tobii or VIVE Pro Eye. Moreover DataGloves and haptics systems, such as Cyberglove or Manus VR Gloves. However, hardware for the detection of facial expressions or EEG-based brain-computer interfaces is not supported.

CyberSession[6] by VTPlus is a program to control empirical data collection with VR, to be applied in the field of experimental psychological, therapeutic, and neurophysiological research (e.g. [5], [6]). But it also misses opportunities for facial expression recognition or shared virtual encounters.

---

[4] https://cognitive3d.com/product/scene-explorer/

[5] https://www.worldviz.com/vizard-virtual-reality-software

[6] https://www.cybersession.info/

The **Virtual Reality Scientific Toolkit**[7] (VRSTK) is our development. While various solutions available are focusing on specific problems we aim for a holistic approach. And frankly, with our development, we are not there yet either. Since the beginning of the development, more and more functions have been added. We are continuously expanding and improving the offer. By sharing our considerations and source code, we hope to foster and encourage development in this direction.

6.1.2 Overview of Functionalities

Fig. 6.1 outlines an overview of different functionalities that VR scientific toolkits can offer. These include *Setup and Control* as well as *Sensing Participants*, *Representation*, and *Data Handling*. For instance, it can be seen that sensor input is required to represent both the avatar of the participant as well as the operant and can also be used for further analysis. The arrows are symbolizing the data flow.

The following introduction of functionality in sections 6.2 to 6.5 reflects either VR test-specific aspects or points not yet widely addressed in current VR research.

6.2 Setup & Control

Elementary user study toolkit features should provide support for planning and setting up the study design (e.g., briefing options, division into sub-experiments, sessions, trials) and allow for customization options to

[7] The latest version of the VRSTK is available for download via GitHub as a Unity package and in source code from https://github.com/ixperience-lab/VRSTK.

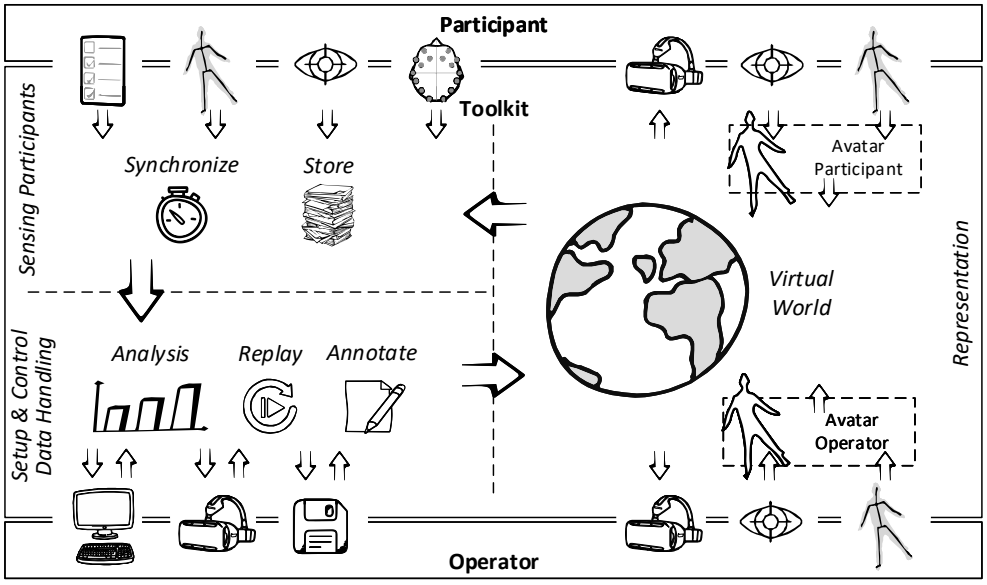

Figure 6.1: Overview of possible funcionalities VR scientific toolkits can provide.

assist researchers in set up test environments, lower the barrier to entry, and save time and resources. The ability to create and customize a *sample scene* and intervene in an ongoing experiment gives the operator a great deal of control over the environment, events, and actions in the scene. By providing a consistent strategy for running experiments, using a toolkit can improve reliability, validity, and accuracy, and lead to more precise results and reduced opportunities for error.

6.2.1 Remote Testing

The ability to run VR tests outside the laboratory environment provides new opportunities to diversify the test population, reduce spending on local test setups, and run studies during lockdowns. This process can be divided into asynchronous studies, where users download and run the

studies themselves[8] and synchronous studies, where one or more participants join a virtual test setup that is under live control of the experimenter. Challenges for remote VR testing include heterogeneity of hardware setups (HMD, controller, graphics card), physical space, no possibility for specific sensor setups, and potential environmental distractions that are difficult to account for and that could affect the validity of data collected and conclusions drawn from it.

An example of using existing social VR platforms such as VRChat was shown in [7]. They found that this approach is fruitful for evaluating prototype interactions and collaborative experiments, but has platform-specific limitations in terms of implementation and data retrieval.

6.2.2 Social Environment

Some areas of research, such as social, communication, or psychological studies, require the study of exchange and interaction between individuals or group of people. In this type of research, the use of confederates, trained actors that behave in a consistent manner, is common. In VR, confederates can be partially replaced by intelligent agents, which might provide a more homogeneous behavior and that might be advantageous in terms of reproducibility [8]. However, when there is a need for natural, controlled interaction that cannot (yet) be adequately replaced by an agent, or when the focus of the experiment lies on simultaneous observation of multiple participants, as may be the case in cooperative or group communication experiments, there is a need to acquire and connect participants from local or remote locations in a shared virtual environment.

[8] as supported by the XR Distributed Research Network https://www.xrdrn.org

6.3 Sensing Participants

VR hardware comes with a rich variety of sensors that are needed to create a responsive environment. While this data can be misused to collect personal and sensitive data about the user without their knowledge, it also provides a rich source of information on a legal basis if consent is given by the user. *Eye-tracking*, for instance, can be used in *foveated rendering* to reduce the rendering workload by reducing the image quality in the peripheral vision and to provide a gaze vector to determine the region or object of interest by the user. *Motion capture* increases the possibilities of interaction in virtual space. While a minimal VR setup with HMDs and hand-held controllers provide limited information for avatar rendering and for examining non-verbal behavior, systems that provide full-body motion capture (either through visual analysis with or without markers or through sensor suits and data gloves) help to increase the perceived sense of immersion, embodiment and identity of the avatar [9]. Depending on the technology used for motion capture, it a rich set of information such as emotions, intentions, relationships, or social status can be derived [10].

Besides the given sensors, additional insights can be gained by including sensors to measure *brain activity* [11], *blood pressure* [12], *heart activity* [13], or *stress levels* [14]. Although this is an established procedure in user studies, it has only recently been applied to VR experiments. This may be caused by the introduction of interference from the VR equipment into the sensitive pickup of such signals.

Audio and video recordings extend the complete coverage of the environment. Although it may sound trivial, this feature often seems to be

neglected by many VR research toolkits. Using additional tools, not synchronized with the toolkit, to capture audio and video creates additional potential for data loss and complicates time alignment.

While this setup already enables a variety of human-centered experiments (e.g., in psychology, ergonomics, social science, or medicine [15]–[17]), much is still unknown for VR environments in terms of how to combine sensor data with standard observational and questionnaire approaches. For example, facial features cannot be directly observed, and post-VR-experience questionnaires might yield different results than in-VR questionnaires.

6.3.1 In-VR Questionnaires

Questionnaires are an indispensable tool for most VR studies to measure participants' responses that cannot be captured well by sensory information. One example that is particularly relevant to conducting experiments in VR is the measurement of presence. The phenomenon of *presence* is used by researchers in various disciplines when conducting experiments in VR, and although much is still unknown about its concept and applicability, it appears to be a necessary but not yet sufficient condition for successful VR experiments [18], [19].

Typically, post-VR-experience surveys are conducted analogously on paper or digitally via a screen. However, research suggests that taking participants out of the virtual environment to answer questionnaires could cause systematic bias in participant responses due to the collapsing presence, in addition to spatial disorientation [20], [21]. Additional inaccuracies may be caused by insufficient reliance on memory or post-event

recollection [22]. Thus, the use of in-VR surveys should be advocated and compared to post-VR surveys to get a feel for potential differences in the findings.

Several in-VR questionnaires have been studied by Alexandrovsky et al. who found comparable completion times and higher enjoyment factor compared to non-VR questionnaires, albeit lower ease of use and higher physical demands within a tolerable range [23].

6.3.2 Facial Features

Facial expressions are considered the most important non-verbal information available for studying human factors and are particularly important as they provide key signals for estimating the emotional and mental states of others [10]. Unlike settings that do not require HMDs to be worn, here at least part of the face is obscured by the headset and thus cannot be observed directly. In addition, facial expressions can be altered by the pressure and constraint that the headset exerts on the head and face, resulting in morphological displacements. This effect must be compensated before being analyzed or interpreted either by algorithms or humans, to not lead to incorrect interpretations and conclusions. Even though, in our opinion considered a critical issue, there exist only limited considerations in this direction; e.g. [24].

Human interpretation of synthesized faces may introduce an additional source of error or uncertainty that, to our knowledge, is not yet fully understood. Other influences, such as the *Kuleshov effect*, which states that the context in which a face is shown has a significant impact on how the face is perceived, could also lead to alternated results [25].

6.4 Representation

How the environment and the entities it contains (e.g., surroundings, objects, participants, agents) are perceived in VR is fundamentally dependent on their visual representation and the possibility of interacting with it.

Although maximum realism is not always the goal, and more abstract representations also have their applications, modern VR applications aim to provide an appealing appearance and create a realistic environment. Realism includes, for example, the visual fidelity of the environment, the behavioral fidelity of an avatar, and the interaction capabilities of a participant. Besides showing virtual objects, *physical objects* can be rendered (as proxies) in a virtual scene.

6.4.1 Embodiment of the Participant

The representation of one's own body has a significant influence on the state of the subjective experience of using and having a body and includes the sense of embodiment, self-location, sense of agency, and sense of body ownership [26], [27]. According to the *Proteus effect* [28], the behavior of an individual, within VR, is modified by the assumed properties of its visual avatar representation. Therefore, any difference in the visual representation of the participant—be it gender, race, musculature, stature, clothing, hair, body jewelry, tattoos—as opposed to the real appearance can have a significant impact on the results. Similar influences can be assumed if a representation is missing or incomplete. While manipulating such parameters on purpose allows for a wide range of human-centered experiments

otherwise hard to impossible to perform (e.g. to foster empathy [29]) [30], particular care has to be taken to not alter the findings incautiously. This could impact not only multi-user settings when it comes to communication and collaboration between multiple embodied avatars but also for communication with computer-controlled embodied agents and even in single-user settings.

Besides the resemblance of visual aspects, the truthful representation of non-verbal communicational signals, such as body posture, gestures, facial expressions is another aspect that needs to be investigated. Considering that real feelings and emotions can be aroused in VR like in the real world [10] it is important to transfer them onto the representation.

6.4.2 Embodiment of the Operator

While in many real environments an operator is present to provide guidance, the feeling of safety, etc., in VR environments—generally—the operator is not visualized, although he/she might be present in the room. This can lead to a strange situation and discomfort, as the subject feels like they do not know what is going on outside the virtual environment. This situation can be especially uncomfortable when the operator starts talking to or touching the participant. To deal with this situation and to be closer to the real environment (depending on the research design), it might be beneficial to include a visual representation of the operator.

The embodied representation of the operator in VR could offer the following advantages:

- the participant's sense of presence and immersion is not interrupted when receiving instructions from outside the virtual world
- the operator has more than the audio channel available to communicate with the participant (without the participant having to remove the HMD)
- the feeling of being watched without knowing who is watching can be reduced

Social psychology research demonstrates the potential importance of the experimenter absence or presence [31]. The influence of the visualization of the experimenter on the experience has not been investigated. The variance in the visual representation and the limited fidelity of nonverbal communication is raising additional uncertainties that need to be explored.

6.5 Data Handling

To enable a detailed evaluation of the experiment, it is important to make all relevant data accessible. The data processing framework must therefore be transparent and collect, process and manage all relevant data for analysis, visualization and sharing either at runtime or post-hoc. A major advantage of experimentation with VR is that any virtual world object can be accurately sensed and combined with the participant data provided, as highlighted in section 6.3. However, many challenges common in data

processing such as synchronization, resampling, preprocessing, and fusion, as well as data management like storing the huge amounts of data in appropriate databases and in standardized formats/protocols persists.

6.5.1 Data Import, Export, and Streaming

The goal of data storage is to ensure excellent archival properties, traceability, transparency, and repeatability. To achieve this goal, data sheets must be standardized, ensure compatibility with existing databases and protocols, and allow easy handling in statistical software or computing language of the experimenter's choice.

In addition to data *export* (i.e. in standardized .json or .csv formats) data *import* (i.e. serial data streams from Arduino sensors) can be a valuable source, e.g. to include the current time of day or weather condition or or to replay a recorded scene. While some data is not time-critical, other data needs to be provided and processed in *real-time* as a data stream, e.g. if the data needs to be closely coupled with other software.

6.5.2 Data Analysis

The use of advanced analytics (e.g., artificial neural networks, hidden Markov models, natural language processing) to automate observation processes can simplify the detection, acquisition, and classification of user states, and can be integrated with modeling approaches, such as

- *behavioral modeling* of real-time visual measurements (gaze, posture, ...),

- *emotion modeling* using electrodermal activity, heart rate, facial expression, and voice characteristics, analyzed in conjunction with event-driven data to derive emotional responses to the virtual environment or avatars,

- *communication and awareness modeling* including speech recognition and analysis for the interpretation of awareness, perception, stress, and engagement, as well as

- *cognitive modeling* to simulate human problem-solving and mental processing to determine cognitive load or discrepancy between an individual's actual state or behavior and the expected state.

The preliminary use case of data analysis methods to support the human observer with additional information to let him/her derive conclusions. In addition to using this source for further interpretation by humans, it can also be used as feedback into the virtual environment to form an interaction loop and change the setup accordingly.

It should be noted that the use of certain algorithms and methods has a significant impact on the classification results and should therefore be used with caution. An example of an emotion detection algorithm that analyzes facial features and returns the likelihood of a person's emotion can be found in [32]. Combining multimodal features such as EEG, speech, facial expressions, text, etc. can improve analysis and recognition, as described in [33].

6.5.3 Replay and Annotation

The recording of the complete VR scene enables the *replay* of test sequences, whereby time and position can be freely changed within the virtual environment. In addition to a simple replay functionality, an important support for reviewing is the ability to *augment* the environment with relevant information, either provided by data analysis or by handwritten *annotations*. For example, the toolkit could provide transcribed audio data, sentiment analysis combined with the participant's biometric data, while the research can jump through the different situations in the same way as skipping through a video file. This feature cannot only be helpful in a playback situation but can also be of great help during supervising an experiment.

6.5.4 Sharing Environment

Replication or reproduction of many scientific studies fails for many reasons [34]. For instance, in social science studies published in Nature and Science it was found that the replication rates is as low as 62% [35]. Two possible causes of the *reproducibility crisis*—as it was called in the early 2010s—are causing a lack of transparency: First, the experimental setup and environment cannot be replicated because the conditions necessary to obtain the results have not been adequately specified. Second, large data sets cannot be processed properly, making it difficult for scientists to review and reanalyze each other's data. Scientific transparency is the ideal that can be supported by pre-registering experiments[9], building software

[9] it allows to publicly disclose research questions and chosen methods before testing the hypothesis, which makes it harder to hide unfavorable results

as open-source, and making the entire experimental setup (including the VR environment) available, not just the (summarized) results.

By publishing the complete experiment as an archived executable, in source code, and raw data, including software, the complete VR environment, and results, the whole process can be made more transparent. Other researchers can carry out the exact experiment in VR as has been done before or adapt it to their own needs. These measures also counteract that tools, used to create the original data, may no longer be available (or only newer versions which might possibly alter the given results).

6.6 Feature Comparison

In this section, we compare the features provided by the different toolkits. For this purpose, the features under investigation have been segmented into four groups as outlined earlier. Comparing Tab. 6.1, we find that some features (e.g., process planning, sample scenes, eye & gaze tracking, or data export) are available in almost any toolkit, while other features (e.g., social environment, in-VR questionnaires, brain activities, facial features, operator representation) are barely supported.

Table 6.1: Comparison of the different features provided by the toolkits.

	UXF	VREX	EVE	Toggle Toolkit	VIZARD	Scene Explorer	Cyber Session	VRSTK
Setup & Control								
Process Planning	●	●	●	●	●	●	●	●
Operator Control	●	○	○	○	●	○	●	●
Sample Scenes	●	●	●	●	●	●	●	●
Scene Customization	○	●	●	●	○	●	○	◑
Remote Testing	●	○	○	○	○	○	○	◑
Social Environment	○	○	○	○	●	●	○	○
Sensing Participants								
In-VR Questionnaires	○	○	●	○	●	●	○	◑
Head-Orientation	●	●	●	○	●	●	●	◑
Controller-Input	●	●	●	●	●	●	●	◑
Eye & Gaze	○	○	●	●	●	●	●	●
Facial Features	○	○	○	○	○	○	○	●
Pose	○	○	○	○	●	○	●	●
Brain Activities	○	○	○	○	○	○	●	●
Other Biosignals	○	○	●	○	●	○	○	◑
Audio	○	○	○	○	●	○	○	◑
Video	○	○	○	○	●	●	○	◑
Representation								
Participant	○	○	○	○	○	○	○	●
Operator	○	○	○	○	○	○	○	●
Physical Objects	○	○	●	○	●	●	●	●
Data Handling								
Import	○	○	○	○	●	●	●	●
Export	●	●	●	●	●	●	●	●
Streaming	○	○	○	●	○	●	●	○
Analysis	○	○	●	○	●	●	○	◑
Scene Replay	○	○	●	○	○	●	○	●
Augmented Replay	○	○	●	○	○	●	○	◑
Annotation	○	○	○	○	○	●	○	◑
Sharing Environment	○	○	●	○	●	○	○	◑
Integration								
Open Source	●	●	●	●	○	○	○	●
Unity	●	●	●	●	○	●	○	●
Unreal	○	○	○	○	○	●	●	○

○ not supported/listed, ◑ partly supported or planned, ● fully supported

Some of the discussed methods show their full effect in combination with other methods; e.g., in remote testing, the real-world environment to which the participant is exposed to is unknown (in contrast to on-site experiments) and could influence his/her decision. By using in-VR surveys, this source of potential influence can be reduced.

Since different toolkits vary in their features and implementation, e.g., in fidelity and visualization of the participant, the toolkit used could have a significant impact on the findings and might vary between toolkits.

6.7 Conclusion & Outlook

By comparing different toolkits, we make the currently available features in different toolkits more transparent and provide a basis for researchers from different fields to decide which toolkit can provide the most benefit for their needs. In addition to analyzing existing toolkits, we have introduced our self-developed toolkit, dubbed VRSTK, available as open-source, that reflects our thoughts. As we have used the VRSTK in many applications, we have been able to confirm that the VRSTK is helpful by simplifying the VR experiment development process as well as for post-experiment analysis.

By proposing and arguing novel features that are not yet present in most toolkits, and by discussing the possibilities and implications, we hope to provide guidance for future software development for VR-based human behavior research. We hope that new sophisticated toolkits for conducting experiments will emerge, or that existing toolkits will be extended, adding functionality and enabling novel combinations of existing functionality

not yet considered by available toolkits. Beyond simplifying setup and test procedures, we hope that this work will also help to support the establishment of processes in VR-based behavioral research to become increasingly comparable and replicable.

As described in [36], there may be a difference between the mental model of the real application and the mental model of its virtual representation. Therefore, transferring findings from VR to the real world cannot be taken for granted and special attention must be paid to whether these findings are actually transferable. Regardless of whether and what kind of toolkit is used, it is necessary to include sanity checks and prove the validity of the findings.

References

[1] J. Brookes, M. Warburton, M. Alghadier, M. Mon-Williams, and F. Mushtaq, "Studying human behavior with virtual reality: The Unity Experiment Framework," en, *Behavior Research Methods*, vol. 52, no. 2, pp. 455–463, Apr. 2020.

[2] M. Vasser, M. Kängsepp, M. Magomedkerimov, *et al.*, "VREX: An open-source toolbox for creating 3D virtual reality experiments," *BMC Psychology*, vol. 5, no. 1, p. 4, Feb. 2017.

[3] P. Ugwitz, A. Šašinková, Č. Šašinka, Z. Stachoň, and V. Juřík, "Toggle toolkit: A tool for conducting experiments in unity virtual environments," en, *Behavior Research Methods*, Jan. 2021.

[4] J. Grübel, R. Weibel, M. H. Jiang, C. Hölscher, D. A. Hackman, and V. R. Schinazi, "Eve: A framework for experiments in virtual environments," in *Spatial Cognition X*, Springer, 2016, pp. 159–176.

[5] J. Reichenberger, S. Porsch, J. Wittmann, V. Zimmermann, and Y. Shiban, "Social Fear Conditioning Paradigm in Virtual Reality: Social vs. Electrical Aversive Conditioning," English, *Frontiers in Psychology*, vol. 8, 2017, Frontiers. DOI: 10.3389/fpsyg.2017.01979.

[6] J. Rodrigues, M. Müller, A. Mühlberger, and J. Hewig, "Mind the movement: Frontal asymmetry stands for behavioral motivation, bilateral frontal activation for behavior," en, *Psychophysiology*, vol. 55, no. 1, e12908, 2018.

[7] D. Saffo, S. Di Bartolomeo, C. Yildirim, and C. Dunne, "Remote and collaborative virtual reality experiments via social VR platforms," in *CHI Conference on Human Factors in Computing Systems*, ACM, 2021.

[8] X. Pan and A. F. d. C. Hamilton, "Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape," *British Journal of Psychology*, vol. 109, no. 3, pp. 395–417, 2018.

[9] L. Almeida, E. Lopes, B. Yalçinkaya, *et al.*, "Towards natural interaction in immersive reality with a cyber-glove," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, ISSN: 2577-1655, Oct. 2019, pp. 2653–2658. DOI: 10.1109/SMC.2019.8914239.

[10] M. Argyle, *Bodily Communication*, en. Routledge, Apr. 2013. DOI: 10.4324/9780203753835.

[11] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: A general-purpose brain-computer interface (BCI) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1034–1043, Jun. 2004.

[12] T. Arakawa, "Recent Research and Developing Trends of Wearable Sensors for Detecting Blood Pressure," en, *Sensors*, vol. 18, no. 9, p. 2772, Sep. 2018, Number: 9 Multidisciplinary Digital Publishing Institute.

[13] V. C. Pezoulas, T. P. Exarchos, and D. I. Fotiadis, "Chapter 2 - types and sources of medical and other related data," in *Medical Data Sharing, Harmonization and Analytics*, V. C. Pezoulas, T. P. Exarchos, and D. I. Fotiadis, Eds., Academic Press, 2020, pp. 19–65.

[14] A. Supratak, C. Wu, H. Dong, K. Sun, and Y. Guo, "Survey on feature extraction and applications of biosignals," in *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges*, A. Holzinger,

Ed. Springer International Publishing, 2016, pp. 161–182. DOI: 10.1007/978-3-319-50478-0_8.

[15] D. Douxchamps and N. Campbell, "Robust Real Time Face Tracking for the Analysis of Human Behaviour," en, in *Machine Learning for Multimodal Interaction*, A. Popescu-Belis, S. Renals, and H. Bourlard, Eds., Springer, 2008, pp. 1–10. DOI: 10.1007/978-3-540-78155-4_1.

[16] D. Roth, D. Mal, C. F. Purps, P. Kullmann, and M. E. Latoschik, "Injecting Nonverbal Mimicry with Hybrid Avatar-Agent Technologies: A Naive Approach," in *Symposium on Spatial User Interaction*, ACM, Oct. 2018, pp. 69–73.

[17] C. Morrison, P. Culmer, H. Mentis, and T. Pincus, "Vision-based body tracking: Turning Kinect into a clinical tool," *Disability and Rehabilitation: Assistive Technology*, vol. 11, no. 6, pp. 516–520, Aug. 2016.

[18] M. J. Schuemie, P. van der Straaten, M. Krijn, and C. A. van der Mast, "Research on Presence in Virtual Reality: A Survey," *CyberPsychology & Behavior*, vol. 4, no. 2, pp. 183–201, Apr. 2001, Mary Ann Liebert, Inc., publishers. DOI: 10.1089/109493101300117884.

[19] M. Price and P. Anderson, "The role of presence in virtual reality exposure therapy," en, *Journal of Anxiety Disorders*, vol. 21, no. 5, pp. 742–751, Jan. 2007. DOI: 10.1016/j.janxdis.2006.11.002.

[20] J. Knibbe, J. Schjerlund, M. Petraeus, and K. Hornbæk, "The dream is collapsing: The experience of exiting vr," in *CHI Conference on Human Factors in Computing Systems*, ACM, 2018, pp. 1–13.

[21] S. Putze, D. Alexandrovsky, F. Putze, S. Höffner, J. D. Smeddinck, and R. Malaka, "Breaking the experience: Effects of questionnaires

in VR user studies," in *CHI Conference on Human Factors in Computing Systems*, ACM, 2020, pp. 1–15.

[22] B. Hodges, G. Regehr, and D. Martin, "Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it," *Academic Medicine*, vol. 76, no. 10, S87–S89, 2001.

[23] D. Alexandrovsky, S. Putze, M. Bonfert, *et al.*, "Examining design choices of questionnaires in VR user studies," in *CHI Conference on Human Factors in Computing Systems*, ACM, 2020, pp. 1–21.

[24] J. Lou, Y. Wang, C. Nduka, *et al.*, "Realistic facial expression reconstruction for VR HMD users," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 730–743, Mar. 2020. DOI: 10.1109/TMM.2019.2933338.

[25] M. Calbi, F. Siri, K. Heimann, *et al.*, "How context influences the interpretation of facial expressions: A source localization high-density eeg study on the Kuleshov effect," *Scientific reports*, vol. 9, no. 1, pp. 1–16, 2019.

[26] K. Kilteni, R. Groten, and M. Slater, "The Sense of Embodiment in Virtual Reality," *Presence: Teleoperators and Virtual Environments*, vol. 21, no. 4, pp. 373–387, Nov. 2012.

[27] O. Blanke and T. Metzinger, "Full-body illusions and minimal phenomenal selfhood," en, *Trends in Cognitive Sciences*, vol. 13, no. 1, pp. 7–13, Jan. 2009. DOI: 10.1016/j.tics.2008.10.003.

[28] N. Yee and J. Bailenson, "The proteus effect: The effect of transformed self-representation on behavior," *Human communication research*, vol. 33, no. 3, pp. 271–290, 2007.

[29] B. K. Wiederhold, "Embodiment empowers empathy in virtual reality," *Cyberpsychology, Behavior, and Social Networking*, vol. 23, pp. 725–726, 2020.

[30] J. M. Loomis, J. J. Blascovich, and A. C. Beall, "Immersive virtual environment technology as a basic research tool in psychology," en, *Behavior Research Methods, Instruments, & Computers*, vol. 31, no. 4, pp. 557–564, Dec. 1999.

[31] M. G. Palmer and C. M. Johnson, "Experimenter presence in human behavior analytic laboratory studies: Confound it?" *Behavior Analysis: Research and Practice*, vol. 19, no. 4, p. 303, 2019.

[32] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, 2018. DOI: 10.3390/s18020401.

[33] Y. Jiang, W. Li, M. S. Hossain, M. Chen, A. Alelaiwi, and M. Al-Hammadi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," *Information Fusion*, vol. 53, pp. 209–221, 2020.

[34] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature News*, vol. 533, no. 7604, p. 452, 2016.

[35] C. F. Camerer, A. Dreber, F. Holzmeister, *et al.*, "Evaluating the replicability of social science experiments in nature and science between 2010 and 2015," *Nature Human Behaviour*, vol. 2, no. 9, pp. 637–644, 2018.

[36] M. Wölfel, "Besonderheiten beim Einsatz von immersiven Augmented und Virtual Reality Lernanwendungen," in *Künstliche Intelligenz in der beruflichen Bildung*, S. Seufert, J. Guggemos, D. Ifenthaler,

and J. Seifried, Eds. in Künstliche Intelligenz in der beruflichen Bildung, Franz Steiner Verlag, 2021.

7 Similarities and Differences between Immersive Virtual Reality, Real World, and Computer Screens: A Systematic Scoping Review in Human Behavior Studies

Publication Note: This chapter is based on the following published work. The content of the chapter is identical to the published article, with only the formatting and numbering being adapted for this dissertation.

Hepperle, D. and Wölfel, M. (2023)
Multimodal Technologies and Interaction 7, no. 6: 56
ISSN: 2414-4088 | DOI: 10.3390/mti7060056
— Open access

Abstract

In the broader field of human behavior studies, there are several trade-offs for on-site experiments. Being tied to a specific location can limit both the availability and diversity of participants. However, current and future technological advances make it possible to replicate real-world scenarios in a virtual environment, up to a certain level of detail. How these differences add up and affect the cross-media validity of findings remains a topic of debate. How a virtual world is accessed, through a

computer screen or a head-mounted display, may have a significant impact. Not surprisingly, the literature has presented various comparisons. However, while previous research has compared the different devices for a specific research question, a systematic review is lacking. To fill this gap, we conducted this review. We identified 1083 articles following the PRISMA guidelines. After screening, 56 articles remained and were compared for a qualitative synthesis that provides the reader with a summary of current research on the differences between head-mounted displays (HMDs), computer screens, and the real world. Overall, the data show that virtual worlds presented in an HMD are more similar to real-world situations than to computer screens. This supports the thesis, that HMDs are more suitable than computer screens for conducting experiments in the field of human behavioral studies.

Keywords Scoping Review – Human Behavioral Studies – Comparison – Immersive Virtual Reality – Head-Mounted Displays – Computer Screen – Real World

7.1 Introduction

To date, numerous experiments in the broader field of human behavior research have been and are being conducted in virtual environments. Virtual environments are used to overcome several trade-offs of on-site experiments. For example, being tied to a specific location can limit both the availability and diversity of participants. Although tools such as Amazon's Mechanical Turk[1], PsyToolkit[2], or Eprime 3.0[3] promise to make it easy to design experiments and collect data, they are mostly limited to content such as text, video, or images. Realistic 3D environments can overcome these limitations and can be used either on a *computer screen* or within a *head-mounted display* (HMD). Whereas the former provides a monoscopic display, the latter provides a stereoscopic display with a large field of view and content that adapts to the position of the head. In both cases, the term virtual reality (VR) is often used. To distinguish between the output devices (which also determines the type of input), "immersive" is often added in the case of HMDs. The higher degrees of freedom combined with a stereoscopic viewport, as offered by HMD VR, can make someone feel as if they are present somewhere else [1]. Combined with a well-implemented 3D world, it can be so convincing that users forget the physical space they are in. Even though the use of HMD VR seems promising, it is important to note that HMD VR, like displays, has several limitations. These may be due to technical limitations (e.g. display resolution, refresh rate, field of view), or conflicting parameters (e.g. vergence–accommodation conflict [2], visuo-proprioceptive conflict [3]) that can lead to different unwanted effects in the way VR is perceived

[1] `https://www.mturk.com/`

[2] `https://www.psytoolkit.org/`

[3] `https://pstnet.com/`

(e.g. uncanny valley effect see e.g. [4], color perception see e.g. [5] or suspension of disbelief). Despite some of these effects and limitations are exclusive to HMDs, others are shared with more established output devices such as screens or *cave automatic virtual environments* (CAVE[4]).

Although there may be a high *internal validity* (results hold true within the environment), *cross-media validity* (results hold true in other environments) cannot be assumed for experiment conducted in a virtual environment. This is especially critical when applying findings from a virtual environment to the real world. In this paper, we provide a systematic review of the evidence by comparing results between HMD VR and the real world, and between HMD VR and screens.

Although it is obvious why comparing HMD VR to the real world is important for human behavioral studies, the rationale for also comparing it to screens may not be immediately obvious. Many of the advantages of using virtual environments in human behavioral research apply to both types of output devices. However, designing, setting up, and conducting HMD VR studies is significantly more challenging than conducting VR studies with a screen. Therefore, evidence on how HMD VR differs needs to be collected to provide a basis for deciding which setup (HMD, screen or real world) to use. In addition, the collected information can serve as a valuable reference for researchers and developers who want to optimize their virtual environments for different applications and use cases.

In 2003, Frank Biocca framed the notion of presence in immersive VR as *"how the mind 'perceives' reality, not reality itself; not physics but psychology; the*

[4] The CAVE is a cube with display screen faces surrounding a viewer [6]. In some cases CAVEs may be a valid alternative to the other entities, however it is not included in this review due to the low number of results.

extended mind, the place where experience, technology, and psychology meet" [7]. This suggests that immersive VR has less to do with a physical setup and more to do with the mind. However, in order to realize the full potential of HMD VR, it is necessary to understand the technical aspects in order to know how to plan and build these experiences to achieve the desired effect.

All individual findings are listed and categorized into major categories (e.g., interaction or perception) and subcategories (e.g., efficiency or presence) for easy reference. In addition, all findings are compared to each other to provide an overview of similar and contrary results. For each paper, we also collected information about the study population, such as the number of participants, gender, and age. Regarding the research methodology, we collected information about the questionnaires used, the study design (within-groups; between-groups), and the software and hardware used to conduct the study. In addition, the goal is to identify any gaps so that new research activities can be positioned accordingly. This is essential to understand this new technology in such a way that the specific needs and possibilities of HMD VR can be addressed and compared to other research environments.

We are interested in understanding if and how HMD VR can be used to conduct research on human behavior. However, this is only possible if the conclusions drawn in (immersive) VR can be applied to the real world. Through posing and answering the following research questions, we aim to highlight the similarities and differences that require further attention when conducting research in virtual environments with the goal of applying these findings to the real world:

RQ1: *"What are the main differences between HMD VR, screen based VR and the real world mentioned in the current literature?"*

RQ2: *"What are the expected consequences of these differences?"*

RQ3: *"How extensive are these differences?"*

Our initial goal was to provide evidence specifically for human behavioral studies, yet the findings are not limited to this specific research discipline. Knowledge of similarities and differences between environments is helpful in all cases where knowledge, insights, etc. need to be transferred from one to another. Examples include education and training (e.g, surgical training [8], pilot training [9], safety training [10]), sports [11] and physiotherapy [12], human factors engineering [13] or for exposure therapies to treat anxiety or similar [14], [15].

7.2 Related Work and Theoretical Foundation

Already in the late 90s and early 2000s, papers were published that attempted to examine the differences between HMD VR and other technologies as well as actual reality. For example, Yoon, Byun, and Chung [16] published a paper in which they examined spatial perception using an HMD VR compared to how it is perceived in the real world. They found that spatial perception in general did not differ from real-world perception, but height estimates did. Although research has examined how HMD VR compares to other entities within a single paper or study, to the best of our knowledge, not much work has been done to provide a comprehensive overview. Santos, Dias, Pimentel, *et al.* [17] took the first

step in mapping the research landscape in this context more than ten years ago, most likely as a byproduct of their original intent to measure navigation performance between desktop systems and HMD VR. As the technology has advanced tremendously in the last 15 to 20 years, one might wonder if the outdated findings are still valid. For example, the HMD used by Yoon, Byun, and Chung, the *V8*[5] from Virtual Research Systems, Inc. came with a 60° diagonal *field of view* (FoV), while current consumer-grade headsets offer a diagonal FoV of 110° (HTC Vive Pro) and in some cases as much as 170° (Pimax 5K+). Fundamental research work, such as Milgram and Kishino's Virtual Reality Continuum introduced in 1994 [18], provides definitions along the reality spectrum. Also, systematic reviews examining the differences between virtual and augmented reality exist [19]. Despite these existing studies, our research seeks to address a notable gap: there has been a lack of exploration regarding the distinctions among real-world entities, immersive virtual reality, and screen environments. This became evident during our exhaustive search, where we were unable to identify any systematic scoping reviews focusing on this particular topic. In order to conduct a systematic scoping review that can contribute to scientific progress by providing a well-researched and aggregated overview of the current research landscape, we followed the guidelines suggested by [20]. They defined the goal of a scoping review as "*to determine what range of evidence (quantitative and/or qualitative) is available on a topic and to represent this evidence visually as a mapping or charting of the located data.*" As some procedures in a systematic scoping review are similar or adopted from a systematic review, we also used the PRISMA guidelines given in [21] for orientation and structure of this work.

[5] `http://www.virtualresearch.com/products/v8.htm`

7.2.1 Categories

To categorize the fundamental aspects of VR, we propose *perception*, *interaction*, and *sensing and reconstruction of reality* as the main categories.

Category I - Perception

Perception can be described as the entirety of impressions that are received by our senses. In general, immersive technologies are able to simulate these impressions to a certain degree. The more and the better this is simulated or faked, the less a person is able to distinguish between the technology and the actual real world.

Category II - Interaction

Interaction is important for any technical system with the goal that users can not only perceive information, but also manipulate it. This cooperation between a technical system and the user is necessary in cases where there is no predefined linear narrative, but where the content can be modified by the user. This category includes results in the area of efficiency (i.e., time to task completion), usability, or workload. More specifically, issues such as object manipulation, navigation, or ease of use fall into this category.

Category III - Sensing & Reconstructing Reality

No matter how hard one tries, it is impossible to completely detach from reality. Whether it's something as basic as moving furniture in your HMD VR installation, or something more subtle like temperature or smell, users are always affected by their physical surroundings. In order to achieve an optimal mapping, it is essential to sense the real world and to reconstruct it within the virtual world in some way. This does not necessarily have to be a 1-by-1 replica, but can vary to some degree [22]–[24]. It does not have to be a static reconstruction of the environment. The use of cameras or other sensor systems combined with machine learning algorithms allows you to sense or scan your environment in real time so that you can implement facial expressions or project full avatar movements and corresponding textures into the HMD VR experience [25], [26].

Subcategories

For each individual article collected in the screening process, a category was noted by the authors without restriction. These were grouped into more similar categories. Subcategories are not limited to a specific main category but can also fall into two main categories. For example, in Table 7.6, efficiency as a subcategory can be perceived efficiency (i.e. result no. 39: "sig. higher felt individual performance in VR") and therefore would fit as a subcategory of perception, but it can also be a subcategory of interaction (i.e. result no. 16 "sig. faster in time to task completion").

7.2.2 Compared Settings

In this work, we focus on the following *intervention types* (IVT) on which the comparison is based: VR (screen), immersive VR (HMD) and real world[6]. For all IVT, it is important to note that we do not differentiate between the specific content, e.g. if it is presented as an image or as a 3D point cloud, we only focus on the device type. We refer to

- *screen* as monoscopic displays in all different sizes as they are commonly used on a PC or tablet.
- *HMD VR* as all kinds of head mounted displays that visually isolate the user from the environment. Content can range from interactive stereoscopic 3D computer graphics to 360° video or photos.
- *real world* as the world that seems to exist.

7.3 Methodology

We conducted a scoping review to map the research landscape of the unique characteristics of HMD VR compared to established technologies and procedures. A VR-related example of a scoping review would be the work of [27], in which they provide an overview of work dealing with VR technology in the assessment and treatment of psychosis. This work is based on the recommended steps as suggested by [28] and [20] and the PRISMA (Preferred Reporting Items for Systematic reviews and Meta–Analyses [PRISMA] by [21]) guidelines for reporting.

[6] Contrary to what was stated in the preregistration, CAVE is no longer considered due to the low number of results

7.3.1 Risk of Bias

As with all systematic scoping reviews, there is the problem of publication bias (also known as the file drawer effect) [29], which (oversimplified) states that studies are less likely to be published if no significant effect is found. This is of particular importance in this study, since we are interested in *differences* as well as whether there are *no differences* to be expected. No differences are of particular interest with regard to real-world comparisons, since a finding of no differences would mean that HMD VR could be a valid substitute for a real-world study.

7.3.2 Query Development and Search

Creating the search query was an iterative process with several loops to define the final search string. The query can be seen in Figure 7.1. For each database (see Table 7.1), all BibTeX entries, including abstracts, were exported and imported into a locally installed version of the open source systematic literature review tool parsif.al, where all duplicates were removed [30]. Following the work of [31], in which they evaluated the respective qualities of 28 academic search systems, the ones listed in Table 7.1 were selected.

Table 7.1: Overview of Search Systems Used

Search System	URL
ACM Digital Library	`https://dl.acm.org/`
Arxiv only 2020	`https://arxiv.org/`
IEEE Xplore	`https://ieeexplore.ieee.org/`
Ovid	`https : / / ovidsp.dc1.ovid.com / ovid - a / ovidweb.cgi`[a]
Scopus	`https://www.scopus.com/home.uri`
Wiley Online Library	`https://onlinelibrary.wiley.com/`

```
("Virtual Reality" OR "HMD" OR "VR" OR "Head
Mounted Display" OR "3D" OR "Stereo 3D")
AND (("Display" OR "Monitor" OR "2D" OR
"Screen") OR ("CAVE") OR ("Real World"))
AND ("Differences" OR "Similarities" OR "Com-
parison" OR "Correlation")
AND ("User Study" OR "Evaluation")
```

Figure 7.1: Search query developed and used by the authors.

The criteria for selecting our search engines were as follows:

- The search engine must be thematically relevant. We included search systems from the fields of computer sciences; social psychological studies, behavioral studies, and health sciences; multidisciplinary ones with a focus on computer science and medicine.
- All search systems need to be able to make use of boolean operators in search strings (we used only OR; AND; NOT) [32].
- Are capable of more complex search terms (e.g., are able to make use of more than seven boolean separated search strings).

Even though a detailed pre-selection has been made, there are still some hurdles to overcome between the different search engines, especially syntactical ones. For example, we decided to search only within the abstracts of the available research articles, which in some cases had to be checked as additional criteria and sometimes could be implemented within the search string. The same happened when we tried to limit the search results to results that were newer than 2013. We considered articles newer than 2013 because that was the year the Oculus Rift DK1 shipped [33].

7.3.3 Preregistration

The study has been pre-registered and is available online at the Open Science Framework[7]. The following derivations of the pre-registration have been made. The entry fields used for collecting the data were supplemented by the entries: "VR hardware used, other hardware, comments, software used, and "what's being compared" to collect more data that might be of interest. The selection criterion "large screen" was not found. Due to the small number of results (1 each), the intervention types CAVE and Audio are not discussed in this paper.

7.4 Screening, Selection, & Assignment Procedure

The process for selecting and rejecting studies can be separated into the following four stages:

Stage 1: Immediately after searching the respective databases, the results were filtered by year (newer than 2013) if this wasn't possible with the search string.

Stage 2: The abstracts of each record were screened according to the following selection and deselection criteria: All articles related to the IVT we defined were selected (see Subsection 7.2.2 for definitions). If any of the research articles used augmented reality (AR) instead of VR, it was rejected. If an article compared both VR and AR, it was not rejected. It is also a balancing act to make the search

[7] URL to pre-registration at the Open Science Foundation: `https://osf.io/gmfns/?view_only=274f99fd32384f42a877526134227337`

query as broad as possible and as narrow as necessary. As a result, many articles were found that made a comparison *with* an HMD VR environment and not *with* another IVT. These articles were also rejected. If something other than the above IVT was compared, it would also be rejected. Languages other than English were rejected. Articles that did not adequately document their research or explain the reasoning behind their conclusions were rejected based on the "unsound methods" rejection criterion.

Stage 3: The accessibility of all papers was checked. At this stage we had to reject two more papers because they were not accessible.

Stage 4: All remaining papers were screened according to the data extraction suggested by [20] and selected or deselected accordingly. The following information was entered into the data extraction form for each paper selected after screening the abstract[8]:

1. Author(s)
2. Year of publication
3. Source of origin / country (if accessible)
4. Aims/purpose
5. Study population
6. Sample Size

[8] Here a derivation to the pre-registration has been made. The entry fields after field 13."*Bibtex entry*" were added because they were mentioned in many study descriptions anyway and are, in our opinion, a valuable addition to the mapping of the research landscape.

7. Methodology
8. Intervention Type (IVT) / Tech. Used
9. Concept
10. Duration of Intervention
11. How outcomes are measured
12. Key findings
13. Bibtex entry
14. What's compared
15. VR hardware used
16. Other hardware
17. Annotations
18. Software used

Most of the listed items are self-explanatory, but for a better understanding of how we classified the results, it is necessary to define the type of data collected under the item "What is compared". Here, the authors noted the specific topics being compared in the article. Later, this information was used as described in Subsection 7.4.1 to derive a main and subcategory that best fit the topic.

7.4.1 Postprocessing

All findings were assigned to one of the three (main)-categories *perception, interaction,* and *sensing & reconstructing reality*. The assignment of each subcategory to the appropriate category was done by two people independently. The cases that differed between the two judgments were discussed again until a decision could be made. When necessary, a subcategory such as efficiency was assigned to more than one main category.

7.4.2 Prisma Flow Diagramm

The flowchart shows the exact number of records found, selected, rejected, or removed due to duplication during the process (see Figure 7.2). The distribution of the results from the different search engines are as follows: Scopus 64%, Wiley Online Library 14%, IEE 13%, OVID 4%, ACM 4%[9], As can be seen, two additional records were selected from other articles because the cited references seemed to be a valuable contribution to this review. After identification, screening and eligibility checking, the results of the 56 articles contribute to this scoping review. Of the 56 articles finally included in this scoping review, nearly 56% compared immerisve VR (HMD) to VR (screen) and 44% compared immersive VR to the real world. Only 3 articles compared all tree settings, immerisve VR, VR, and the real world.

[9] Numbers are rounded for better readability

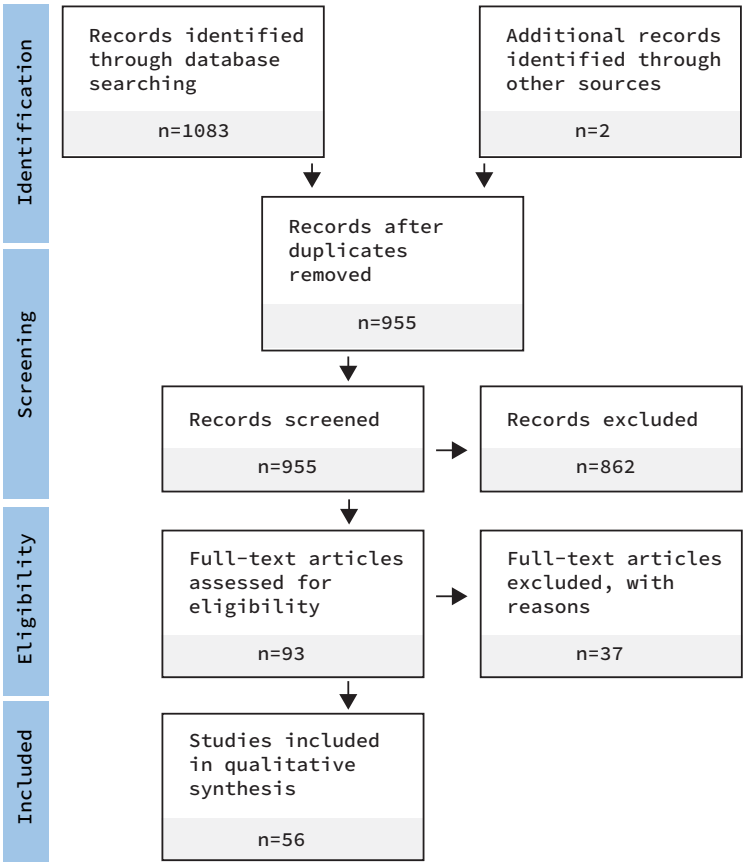

Figure 7.2: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram for the scoping review process. [21]

7.5 Results

Most of the 56 studies included in the synthesis report more than one finding, resulting in a total of 163 findings, some of which can be compared to other findings[10]. Each finding has been assigned one category and one subcategory.

To improve comprehension, individual findings in the tables 7.4, 7.5, and 7.6 are categorized and visually distinguished by icons that vary in shape

[10] Sometimes, a single research paper answered more than one question. So, we ended up with more findings than the total number of papers we looked at.

and color. This helps to indicate whether the result is from the perspective of the HMD VR:

- ▲ *advantageous* in relation to the screen or real world,
- ▼ *disadvantageous* in relation to the screen or real world,
- ► *similar* in relation to the screen or real world if there is no significant difference, and
- ■ *undecided*, if no clear tendency can be inferred, but there is a significant difference.

Table 7.4 shows the number of results for each main- and subcategory. For a better understanding, the listings in the table are to be read from the perspective of HMD VR. For example, in the first row, in the category *interaction*, the results related to the subcategory *efficiency* in comparison to the real world show that there are 3 results in favor of HMD VR, 10 results show no difference between HMD VR and the real world, 0 results are undecided and 4 results are against HMD VR. Similarly, for screen, 10 findings are favorable to HMD VR, 7 show no significant difference between HMD VR and screen, 0 are undecided, and in 2 findings it is stated that the HMD VR related task was less efficient than in the screen environment. For each subcategory, the results are summed and the distribution between favorable, similar, undecided, and unfavorable findings is expressed as a percentage.

Table 7.5 lists all results comparing HMD VR with real world results, Table 7.6 lists all results comparing HMD VR with results obtained in

a screen environment. The "Corr." and "Contr." columns group results together and compare whether they correlate or contradict each other.

7.5.1 Hard- & Software Setup

To provide a comprehensive overview of the hardware and software used, the available data from the papers are summarized as follows: 42 times an Oculus device was used[11], 33 times a device from the HTC Vive family was used, four times the Samsung Gear VR, one time the Google Daydream, and one time the Pimax 5k HMD. In 17 cases the hardware was not specified. In the case of software, we can see that—if specified—92% (44) of the studies are created using the Unity game engine, while only 8% (4) use Unreal[12].

7.5.2 Study Population and Duration

As shown in Table 7.2, the majority of participants where gender was specified (n: 1051; f: 405; m: 644; d: 2) were male (61.3%). The number of participants ranged from as few as 3 to as many as 200 (gender not specified). The age of the participants ranged from 17 to 85 years. In terms of study duration, one outlier included an observation period of a full day (8 hours). Excluding this outlier, the average study duration was 41 minutes (SD: 27).

[11] we do not count the Samsung Gear VR as Oculus, even though it is co-developed by Oculus

[12] Please note that the numbers may not add up as expected, since in some works two different HMDs were used

Table 7.2: Overview of study population of accepted articles. Some studies did not mention age and information about participants, therefore these could not be taken into account in the calculation.

	Male	**Female**	**Diverse**	**Not defined**	**Age min–max**
Avgerage	16.95	10.66	0.05	27.61	22 – 39
SD	13.45	11.35	0.21	4.14	8 – 18
n	38	38	2	22	25
$\sum$	644	405	2	574	x

7.5.3 Questionnaires Used

The questionnaires used are rather fragmented. Nevertheless, almost 70% (n: 39) of the 56 papers used a questionnaire. All questionnaires that were not developed by the authors themselves and cited accordingly can be found in Table 7.3. The questionnaires used in the 56 works assess aspects such as task load, presence, usability, user experience, engagement, and simulator sickness. The most commonly used questionnaire is the NASA Task Load Index (TLX) [34], indicating that workload or cognitive load is an important factor examined in these studies. Other commonly used questionnaires, such as Witmer & Singer's Presence Questionnaire [35] and the System Usability Scale (SUS) [36], indicate that researchers are also interested in understanding the sense of presence and usability of the systems under study.

Table 7.3: Number of Questionnaire usages found within the 56 articles that are included in this review. No.: Number of usages counted over all articles collected. Origin: Inventor of the questionnaire. Used In: Articles in which the questionnaires are used.

No.	Questionnaire	Origin	Used In
7	Task Load Index (TLX) by NASA	[34]	[37]–[43]
4	System Usability Scale (SUS)	[36]	[38], [44]–[46]
2	Witmer & Singer's Presence Questionnaire	[35]	[42], [47]
2	User Experience Questionnaire (UEQ)	[48]	[49], [50]
2	IBM CSUQ System Usability	[51]	[42], [47]
1	IGroup Presence Questionnaire (IPQ)	[52]	[39], [53]
1	After-Scenario Questionnaire (Satisfaction)	[54]	[53]
1	Immersive Experience Questionnaire (IEQ)	[55]	[56]
1	ITC-Sense of Presence Inventory	[57]	[50]
1	Player Experience of Need Satisfaction (PENS) Questionnaire	[58]	[59]
1	Self-Assessment Manikin (SAM)	[60]	[61]
1	Simulator Sickness Questionnaire	[62]	[63]
1	Temple Presence Inventory (TPI)	[64]	[65]
1	Intrinsic Motivation Inventory (IMI)	[66]	[38]
1	Virtual Reality Sickness Questionnaire (VRSQ)	[67]	[38]
1	User Engagement Scale	[68]	[69]
1	Satisfaction and Self-Confidence in Learning (SSCL)	[70]	[71]

7.5.4 Study Design

In the case of study design, a between-group design was used 36 times and a within-group design was used 40 times. The almost even distribution shows that research in HMD VR is interested in individual differences or changes within individuals over time, as well as comparing different groups. In addition, within-group designs can be useful in situations where it's difficult or impossible to recruit enough participants from a particular group.

7.5.5 Mapping the Field

In summary, we get an overview on which topic most research has been done. We can see that more than 61.3% (n: 100) of the results refer to a comparison between HMD VR and the screen environment, while only 38.7% (n: 63) of the results compare HMD VR with a real world scenario. Most of the research done in the HMD VR × screen setting falls into the category *perception* (n: 57) next to *interaction* (n: 32) and 15 of the findings relate to the category *sensing and reconstructing reality*. In the HMD VR × real comparison, however, the results are more evenly distributed across the categories. Similar to HMD VR × screen, *sensing and reconstructing reality* is the category with the fewest results (n: 16). However, the number of results for the categories *interaction* and *perception* are reversed. This means that for *interaction* there are the most (n: 31) and for *perception* the second most (n: 17) results for HMD vs. real.

In both comparison settings, the most researched subcategory is *efficiency* with 19 results for HMD VR × screen and 17 results for HMD VR × real world. Other highly investigated subcategories for HMD VR × screen are *workload* with 11 results (in the main category interaction, 10 in perception), *presence* with 8, and *learning* with 7. For HMD VR × real world they are *workload* with 7 results (6 for interaction, 1 for perception), *engagement* with 5 results, and *spatial perception* with 5 results. In addition, the range of questionnaires used in these papers highlights the complex nature of studying immersive virtual reality, computer screens, and the real world, as well as the need for multiple instruments to capture the different dimensions of user experience in these environments.

Table 7.4: The table shows the number of results distributed among each sub-category comparing HMD VR to real world and screen. **[▲]**: Nr. of results that are advantageous towards HMD VR; **[▶]**: Nr. of similar results (no sig. difference found); **[■]**: Nr. of indecisive results – not able to infer a tendency; **[▼]**: Nr. of results in which HMD VR is a drawback.

		HMD VR in comparison to:							
	Category	Real World				Screen			
	Sub-Cat.	▲	▶	■	▼	▲	▶	■	▼
Interaction	Efficiency	3	10	0	4	10	7	0	2
	Interaction	0	3	0	0	2	0	0	0
	Overview	0	0	0	0	2	0	0	0
	Physical Demand	0	0	0	0	0	0	0	1
	Simulator Sick.	0	0	0	1	0	0	0	0
	Usability	0	1	0	1	0	3	0	2
	Usefulness	0	1	0	0	0	0	0	0
	User Experience	0	1	0	0	0	0	0	0
	Workload	1	2	0	3	0	1	0	0
	∑	4	18	0	9	14	11	0	5
	%	13	58	0	29	47	38	0	16
Perception	Aesthetics	0	1	0	0	0	0	0	0
	Accuracy	0	0	0	0	1	0	0	0
	Color	0	0	0	0	0	0	2	0
	Efficiency	0	0	0	0	3	0	1	0
	Emotions	0	1	1	0	0	0	0	0
	Engagement	0	4	0	1	3	3	0	0
	Experience	0	0	0	0	1	0	0	0
	Frustration	0	0	0	0	3	0	0	1
	Immersion	0	0	0	0	3	0	0	1
	Learning	0	2	0	0	2	2	0	3
	Motion Sickness	0	1	0	1	0	0	0	0
	Perception	0	0	0	0	0	1	0	0
	Presence	1	1	0	0	6	2	0	0
	Qual. of Exp.	0	0	0	0	0	0	0	1
	Realism	0	1	0	1	1	0	0	0
	Satisfaction	0	0	0	0	2	0	0	0
	Simulator Sickness	0	0	0	0	0	0	0	1
	Spatial Perception	0	0	0	0	1	0	0	0
	Workload	1	0	0	0	4	2	0	4
	∑	2	11	1	3	33	10	3	11
	%	12	65	6	18	58	18	5	19
Sensing and Reconstr.	Accuracy	0	0	0	0	1	1	1	0
	Autonomy	0	0	0	0	1	0	0	0
	Efficiency	0	0	0	0	0	1	0	0
	Flexibility	1	0	0	0	0	0	0	0
	Haptics	0	1	0	0	0	0	0	0
	Interaction	1	0	0	0	0	0	0	0
	Learning	1	0	1	0	0	0	0	0
	Locomotion	0	0	0	1	0	0	1	0
	Overview	0	0	0	0	1	0	0	0
	Physi. Response	0	0	0	0	0	1	0	0
	Realism	0	1	0	0	0	0	0	0
	Reconstruction	1	0	1	0	0	0	0	0
	Spatial Perception	0	2	0	3	0	1	0	1
	Transferability	0	0	1	0	0	0	0	0
	Usability	0	0	0	1	0	0	0	0
	∑	4	4	3	5	3	4	2	1
	%	25	25	19	31	30	40	20	10

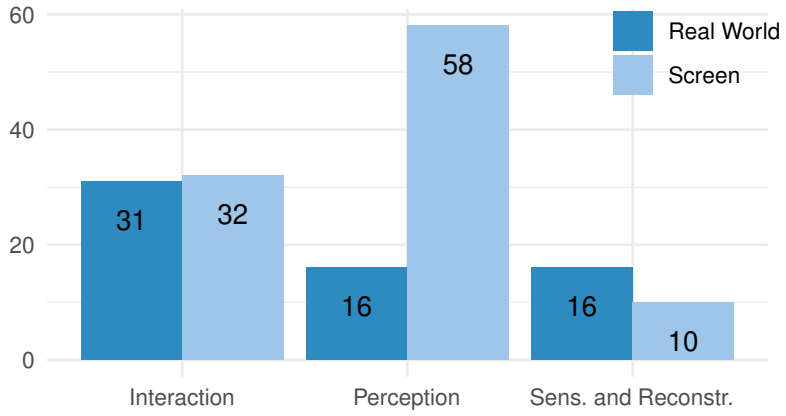

Figure 7.3: Number of results for the three main categories by intervention type

7.5.6 Advantages and Disadvantages in General

One can conclude from the results that HMD VR is advantageous in 37% (n: 61; 51 × screen; 10 × real world) of the 163 results and disadvantageous in 21% (n: 34) of the cases (17 × screen; 16 × real world). 5.5% (n: 9) cases cannot be classified due to the nature of the finding (5 × screen; 4 × real world). For example, finding number 34 "sig. stronger fear" in HMD VR × real world may be positive if VR elicits higher emotional arousal, but fear may also be a disadvantage. The other 36% (n: 59) findings (26 × screen; 33 × real world) are categorized as *no difference*[13]. As a main finding, we observe a high number of *similarities* in the results for HMD VR × real world. More than 50% (n: 33) of the 63 results for VR × real world show no significant difference. This is of great interest because some differences are likely due to technical limitations that may be resolved in the future.

We argue that it is of utmost importance to understand and evaluate the specific characteristics when looking first at the IVT, such as real world or screen, and then at the category. In some cases, a portion of

[13] When assigned to the "no difference" category, no significant difference could be found between HMD VR and the compared entity.

the subcategories may be the same for another category. This is due to the nature of the research and sometimes comes to its limits in breaking results down. Here we invite the reader to take a closer look at the work in question, as it cannot be described in more detail within the scope of this work.

7.5.7 Advantages and Disadvantages per Category

Examining the advantages, disadvantages, and similarities per category, allows us to gain more information about **RQ 1:** *"What are the main differences between HMD VR, screen based VR and the real world mentioned in the current literature?"*. Note that in this section we're interested in both differences (advantages and disadvantages) and similarities. If the mentioned percentages don't add up to 100%, this is either due to rounding or to the fact that for this category there are results categorized as "indecisive", which cannot be assigned to either differences or similarities.

Category Interaction

When comparing HMD VR × real world, 13% (n: 4) of the results showed HMD VR to be advantageous, and 29% (n: 9) of the results showed HMD VR to be disadvantageous. Almost 60% (n: 18) of the results showed no significant difference. This means that there are 42% (n: 13) differences compared to 58% (n: 18) similarities. Overall, we found more similarities than differences for HMD VR × real world in the *interaction* category. On the other hand, when comparing HMD VR × screen, we find 62% differences (n: 20) and 37.5% similarities (n: 12). Therefore, the results suggest

that HMD VR is more similar to real-world environments than to screen environments in terms of the *interaction* category.

Category Perception

In the category *perception* we find similar tendencies as for *interaction*. We observe 64% (n: 11) results that are similar between HMD VR × real world and only 29% (n: 5) that are different between the two entities. For HMD VR × screen we see 77% (n: 44) differences and 18% (n: 10) similarities. Therefore, in the category *perception,* the results suggest that HMD VR is more similar to the real world than to screen environments.

Category Sensing & Reconstructing Reality

Sensing & reconstructing reality is the only category that shows more differences 56% (n: 9) than similarities 25% (n: 4) for HMD VR × real world. For the HMD VR × screen comparison, we see equally distributed differences 44% (n: 4) and similarities 44% (n: 4). Also, worth mentioning is the fact that we found only one case where HMD VR was worse than screen for the category *sensing and reconstructing reality*

7.5.8 Possible Consequences

Answering **RQ 2:** *"What are the consequences of these differences?"* we observe, when comparing HMD VR with findings from the real world and from the screen, that there are *more similarities between HMD VR × real world*

for the subcategories interaction and perception than we find similarities between HMD VR × screen. For sensing and reconstructing reality, the *similarities and differences are more evenly distributed*. Overall, the results indicate that HMD VR environments tend to be more similar to real-world environments than screen-based environments in terms of interaction and perception. This may be useful information for designers and researchers looking to create more immersive and realistic virtual experiences.

When considering **RQ 3:** *"How elaborate are these differences?"*, it seems that the similarities outweigh the differences between HMD VR × real world. This suggests the potential for using HMD VR as a platform for experimentation, as noted in chapter 1 For HMD VR × screen, the differences outweigh the similarities, which could be an important sign in cases where existing studies on a screen could be transferred to an HMD VR scenario. However, it remains uncertain how detailed these differences are. A particular drawback in this regard is the measured effect. As mentioned above, some reported sample sizes are rather small, which is in line with the findings of [72]. In order to better understand how studies are conducted in this area, future work should consider effect sizes and study design.

7.5.9 Corresponding and Contradictory Findings

For each finding, we list in columns 5 and 6 of Table 7.5 and Table 7.6 the results that are related to each other either by supporting the same hypothesis or by presenting conflicting results. Since new results are usually easier to publish than successful and unsuccessful replications

of an experiment, we have not found a 1:1 replication of an experiment that would prove a result to be more robust. To provide an overview, we believe it is useful to relate results that are in the same category and subcategory, as long as the detailed findings are thematically similar. We consider the analysis of corr. and condr. results to be an initial guide for future studies and to provide a brief overview of the current research situation, but a close examination of the respective work is required.

Single Findings HMD VR × Real World

To get a better understanding of corr. and contr. findings, we will take a closer look at each finding and list them accordingly. As mentioned before, it is of utmost importance to take a closer look at the cited literature, as these are not generalizable results, but only specific cases in which this finding applies. For example, one of the studies examined forklift operator behavior and showed high correlations with behavior observed in real-world situations.

Efficiency Most studies report no significant differences in task completion time (Nr. 3, 6, 9, 11), error rates (Nr. 12) or entry accuracy (13). Eye-gaze input (Nr. 15), felt individual performance (Nr. 4) and task-related focus (Nr. 5) are reported to be advantageous in HMD VR. HMD VR is disadvantageous for some studies found reaches to be less efficient (Nr. 7) and higher time to task completion in VR (Nr. 8) slower object placement (Nr. 1), slower touch input (Nr. 14).

Interaction Interaction skills show no significant difference (Nr. 18, 19) and similar qualitative feedback (Nr. 20) between VR and the real world

Simulator Sickness Higher simulator sickness is reported in VR (Nr. 21)

Usability Usability results are mixed, with no significant differences found in some studies (Nr. 22) and lower scores for ease of use in VR in others (Nr. 23)

Usefulness VR-based aging simulation is found to have the same potential as real-world aging suits in terms of usefulness (Nr. 24)

User Experience No significant difference in user experience is reported between VR and the real world (Nr. 25)

Workload Workload results are mixed, with some studies reporting no significant differences in cognitive load (Nr. 26, 30) and others reporting higher mental demand (Nr. 27) and lower workload in VR (Nr. 31)

Aesthetics No difference in aesthetic preferences between VR and the real world (Nr. 32)

Emotions Emotion findings are mixed, with no significant difference between VR and video for most emotion arousal (Nr. 33) but stronger fear in VR (Nr. 34)

Engagement Engagement findings are varied, with no difference in engagement (Nr. 35), rapport (Nr. 37), co-presence (Nr. 38) and interpersonal trust (Nr. 39). Yet, one study reported lower engagement in VR (Nr. 36),

Learning No significant learning differences between learning (Nr. 40, 41, 52) but contradicting results exist (Nr. 51)

Motion Sickness More symptoms of "focus difficulty", "general discomfort", "nausea", and "headache" in VR (Nr. 42), but no difference in accommodation response (Nr. 43)

Presence Presence findings are mixed, with no significant difference in presence (Nr. 44) but a higher sense of presence in VR (Nr. 45)

Realism No significant differences between evaluation based on real user (supernumerary) in real world and avatars (Nr. 46), but lower natural feeling in VR (Nr. 47)

Single Findings HMD VR × Screen

Similar as in Subsubsection 7.5.9 a brief overview of the findings with more than 1 results will be discussed:

Efficiency With 10 results in favour for HMD VR, results in sub-category efficiency shows a clear tendency towards HMD VR.

Overview Overview also leans towards VR, with results showing that data overview and data depiction (Nr. 22, 23) are more intuitive in VR.

Immersion, Experience Studies report higher immersion in VR (Nr. 50, 55, 56, 57) and lower frustration levels (Nr. 51), but also disadvantages like lower quality of experience (Nr. 75, 79) and a decrease in immersion at the narrative level (Nr. 58)

Learning Learning presents mixed results. Some studies suggest no significant differences in correct insights (59) other suggest less correct insights in VR (60). Others report fewer deep insights from VR (62), less learning in VR (64), but also higher recall of information about tasks in VR (63) and higher motivation in learning (65)

Presence Presence in VR is generally found to be higher (68, 69, 70, 71, 73, 74), although two studies report no significant difference (67, 72)

Satisfaction Data exploration is considered more satisfying in VR (77), and VR is found to be more engaging (78)

Workload Workload results are mixed, with some studies reporting lower workload in VR (82, 84, 88), but others indicating higher cognitive load (85, 86, 89).

7.6 Discussion and Future Directions

With this work we provide an overview of 163 findings from 56 papers concerning the current research landscape on differences and similarities between HMD VR and the entities real world and screen. All findings are grouped into three main categories *interaction*, *perception* and *sensing & reconstructing reality*, which are further subdivided into more elaborate subcategories to evaluate differences and similarities in more detail. The

study presents a summary of the research used questionnaires (see Subsection 7.5.3) and applied study design (see Subsection 7.5.4), population (see Subsection 7.5.2), and hard-& software setup (see Subsection 7.5.1) for studies that have been conducted in the area of virtual reality research. Researchers can build on this knowledge and design more effective and rigorous experiments. In addition, the findings from the scoping review may indicate the extent to which cross-media validity can be assumed or needs to be questioned, so that findings in one environment may or may not be transferable to another. The review of questionnaires and hardware helps to select the most appropriate measurement for their own studies, while the summary of population characteristics can help to understand the degree of generalizability of the results.

All findings are listed and related to other findings because, as is often the case in science, there is no single truth that can be taken for granted, but rather many different aspects that need to be considered. We have identified the following three most important findings:

- In proportion, there are *more findings that show similarities* between HMD VR × real world than there are findings that show *differences* between the HMD VR and the real world. Especially for the category "interaction" as well as for the category "perception". Only in the category "sensing and reconstructing reality" we find more differences than similarities. This is different for HMD VR × screen, where we collected *more findings showing differences* between the HMD VR × screen environment for the categories *interaction* and *perception*. The category *sensing & reconstructing reality* is evenly distributed.

- For both entities there are findings that need to be considered further. For example, in HMD VR × Screen, learning shows mixed results (2 in favor of HMD VR, 2 undecided, and 3 against). This may indicate that typical learning scenarios cannot be transferred "as is" to HMD VR, but that content and presentation type have to be adapted to the particularities of the system in order to take advantage of the specific benefits of HMD VR. This is different for HMD VR × real world where we find 2 results that now show differences between the two entities that could mean easier adoption.

- When we compare results from HMD VR with those from the real world, we observe numerous findings reporting increased symptoms of "focus difficulty", "general discomfort", "nausea", and "headache". As technology advances, we anticipate significant improvements in the design and functionality of VR systems. We predict these advancements will effectively mitigate these prevalent issues through improved display technology, enhanced ergonomics, which includes reduced weight, an elevated user experience, and greater customization, as well as innovative algorithmic solutions.

- With an average of 28 participants (SD: 22), the study population is rather small and predominantly male.

In addition, we see an increase in software that supports setting up and conducting user studies in HMD VR for different disciplines such as "toggle toolkit" [73], "EVE" [74] or "VREX" [75]. An overview of current toolkits can be found here [76]. Platforms such as these can not only help to create a research environment, but can also help to standardize and optimize

recurring features, such as the implementation of a questionnaire within the HMD VR environment.

To provide an outlook, we emphasize that the current ability of HMD VR to elicit responses and sensations that are close or similar to real-life experiences implies that HMD VR offers applications and uses beyond the often stated "gaming" purpose. In particular, HMD VR may offer promising opportunities in fields such as medicine, psychology, and other areas related to human behavior.

Although both screens and HMDs can be categorized as *technology* it is important to note that the two entities should not be treated as interchangeable. The results have shown that in many cases the outcomes are significantly different between the two entities. This does not mean that using screens to answer research questions is not a valuable approach, but it cannot replace HMDs for the reasons shown. We argue that each purpose must be evaluated individually, and efforts have to be weighed up against each other. In most cases it can be assumed that HMD tends to produce more similar results than screens.

At present, we are nowhere near a complete understanding of how immersive VR findings can be applied to real-world outcomes, but with this work we have taken a first step to provide direction for interested researchers. Increased research interest, combined with technical advances, will provide new opportunities to support knowledge transfer between HMD VR and the real world, and also to add value to established research practices, especially in cases where:

- (attention) control is important (e.g., phobia therapy or learning situations),

- participants are exposed to dangerous situations (e.g., firefighter training)
- replication and sharing is useful (applies to almost any discipline except sensitive data such as patient information),
- processes are difficult or impossible to perform in the real world (e.g., taking participants "back in time" as in reminiscence therapy), and
- cost-efficiency is desired (e.g., participants could be recruited from anywhere in the world as long as they own a HMD).

We are confident that with recent and upcoming advances in the technology, combined with a good understanding of it, the use of immersive VR will grow for various fields that require the application of virtual studies, training, etc. to the real world. With this work, we provide a first step towards establishing a guide for a better understanding of the technology in relation to established environments, so that the respective advantages and disadvantages can be understood and implemented accordingly.

Table 7.5: HMD VR compared to real world:
[▲]: HMD VR is advantageous [▼]: HMD VR is disadvantageous [▶]: No difference between conditions [■]: Indecisive [**Corr.**] and [**Contr.**] indicate whether the finding confirms or contradicts existing findings.

	Sub-Cat.		Finding	Nr.	Corr.	Contr.	Ref.
Interaction	Efficiency	▼	Sig. slower in object placement	1			[77]
	Efficiency	▶	VR based aging simulation has same potential as RR aging suits in terms of effectiveness;	2			[78]
	Efficiency	▶	No sig. difference in time to task completion	3	9	8;29	[42]
	Efficiency	▲	Sig. higher felt individual performance in VR	4			[42]
	Efficiency	▲	Higher task-related focus in VR	5			[79]
	Efficiency	▶	No difference in Task Completion Time when adding visuo-haptic feedback	6			[80]
	Efficiency	▼	Reaches were less efficient in the VR	7			[80]
	Efficiency	▼	Higher time to task completion in VR	8	29	9	[80]
	Efficiency	▶	No sig. difference in time to task completion	9	3	8;29	[81]
	Efficiency	▶	No sig. difference in score	10			[81]
	Efficiency	▶	No sig. difference in reading performance	11			[37]
	Efficiency	▶	No sig. difference in error rates	12			[37]
	Efficiency	▶	No sig. differences for entry accuracy	13			[82]
	Efficiency	▼	Sig. slower touch input in VR	14			[82]
	Efficiency	▲	Sig. faster eye-gaze input in VR	15			[82]
	Efficiency	▶	No sig. difference in finding an object	16			[77]
	Efficiency	▶	No sig. difference for grasping time and head movement	17			[77]
	Interaction	▶	No sig. difference in interaction skills	18	19;22	23	[83]
	Interaction	▶	Operation behavior of the same task in VE is highly correlated to that in RR (r> 0.90), which suggests VR successfully induces operation behavior which is like the real operation behavior.	19	18;22	23	[84]
	Interaction	▶	Similar qualitative feedback in VR and real world condition	20			[82]
	Sim. Sick.	▼	Sig. higher simulator sickness	21			[37]
	Usability	▶	No sig. diff in usability	22	18;19	23	[42]
	Usability	▼	Sig. lower score for ease of use	23		18;19;22	[49]
	Usefulness	▶	VR based aging simulation has same potential as RR aging suits in terms of usefulness	24			[78]
	User Exp.	▶	No sig. difference in user experience	25			[50]
	Workload	▶	No sig. difference in cognitive load	26			[81]
	Workload	▼	Sig. higher metal demand in VR	27			[37]
	Workload	▼	Sig. higher physical demand in VR	28			[37]
	Workload	▼	Sig. higher time to task completion	29	8	3;9	[37]
	Workload	▶	No sig. difference in workload	30		31	[82]
	Workload	▲	Sig. lower workload in VR	31		30	[42]
Perception	Aesthetics	▶	No difference in aesthetics preferences	32			[85]
	Emotions	▶	No sig. difference between VR and video for each emotion arousal except fear	33	34		[86]
	Emotions	■	Sig. stronger fear in VR	34	33		[86]
	Engagement	▶	No difference in engagement	35		36	[65]
	Engagement	▼	Sig. lower engagement in VR	36	35		[69]
	Engagement	▶	No difference in rapport	37			[69]
	Engagement	▶	No difference in co-presence	38	44	45	[69]
	Engagement	▶	No difference in interpersonal trust	39			[69]
	Learning	▶	No learning differences between learning additive manufacturing in RR and VR	40	41;52	51	[87]
	Learning	▶	No difference in learning success	41	40;52	51	[88]
	Motion Sick.	▼	Sig. more symptoms of "focus difficulty"; "general discomfort"; "nausea"; "headache" for VR	42			[89]
	Motion Sick.	▶	No difference on accommodation response	43			[89]

Table 7.5: HMD VR compared to real world:
[▲]: HMD VR is advantageous [▼]: HMD VR is disadvantageous [▶]: No difference between conditions [■]: Indecisive [**Corr.**] and [**Contr.**] indicate whether the finding confirms or contradicts existing findings.

	Sub-Cat.		Finding	Nr.	Corr.	Contr.	Ref.
Perception	Presence	▶	No sig. diff in presence	44	38	45	[42]
	Presence	▲	Higher sense of presence in VR	45		44	[65]
	Realism	▶	No sig. differences between evaluation based on real user (supernumerary) in real world and avatars	46			[82]
	Realism	▼	Sig. lower natural feeling	47			[49]
Sens. and Recons.	Flexibility	▲	VR is advantageous compared to aging suits in terms of flexibility	48			[78]
	Haptics	▶	No sig. difference in material identification when using the TAGlove compared to perceiving the real physical objects.	49			[43]
	Interaction	▲	VR improves the external validity	50			[90]
	Learning	▲	VR kinesthetic experiences were more memorable and helped participants retain a larger number of words, despite any confounding elements that hindered their initial learning gain.	51		40;41;52	[91]
	Learning	■	Participants first remembered sig. more words in the text-only conditon (RR). A week later, the amount of words remembered between text-only and VR with kinesthetic motion was equal	52	40;41		[91]
	Locomotion	▼	Significantly higher travel times in VR	53			[92]
	Realism	▶	No difference in realism	54			[65]
	Reconstruction	■	Transfer of motor skills from RR to VR not given	55			[93]
	Reconstruction	▲	VR studies completely support literature on real-life bike rides	56			[94]
	Spatial Perc.	▼	VR less accurate in distance estimation	57	58		[80]
	Spatial Perc.	▼	VR less correct in depth judgements	58	57		[80]
	Spatial Perc.	▶	No difference in distance estimation when adding visuo haptic feedback	59			[80]
	Spatial Perc.	▼	Sig. difference in behaviour	60			[77]
	Spatial Perc.	▶	No sig. difference in distance traveled	61			[77]
	Transferability	■	Difference between therapist with experience in handling VR to therapists that had no prior experience. Therapists with experience handled the patients the same as in conventional therapy whereas without experience they did not	62			[95]
	Usability	▼	VR generates less answers directly related with the mockup and more related to the surrounding	63			[90]

Table 7.6: HMD VR compared to screen environment.

[▲]: HMD VR is advantageous | [▼]: HMD VR is disadvantageous [▶]: No difference between conditions [■]: Indecisive [**Corr.**] and [**Contr.**] indicate whether the finding confirms or contradicts existing findings.

	Sub-Cat.		Finding	Nr.	Corr.	Contr.	Ref.
Interaction	Efficiency	▼	Sig. slower filling out questionnaire in VR	1	5	3;7;10;12;17	[96]
	Efficiency	▲	Data exploration to be more successful in VR	2	11;22;23;77;4		[41]
	Efficiency	▶	No sig. difference in time to task completion	3	7;9;10;12	5;1;17	[42]
	Efficiency	▶	Data distinction similar	4			[40]
	Efficiency	▼	Time to task completion larger (slower) in VR	5	1	3;7;10;12;17	[97]
	Efficiency	▲	Performed better for Design Thinking tasks in VR	6			[98]
	Efficiency	▶	No difference in time to task completion	7	3;9;10;12	5;1;17	[99]
	Efficiency	▲	Reduced task error rate in VR	8			[99]
	Efficiency	▶	No differences in task completion time	9	3;7;10;12	5;1;17	[46]
	Efficiency	▶	No sig. difference in time to task completion	10	3;7;9;12	5;1;17	[47]
	Efficiency	▲	VR more efficient in data exploration	11	2;22;23;4		[100]
	Efficiency	▶	No sig. difference in time to task completion	12	3;7;9;10	5;1;17	[81]
	Efficiency	▶	No sig. difference in score	13			[81]
	Efficiency	▲	Sig. faster in annotation task	14			[101]
	Efficiency	▲	Sig. faster in counting	15			[101]
	Efficiency	▲	Sig. faster in time to task completion	16	17	3;5;7;9;10;12	[102]
	Efficiency	▲	Sig. faster in time to task completion	17	16	3;5;7;9;10;12	[38]
	Efficiency	▲	Sig. performance increase	18	14;15;19		[38]
	Efficiency	▲	Sig. faster in VR	19	14;15;18		[103]
	Interaction	▲	Interaction is more intuitive in VR	20	21;22;23	26	[40]
	Interaction	▲	Better interaction quality	21	20		[45]
	Overview	▲	Data overview is easier in VR	22	20;23	26	[40]
	Overview	▲	Data depiction more intuitive in VR	23	20;22	26	[40]
	Phys. Demand	▼	VR data exploration required significantly more physical demand	24		82	[41]
	Usability	▶	No sig. difference in usability	25	26;27;32	28;29;31	[42]
	Usability	▶	No difference in intuitive controls	26	25;27;32	28;29;31	[59]
	Usability	▶	No sig. difference in usability	27	25;26;32	28;29;31	[47]
	Usability	▼	Sig. lower score in System Usability Scale Questionnaire	28	29	25;26;27	[44]
	Usablity	▼	VR is sig. harder to use	29	28	25;26;27;31	[104]
	Workload	▶	No sig. difference in cognitive load	30			[81]
	Usability	▲	Sig. better usable	31	20	25;26;27;28	[38]
	Usability	▶	No sig. difference in usability	32	25;26;32	28;29;31	[96]
Perception	Accuracy	▲	Participants were better in estimating size in larger scales in VR	33	34;35;36	99	[105]
	Accuracy	▲	Participants were better in estimating size in smaller scales in VR	34	33;35;36	99	[105]
	Accuracy	▲	Less error in height estimation in VR	35	33;34;36	99	[105]
	Accuracy	▲	Sig. lower error rate for shape and distance estimation	36	33;34;34	99	[103]
	Color	■	Higher luminance and chroma perception in VR	37			[106]
	Color	■	Higher amount of retinal illuminance in VR	38			[106]
	Efficiency	▲	Sig. higher felt individual performance in VR	39	40;42		[42]
	Efficiency	▲	VR improves perceived collaborative success	40	39;42		[99]
	Efficiency	▲	Sig. better perceived content organization	41	77		[71]
	Efficiency	■	Participants reported subjectively that they performed best in rich VR environment while they actually were not	42	39;40		[105]
	Engagement	■	Spent more time on the storytelling process when using VR	43			[56]
	Engagement	▲	Sig. higher engagement in VR	44	54		[69]
	Engagement	▶	No difference in rapport	45			[69]
	Engagement	▶	No difference in co-presence	46			[69]
	Engagement	▶	No difference in interpersonal trust	47			[69]
	Engagement	▲	VR was considered more engaging	48			[107]
	Engagement	▲	Sig. more interest and enjoyment	49	44;54		[38]

Table 7.6: HMD VR compared to screen environment.
[▲]: HMD VR is advantageous [▼]: HMD VR is disadvantageous [▶]: No difference between conditions [■]: Indecisive [**Corr.**] and [**Contr.**] indicate whether the finding confirms or contradicts existing findings.

	Sub-Cat.		Finding	Nr.	Corr.	Contr.	Ref.
Perception	Experience	▲	Higher immersion in VR	50	55;56;57	58	[45]
	Frustration	▲	Lower frustration levels in VR	51			[40]
	Frustration	▲	Sig. higher in perceived enjoyment	52			[71]
	Frustration	▼	Sig. higher frustration	53		44;54	[96]
	Frustration	▲	Sig. more fun in VR	54	44; 49		[108]
	Immersion	▲	Data immersion is larger in VR	55	50;56;57	58	[40]
	Immersion	▲	More immersive experience in VR	56	50;55;57	58	[56]
	Immersion	▲	Perceptual immersion higher in VR	57	50;55;56	58	[61]
	Immersion	▼	Immersion on narrative level lower in VR	58		50;55;56;57	[61]
	Learning	▶	No differences in correct insights	59		60	[41]
	Learning	▼	Less incorrect insights through VR	60		59	[41]
	Learning	▶	No differences in hypotheses generated	61		62	[41]
	Learning	▼	Fewer deep insights from within VR	62		61	[41]
	Learning	▲	User in VR can recall more information	63		64	[47]
	Learning	▼	Learned less in VR	64		63	[109]
	Learning	▲	Sig. higher motivation in learning	65			[71]
	Perception	▶	No difference in mesh resolution preferences	66			[110]
	Presence	▶	No sig. difference in presence	67	72	68;69;70;71;73	[42]
	Presence	▲	Higher presence in VR	68	69;70;71;73	67;72	[59]
	Presence	▲	Higher presence in VR condition	69	68;70;71;73	67;72	[47]
	Presence	▲	Higher presence in VR	70	69;70;71;73	67;72	[109]
	Presence	▲	Sig. stronger sense of presence	71	69;70;71;73	67;72	[38]
	Presence	▶	No sig. difference in presence	72	67	68;69;70;71;73	[96]
	Presence	▲	Sig. higher feeling of professor talking	73	69;70;71;73	67;72	[108]
	Presence	▲	Sig. higher feeling of talking to class with others	74	69;70;71;73	67;72	[108]
	Experience	▼	VR offers lower quality of experience	75	79		[104]
	Realism	▲	Meshes were perceived sig. more realistic	76			[110]
	Satisfaction	▲	Data exploration to be more satisfying in VR	77	2;11;22;23		[41]
	Satisfaction	▲	VR the most engaging	78	49		[97]
	Sim. Sick.	▼	VR induced sig. higher simulator sickness	79	75		[63]
	Spat. Perc.	▲	Better spatial perception in VR	80	33;34;35;36	99	[111]
	Workload	▼	VR shows elevation in electrodermal activity	81			[104]
	Workload	▲	Sig. lower workload in VR	82	84;88	83	[42]
	Workload	▶	No differences in workload	83		82	[40]
	Workload	▲	VR required less effort	84	82;88	83	[97]
	Workload	▼	Higher cognitive load in VR	85	86;89		[112]
	Workload	▼	Higher cognitive load in VR	86	85;89		[109]
	Workload	▶	No sig. difference in physical performance	87			[113]
	Workload	▲	Sig. lower effort	88	82;84	83	[38]
	Workload	▼	Sig. higher mental demand in VR	89	85;86		[96]
	Workload	▲	Sig. higher concentration rate in VR	90			[108]
Sens. Rec.	Accuracy	■	Perceived accuracy higher despite similar results	91			[97]
	Accuracy	▶	No differences in completion accuracy	92			[46]
	Accuracy	▲	Higher classification accuracy (EEG) in VR	93			[112]
	Autonomy	▲	Higher Autonomy in VR	94			[59]
	Efficiency	▶	No sig. differences in lane change performance	95			[63]
	Locomotion	■	Users in VR condition walked further	96			[47]
	Overview	▲	VR improves quality of view	97			[99]
	Phys. Resp.	▶	No sig. differences regarding physiological repsonses	98	87		[63]
	Spat. Perc.	▶	No difference in distance perception between all conditions	99		33;34;35;36;80	[97]
	Spat. Perc.	▼	Sig. lower realism in VR	100			[96]

Author Contributions "Conceptualization, Daniel Hepperle and Matthias Wölfel.; methodology, Daniel Hepperle.; software, Daniel Hepperle; validation, Daniel Hepperle, and Matthias Wölfel.; formal analysis, Daniel Hepperle; investigation, Daniel Hepperle; resources, Daniel Hepperle; data curation, Daniel Hepperle; writing—original draft preparation, Daniel Hepperle; writing—review and editing, Daniel Hepperle and Matthias Wölfel; visualization, Daniel Hepperle; supervision, Daniel Hepperle, and Matthias Wölfel; All authors have read and agreed to the published version of the manuscript."

Funding No funding

Conflicts of Interest "The authors declare no conflict of interest."

References

[1] J. Zheng, K. Chan, and I. Gibson, "Virtual reality," *IEEE Potentials*, vol. 17, no. 2, pp. 20–23, 1998. DOI: 10.1109/45.666641.

[2] J. P. Wann, S. Rushton, and M. Mon-Williams, "Natural problems for stereoscopic depth perception in virtual environments," vol. 35, no. 19, pp. 2731–2736, 1995. DOI: 10.1016/0042-6989(95)00018-u.

[3] C. Fossataro, A. Rossi Sebastiano, G. Tieri, *et al.*, "Immersive virtual reality reveals that visuo-proprioceptive discrepancy enlarges the hand-centred peripersonal space," *Neuropsychologia*, vol. 146, p. 107 540, 2020. DOI: https://doi.org/10.1016/j.neuropsychologia.2020.107540.

[4] D. Hepperle, H. Ödell, and M. Wölfel, "Differences in the uncanny valley between head-mounted displays and monitors," in *2020 International Conference on Cyberworlds (CW)*, IEEE, Sep. 2020. DOI: 10.1109/cw49994.2020.00014.

[5] A. Siess and M. Wölfel, "User color temperature preferences in immersive virtual realities," *Computers & Graphics*, vol. 81, pp. 20–31, Jun. 2019. DOI: 10.1016/j.cag.2019.03.018.

[6] C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J. C. Hart, "The CAVE: Audio visual experience automatic virtual environment," *Communications of the ACM*, vol. 35, no. 6, pp. 64–72, Jun. 1992. DOI: 10.1145/129888.129892.

[7] F. Biocca, "Media and the laws of the mind (preface)," in *Being There Concepts, Effects and Measurements of User Presence in Synthetic Environments*, G. Riva, F. Davide, and W. IJsselsteijn, Eds., 5th ed., vol. 5, IOS Press, 2003, pp. V–VII.

[8] R. Lohre, A. J. Bois, G. S. Athwal, D. P. Goel, on behalf of the Canadian Shoulder, and E. S. (CSES)*, "Improved complex skill acquisition by immersive virtual reality training: A randomized controlled trial," *JBJS*, vol. 102, no. 6, 2020.

[9] I. S. Cardenas, C. N. Letdara, B. Selle, and J.-H. Kim, "Immersifly: Next generation of immersive pilot training," in *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2017, pp. 1203–1206. DOI: 10.1109/CSCI.2017.212.

[10] A. Grabowski and J. Jankowski, "Virtual reality-based pilot training for underground coal miners," *Safety Science*, vol. 72, pp. 310–314, 2015. DOI: https://doi.org/10.1016/j.ssci.2014.09.017.

[11] D. L. Neumann, R. L. Moffitt, P. R. Thomas, *et al.*, "A systematic review of the application of interactive virtual reality to sport," *Virtual Reality*, vol. 22, no. 3, pp. 183–198, Sep. 2018. DOI: 10.1007/s10055-017-0320-5.

[12] M. Bordeleau, A. Stamenkovic, P.-A. Tardif, and J. Thomas, "The use of virtual reality in back pain rehabilitation: A systematic review and meta-analysis," *The Journal of Pain*, 2021. DOI: https://doi.org/10.1016/j.jpain.2021.08.001.

[13] M. Oberhauser and D. Dreyer, "A virtual reality flight simulator for human factors engineering," *Cognition, Technology & Work*, vol. 19, no. 2, pp. 263–277, Sep. 2017. DOI: 10.1007/s10111-017-0421-7.

[14] R. Tadayon, C. Gupta, D. Crews, and T. McDaniel, "Do trait anxiety scores reveal information about our response to anxious situations? a psycho-physiological vr study," in *Proceedings of the 4th International Workshop on Multimedia for Personal Health and Health*

Care, ser. HealthMedia '19, Nice, France: Association for Computing Machinery, 2019, pp. 16–23. DOI: 10.1145/3347444.3356239.

[15] A. Rizzo, J. Cukor, M. Gerardi, *et al.*, "Virtual reality exposure for ptsd due to military combat and terrorist attacks," *Journal of Contemporary Psychotherapy*, vol. 45, no. 4, pp. 255–264, Dec. 2015. DOI: 10.1007/s10879-015-9306-3.

[16] J. Yoon, E. Byun, and N. S. Chung, "Comparison of space perception between a real environment and a virtual environment," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 44, no. 5, pp. 515–518, 2000. DOI: 10.1177/154193120004400508.

[17] B. S. Santos, P. Dias, A. Pimentel, *et al.*, "Head-mounted display versus desktop for 3d navigation in virtual reality: A user study," *Multimedia Tools and Applications*, vol. 41, no. 1, pp. 161–181, Aug. 2008. DOI: 10.1007/s11042-008-0223-2.

[18] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino, "Augmented reality: A class of displays on the reality-virtuality continuum," in *SPIE Proceedings*, H. Das, Ed., SPIE, Dec. 1995. DOI: 10.1117/12.197321.

[19] M. J. Liberatore and W. P. Wagner, "Virtual, mixed, and augmented reality: A systematic review for immersive systems research," *Virtual Reality*, vol. 25, no. 3, pp. 773–799, Jan. 2021. DOI: 10.1007/s10055-020-00492-0.

[20] M. D. Peters, C. M. Godfrey, H. Khalil, P. McInerney, D. Parker, and C. B. Soares, "Guidance for conducting systematic scoping reviews," *International Journal of Evidence-Based Healthcare*, vol. 13, no. 3, pp. 141–146, Sep. 2015. DOI: 10.1097/xeb.0000000000000050.

[21] D. Moher, A. Liberati, J. Tetzlaff, and D. G. A. and, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Medicine*, vol. 6, no. 7, e1000097, Jul. 2009. DOI: 10.1371/journal.pmed.1000097.

[22] A. Hettiarachchi and D. Wigdor, "Annexing reality," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, May 2016. DOI: 10.1145/2858036.2858134.

[23] A. L. Simeone, E. Velloso, and H. Gellersen, "Substitutional reality," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, ACM Press, 2015. DOI: 10.1145/2702123.2702389.

[24] M. Azmandian, M. Hancock, H. Benko, E. Ofek, and A. D. Wilson, "Haptic retargeting," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, May 2016. DOI: 10.1145/2858036.2858226.

[25] Y. Guo, J. Zhang, J. Cai, B. Jiang, and J. Zheng, "Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 6, pp. 1294–1307, 2019.

[26] P. Caserman, A. Garcia-Agundez, R. Konrad, S. Göbel, and R. Steinmetz, "Real-time body tracking in virtual reality using a vive tracker," *Virtual Reality*, vol. 23, no. 2, pp. 155–168, Nov. 2018. DOI: 10.1007/s10055-018-0374-z.

[27] M. Rus-Calafell, P. Garety, E. Sason, T. J. K. Craig, and L. R. Valmaggia, "Virtual reality in the assessment and treatment of psychosis: A systematic review of its utility, acceptability and effectiveness,"

Psychological Medicine, vol. 48, no. 3, pp. 362–391, Jul. 2017. DOI: 10.1017/s0033291717001945.

[28] Z. Munn, M. D. J. Peters, C. Stern, C. Tufanaru, A. McArthur, and E. Aromataris, "Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach," *BMC Medical Research Methodology*, vol. 18, no. 1, Nov. 2018. DOI: 10.1186/s12874-018-0611-x.

[29] R. Rosenthal, "The file drawer problem and tolerance for null results.," *Psychological Bulletin*, vol. 86, no. 3, pp. 638–641, May 1979. DOI: 10.1037/0033-2909.86.3.638.

[30] V. Freitas, *Parsifal*, https://github.com/vitorfs/parsifal, last accessed: 2023-04-04, Mar. 11, 2020.

[31] M. Gusenbauer and N. R. Haddaway, "Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, PubMed, and 26 other resources," *Research Synthesis Methods*, vol. 11, no. 2, pp. 181–217, Jan. 2019. DOI: 10.1002/jrsm.1378.

[32] C. L. Cole, A. S. Kanter, M. Cummens, S. Vostinar, and F. Naeymi-Rad, "Using a terminology server and consumer search phrases to help patients find physicians with particular expertise," *Studies in Health Technology and Informatics*, vol. 107, no. MEDINFO 2004, pp. 492–496, 2004. DOI: 10.3233/978-1-60750-949-3-492.

[33] P. James, *3 years ago the oculus rift dk1 shipped, here's a quick look back*, https://www.roadtovr.com/3-years-ago-the-oculus-rift-dk1-shipped-heres-a-quick-look-back/, last accessed 2020-09-08, Mar. 28, 2016.

[34] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Human mental workload*, vol. 1, no. 3, pp. 139–183, 1988.

[35] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence*, vol. 7, no. 3, pp. 225–240, 1998. DOI: 10.1162/105474698565686.

[36] J. Brooke, ""sus: A "quick and dirty" usability scale," in *Usability Evaluation in Industry*, P. W. J. B. T. B. A. W. A. L. McClelland, Ed., Taylor and Francis Group, 1986.

[37] S. Auer, J. Gerken, H. Reiterer, and H.-C. Jetter, "Comparison between virtual reality and physical flight simulators for cockpit familiarization," in *Mensch und Computer 2021*, ACM, Sep. 2021. DOI: 10.1145/3473856.3473860.

[38] A. Elor, T. Thang, B. P. Hughes, *et al.*, "Catching jellies in immersive virtual reality: A comparative teleoperation study of ROVs in underwater capture tasks," in *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, ACM, Dec. 2021. DOI: 10.1145/3489849.3489861.

[39] R. M. S. Clifford, T. McKenzie, S. Lukosch, R. W. Lindeman, and S. Hoermann, "The effects of multi-sensory aerial firefighting training in virtual reality on situational awareness, workload, and presence," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, IEEE, Mar. 2020. DOI: 10.1109/vrw50115.2020.00023.

[40] S. V. Broucke and N. Deligiannis, "Visualization of real-time heterogeneous smart city data using virtual reality," in *2019 IEEE International Smart Cities Conference (ISC2)*, 2019, pp. 685–690.

[41] P. Millais, S. L. Jones, and R. Kelly, "Exploring data in virtual reality," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, ACM, Apr. 2018. DOI: 10.1145/3170427.3188537.

[42] S. Narasimha, E. Scharett, K. C. Madathil, and J. Bertrand, "Wersort: Preliminary results from a new method of remote collaboration facilitated by fully immersive virtual reality," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1, pp. 2084–2088, 2018. DOI: 10.1177/1541931218621470. eprint: https://doi.org/10.1177/1541931218621470.

[43] S. Cai, P. Ke, T. Narumi, and K. Zhu, "ThermAirGlove: A pneumatic glove for thermal perception and material identification in virtual reality," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, IEEE, Mar. 2020. DOI: 10.1109/vr46266.2020.00044.

[44] O. C. Bahceci, A. Pena-Rios, V. Gupta, A. Conway, and G. Owusu, "Work-in-progress-using immersive virtual reality in field service telecom engineers training," in *2021 7th International Conference of the Immersive Learning Research Network*, IEEE, May 2021. DOI: 10.23919/ilrn52045.2021.9459243.

[45] Z. Li, J. Wang, Z. Yan, X. Wang, and M. S. Anwar, "An interactive virtual training system for assembly and disassembly based on precedence constraints," in *Advances in Computer Graphics*, Springer International Publishing, 2019, pp. 81–93. DOI: 10.1007/978-3-030-22514-8_7.

[46] F. Pece, J. Tompkin, H. Pfister, J. Kautz, and C. Theobalt, "Device effect on panoramic video+context tasks," vol. 2014-November, 2014. DOI: 10.1145/2668904.2668943.

[47] J. Harman, R. Brown, and D. Johnson, "Improved memory elicitation in virtual reality: New experimental results and insights," in *Human-Computer Interaction - INTERACT 2017*, Springer International Publishing, 2017, pp. 128–146. DOI: 10.1007/978-3-319-67684-5_9.

[48] B. Laugwitz, T. Held, and M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2008, pp. 63–76. DOI: 10.1007/978-3-540-89350-9_6.

[49] L. Schölkopf, M. Lorenz, M. Stamer, *et al.*, "Haptic feedback is more important than VR experience for the user experience assessment of in-car human machine interfaces," *Procedia CIRP*, vol. 100, pp. 601–606, 2021. DOI: 10.1016/j.procir.2021.05.130.

[50] I. Pettersson, M. Karlsson, and F. T. Ghiurau, "Virtually the same experience?" In *Proceedings of the 2019 on Designing Interactive Systems Conference*, ACM, Jun. 2019. DOI: 10.1145/3322276.3322288.

[51] J. R. Lewis, "IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use," *International Journal of Human-Computer Interaction*, vol. 7, no. 1, pp. 57–78, Jan. 1995. DOI: 10.1080/10447319509526110.

[52] T. Schubert, F. Friedmann, and H. Regenbrecht, "The experience of presence: Factor analytic insights," *Presence: Teleoperators and Virtual Environments*, vol. 10, no. 3, pp. 266–281, Jun. 2001. DOI: 10.1162/105474601300343603.

[53] D. Pinto, B. Peixoto, A. Krassmann, M. Melo, L. Cabral, and M. Bessa, "Virtual reality in education: Learning a foreign language," in *Advances in Intelligent Systems and Computing*, Springer International

Publishing, 2019, pp. 589–597. DOI: 10.1007/978-3-030-16187-3_57.

[54] J. R. Lewis, "AN AFTER-SCENARIO QUESTIONNAIRE FOR USABILITY STUDIES," *ACM SIGCHI Bulletin*, vol. 23, no. 4, p. 79, Oct. 1991. DOI: 10.1145/126729.1056077.

[55] C. Jennett, A. L. Cox, P. Cairns, *et al.*, "Measuring and defining the experience of immersion in games," *International Journal of Human-Computer Studies*, vol. 66, no. 9, pp. 641–661, Sep. 2008. DOI: 10.1016/j.ijhcs.2008.04.004.

[56] H. Liang, J. Chang, S. Deng, C. Chen, R. Tong, and J. J. Zhang, "Exploitation of multiplayer interaction and development of virtual puppetry storytelling using gesture control and stereoscopic devices," *Computer Animation and Virtual Worlds*, vol. 28, no. 5, e1727, Aug. 2016. DOI: 10.1002/cav.1727.

[57] J. Lessiter, J. Freeman, E. Keogh, and J. Davidoff, "A cross-media presence questionnaire: The ITC-sense of presence inventory," *Presence: Teleoperators and Virtual Environments*, vol. 10, no. 3, pp. 282–297, Jun. 2001. DOI: 10.1162/105474601300343612.

[58] R. M. Ryan, C. S. Rigby, and A. Przybylski, "The motivational pull of video games: A self-determination theory approach," *Motivation and Emotion*, vol. 30, no. 4, pp. 344–360, Nov. 2006. DOI: 10.1007/s11031-006-9051-8.

[59] A. Perrin, T. Ebrahimi, S. Zadtootaghaj, S. Schmidt, and S. Müller, "Towards the need satisfaction in gaming: A comparison of different gaming platforms," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–3.

[60] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-/assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994. DOI: 10.1016/0005-7916(94)90063-9.

[61] T. Marques, M. Vairinhos, and P. Almeida, "How vr 360º impacts the immersion of the viewer of suspense av content," in *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*, ser. TVX '19, Salford (Manchester), United Kingdom: Association for Computing Machinery, 2019, pp. 239–246. DOI: 10.1145/3317697.3325120.

[62] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *The International Journal of Aviation Psychology*, vol. 3, no. 3, pp. 203–220, Jul. 1993. DOI: 10.1207/s15327108ijap0303_3.

[63] F. Weidner, A. Hoesch, S. Poeschl, and W. Broll, "Comparing vr and non-vr driving simulations: An experimental user study," 2017, pp. 281–282. DOI: 10.1109/VR.2017.7892286.

[64] M. Lombard, T. Bolmarcich, and L. Weinstein, "Measuring presence: The temple presence inventory," Jan. 2009.

[65] C. Bishop, A. Esteves, and I. McGregor, "Head-mounted displays as opera glasses: Using mixed-reality to deliver an egalitarian user experience during live events," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017*, ACM Press, 2017. DOI: 10.1145/3136755.3136781.

[66] D. Markland and L. Hardy, "On the factorial and construct validity of the intrinsic motivation inventory: Conceptual and operational

concerns," *Research Quarterly for Exercise and Sport*, vol. 68, no. 1, pp. 20–32, Mar. 1997. DOI: 10.1080/02701367.1997.10608863.

[67] H. K. Kim, J. Park, Y. Choi, and M. Choe, "Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment," *Applied Ergonomics*, vol. 69, pp. 66–73, May 2018. DOI: 10.1016/j.apergo.2017.12.016.

[68] H. L. O'Brien and E. G. Toms, "The development and evaluation of a survey to measure user engagement," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 1, pp. 50–69, Oct. 2009. DOI: 10.1002/asi.21229.

[69] M. Sanaei, M. Machacek, J. C. Eubanks, P. Wu, J. Oliver, and S. B. Gilbert, "The effect of training communication medium on the social constructs co-presence, engagement, rapport, and trust," in *28th ACM Symposium on Virtual Reality Software and Technology*, ACM, Nov. 2022. DOI: 10.1145/3562939.3565686.

[70] A. E. Franklin, P. Burns, and C. S. Lee, "Psychometric testing on the NLN student satisfaction and self-confidence in learning, simulation design scale, and educational practices questionnaire using a sample of pre-licensure novice nurses," *Nurse Education Today*, vol. 34, no. 10, pp. 1298–1304, Oct. 2014. DOI: 10.1016/j.nedt.2014.06.011.

[71] T. Hoang, S. Greuter, and S. Taylor, "An evaluation of virtual reality maintenance training for industrial hydraulic machines," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, IEEE, Mar. 2022. DOI: 10.1109/vr51125.2022.00077.

[72] M. Lanier, T. Waddell, M. Elson, D. Tamul, J. Ivory, and A. Przybylski, "Virtual reality check: Statistical power, reported results, and

the validity of research on the psychology of virtual reality and immersive environments," *Computers in Human Behavior*, vol. 100, Jun. 2019. DOI: 10.1016/j.chb.2019.06.015.

[73] P. Ugwitz, A. Šašinková, Č. Šašinka, Z. Stachoň, and V. Juřík, "Toggle toolkit: A tool for conducting experiments in unity virtual environments," en, *Behavior Research Methods*, Jan. 2021.

[74] J. Grübel, R. Weibel, M. H. Jiang, C. Hölscher, D. A. Hackman, and V. R. Schinazi, "Eve: A framework for experiments in virtual environments," in *Spatial Cognition X*, Springer, 2016, pp. 159–176.

[75] M. Vasser, M. Kängsepp, M. Magomedkerimov, *et al.*, "VREX: An open-source toolbox for creating 3D virtual reality experiments," *BMC Psychology*, vol. 5, no. 1, p. 4, Feb. 2017.

[76] M. Wolfel, D. Hepperle, C. F. Purps, J. Deuchler, and W. Hettmann, "Entering a new dimension in virtual reality research: An overview of existing toolkits, their features and challenges," in *2021 International Conference on Cyberworlds (CW)*, IEEE, Sep. 2021. DOI: 10.1109/cw52790.2021.00038.

[77] N. Takahashi, T. Inamura, Y. Mizuchi, and Y. Choi, "Evaluation of the difference of human behavior between VR and real environments in searching and manipulating objects in a domestic environment," in *2021 30th IEEE International Conference on Robot Human Interactive Communication*, IEEE, Aug. 2021. DOI: 10.1109/ro-man50785.2021.9515393.

[78] C. Zavlanou and A. Lanitis, "Product packaging evaluation through the eyes of elderly people: Personas vs. aging suit vs. virtual reality aging simulation," in *Human Systems Engineering and Design*,

Springer International Publishing, Oct. 2018, pp. 567–572. DOI: 10.1007/978-3-030-02053-8_86.

[79] H. Han, A. Lu, and U. Wells, "Under the movement of head: Evaluating visual attention in immersive virtual reality environment," in *2017 International Conference on Virtual Reality and Visualization (ICVRV)*, 2017, pp. 294–295.

[80] E. Ebrahimi, S. V. Babu, C. C. Pagano, and S. Jörg, "An empirical evaluation of visuo-haptic feedback on physical reaching behaviors during 3d interaction in real and immersive virtual environments," *ACM Trans. Appl. Percept.*, vol. 13, no. 4, Jul. 2016. DOI: 10.1145/2947617.

[81] J. Mathur, S. R. Miller, T. W. Simpson, and N. A. Meisel, "Identifying the effects of immersion on design for additive manufacturing evaluation of designs of varying manufacturability," in *Volume 5: 27th Design for Manufacturing and the Life Cycle Conference (DFMLC)*, American Society of Mechanical Engineers, Aug. 2022. DOI: 10.1115/detc2022-90063.

[82] F. Mathis, K. Vaniea, and M. Khamis, "RepliCueAuth: Validating the use of a lab-based virtual reality setup for evaluating authentication systems," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ACM, May 2021. DOI: 10.1145/3411764.3445478.

[83] C. Ma and T. Han, "Combining virtual reality (vr) technology with physical models – a new way for human-vehicle interaction simulation and usability evaluation," in *HCI in Mobility, Transport, and Automotive Systems*, H. Krömker, Ed., Cham: Springer International Publishing, 2019, pp. 145–160.

[84] J. Y. Chew, K. Okayama, T. Okuma, M. Kawamoto, H. Onda, and N. Kato, "Development of a virtual environment to realize human-machine interaction of forklift operation," in *2019 7th International Conference on Robot Intelligence Technology and Applications (RiTA)*, 2019, pp. 112–118.

[85] S. Verwulgen, S. V. Goethem, G. Cornelis, J. Verlinden, and T. Coppens, "Appreciation of proportion in architecture: A comparison between facades primed in virtual reality and on paper," in *Advances in Human Factors in Wearable Technologies and Game Design*, Springer International Publishing, Jun. 2019, pp. 305–314. DOI: 10.1007/978-3-030-20476-1_31.

[86] D. Liao, W. Zhang, G. Liang, *et al.*, "Arousal evaluation of vr affective scenes based on hr and sam," in *2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)*, vol. 1, 2019, pp. 1–4.

[87] J. K. Ostrander, C. S. Tucker, T. W. Simpson, and N. A. Meisel, "Evaluating the effectiveness of virtual reality as an interactive educational resource for additive manufacturing," in *Volume 3: 20th International Conference on Advanced Vehicle Technologies 15th International Conference on Design Education*, American Society of Mechanical Engineers, Aug. 2018. DOI: 10.1115/detc2018-86036.

[88] T. Keller, E. Brucker-Kley, and C. Wyder, "Virtual reality and its impact on learning success," English, in *Proceedings of the 16th International Conference Mobile Learning 2020, ML 2020*, 2020, pp. 78–86.

[89] J. Guo, D. Weng, H. Fang, *et al.*, "Exploring the differences of visual discomfort caused by long-term immersion between virtual environments and physical environments," in *2020 IEEE Conference*

on Virtual Reality and 3D User Interfaces (VR), IEEE, Mar. 2020. DOI: 10.1109/vr46266.2020.00065.

[90] F. Diederichs, F. Niehaus, and L. Hees, "Guerilla evaluation of truck HMI with VR," in *Virtual, Augmented and Mixed Reality. Design and Interaction*, Springer International Publishing, 2020, pp. 3–17. DOI: 10.1007/978-3-030-49695-1_1.

[91] C. Vazquez, L. Xia, T. Aikawa, and P. Maes, "Words in motion: Kinesthetic language learning in virtual reality," in *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, IEEE, Jul. 2018. DOI: 10.1109/icalt.2018.00069.

[92] P. Agethen, M. Link, F. Gaisbauer, T. Pfeiffer, and E. Rukzio, "Counterbalancing virtual reality induced temporal disparities of human locomotion for the manufacturing industry," in *Proceedings of the 11th Annual International Conference on Motion, Interaction, and Games*, ACM, Nov. 2018. DOI: 10.1145/3274247.3274517.

[93] A. Franzluebbers and K. Johnsen, "Performance benefits of high-fidelity passive haptic feedback in virtual reality training," in *Proceedings of the Symposium on Spatial User Interaction - SUI '18*, ACM Press, 2018. DOI: 10.1145/3267782.3267790.

[94] S. Bialkova and D. Ettema, "Cycling renaissance: The vr potential in exploring static and moving environment elements," in *2019 IEEE 5th Workshop on Everyday Virtual Reality (WEVR)*, 2019, pp. 1–6.

[95] D. J. R. Christensen and M. B. Holte, "The impact of virtual reality training on patient-therapist interaction," in *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, Springer International Publishing, 2018, pp. 127–138. DOI: 10.1007/978-3-319-76908-0_13.

[96] S. Safikhani, M. Holly, A. Kainz, and J. Pirker, "The influence of in-VR questionnaire design on the user experience," in *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, ACM, Dec. 2021. DOI: 10.1145/3489849.3489884.

[97] J. A. Wagner Filho, M. F. Rey, C. M. D. S. Freitas, and L. Nedel, "Immersive visualization of abstract information: An evaluation on dimensionally-reduced data scatterplots," in *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2018, pp. 483–490.

[98] M. Petrykowski, P. Berger, P. Hennig, and C. Meinel, "Digital collaboration with a whiteboard in virtual reality," in *Proceedings of the Future Technologies Conference (FTC) 2018*, Springer International Publishing, Oct. 2018, pp. 962–981. DOI: 10.1007/978-3-030-02686-8_72.

[99] S. Kratz and F. Rabelo Ferriera, "Immersed remotely: Evaluating the use of head mounted devices for remote collaboration in robotic telepresence," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 638–645.

[100] B. J. H. Andersen, A. T. A. Davis, G. Weber, and B. C. Wunsche, "Immersion or diversion: Does virtual reality make data visualisation more effective?" In *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, IEEE, Jan. 2019. DOI: 10.23919/elinfocom.2019.8706403.

[101] A. Franzluebbers, C. Li, A. Paterson, and K. Johnsen, "Virtual reality point cloud annotation," in *Proceedings of the 2022 ACM Symposium on Spatial User Interaction*, ACM, Dec. 2022. DOI: 10.1145/3565970.3567696.

[102] X. Zhang, W. He, and S. Wang, "Manual preliminary coarse alignment of 3d point clouds in virtual reality," in *Communications in Computer and Information Science*, Springer International Publishing, 2021, pp. 424–432. DOI: 10.1007/978-3-030-90176-9_55.

[103] J. Hombeck, M. Meuschke, L. Zyla, *et al.*, "Evaluating perceptional tasks for medicine: A comparative user study between a virtual reality and a desktop application," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, IEEE, Mar. 2022. DOI: 10.1109/vr51125.2022.00071.

[104] C. Keighrey, R. Flynn, S. Murray, and N. Murray, "A physiology-based QoE comparison of interactive augmented reality, virtual reality and tablet-based applications," *IEEE Transactions on Multimedia*, pp. 1–1, 2020. DOI: 10.1109/tmm.2020.2982046.

[105] D. Watson, G. Fitzmaurice, and J. Matejka, "How tall is that bar chart? virtual reality, distance compression and visualizations," en, 2021. DOI: 10.20380/GI2021.29.

[106] T. Nishimura, K. Hirai, and T. Horiuchi, "Color perception comparison of scene images between head-mounted display and desktop display," *Proceedings of the International Display Workshops*, p. 1148, Nov. 2019. DOI: 10.36463/idw.2019.1148.

[107] M. Nebeling, S. Rajaram, L. Wu, Y. Cheng, and J. Herskovitz, "XRStudio: A virtual production and live streaming system for immersive instructional experiences," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ACM, May 2021. DOI: 10.1145/3411764.3445323.

[108] R. Fujii, H. Hirose, S. Aoyagi, and M. Yamamoto, "On-demand lectures that enable students to feel the sense of a classroom with students who learn together," in *Human Interface and the Management of Information. Information Presentation and Visualization*, Springer International Publishing, 2021, pp. 268–282. DOI: 10.1007/978-3-030-78321-1_21.

[109] G. Makransky, T. S. Terkildsen, and R. E. Mayer, "Adding immersive virtual reality to a science lab simulation causes more presence but less learning," *Learning and Instruction*, vol. 60, pp. 225–236, Apr. 2019. DOI: 10.1016/j.learninstruc.2017.12.007.

[110] J. Thorn, R. Pizarro, B. Spanlang, P. Bermell-Garcia, and M. Gonzalez-Franco, "Assessing 3d scan quality through paired-comparisons psychophysics," in *Proceedings of the 2016 ACM on Multimedia Conference - MM '16*, ACM Press, 2016. DOI: 10.1145/2964284.2967200.

[111] N. Horvat, S. Škec, T. Martinec, F. Lukacevic, and M. Perišic, "Comparing virtual reality and desktop interface for reviewing 3d cad models," vol. 2019-August, 2019, pp. 1923–1932. DOI: 10.1017/dsi.2019.198.

[112] Z. Qadir, E. Chowdhury, L. Ghosh, and A. Konar, "Quantitative analysis of cognitive load test while driving in a VR vs non-VR environment," in *Lecture Notes in Computer Science*, Springer International Publishing, 2019, pp. 481–489. DOI: 10.1007/978-3-030-34872-4_53.

[113] V. Colombo, G. Bocca, M. Mondellini, M. Sacco, and A. Aliverti, "Evaluating the effects of virtual reality on perceived effort during cycling: Preliminary results on healthy young adults," in *2022 IEEE*

International Symposium on Medical Measurements and Applications (MeMeA), IEEE, Jun. 2022. DOI: 10.1109/memea54994.2022.9856467.

8 Exploring Ecological Validity: A Comparative Study of the Mere Exposure Effect on Screens and in Immersive Virtual Reality

Publication Note: This chapter is based on the following published work. The content of the chapter is identical to the published article, with only the formatting and numbering being adapted for this dissertation.

Hepperle, D., Wölfel, M. (2025).
In: Bebis, G., et al. Advances in Visual Computing. ISVC 2024.
Lecture Notes in Computer Science, vol 15047. Springer, Cham.
https://doi.org/10.1007/978-3-031-77389-1_17

Abstract

This study examines the ecological validity of the mere exposure effect across different media, comparing the effects on computer screens and in immersive virtual reality (IVR). Both experiments were conducted remotely, maintaining a similar design to that of the original study but differing in terms of the technological platforms utilized. The results of our study indicate that increased exposure consistently enhances liking in both experimental settings, thereby reinforcing the robustness of the mere exposure effect across different modalities. Furthermore, our results

align closely with those reported by Mrkva and van Boven from 2020 which we replicated and the meta-analyses by Montoya et al. from 2017. By demonstrating that findings are statistically similar between media, this research contributes to the understanding of ecological validity. It also suggests that IVR can effectively replicate findings from traditional experimental setups. Further investigation into factors such as graphics quality, interaction complexity, and the influence of the uncanny valley is necessary to fully leverage the potentials and pitfalls of IVR in behavioral research.

Keywords Mere Exposure – Immersive Virtual Reality – Computer Screens.

8.1 Introduction

The introduction of novel technologies raises novel questions and opportunities. Digital innovations have been rapidly adopted and became indispensable to the daily practices of most social science researchers. Traditional methods that relied on paper forms, static images, and face-to-face interviews are increasingly supplemented by dynamic and interactive formats such as online forms, videos, and video conferences [1]. This shift is driven by several "quality of life" factors: online forms are simpler to maintain, require less time to process compared to their paper counterparts, can be shared and published more easily, and are accessible to a broader pool of participants. Beyond these practical benefits, digital technologies enhance the documentation of research activities, increasing scientific rigor and fostering replications [2], [3]. Interactive elements in stimuli exposure, such as videos and virtual simulations, provide researchers with better control over the experimental environment. These tools enable precise manipulation of variables and consistent presentation of stimuli, leading to more accurate and reliable data [4].

Using virtual environments, on 2D computer screens, to conduct social science and psychological experiments had been used for decades. However, immersive technology such as head-mounted displays (HMD) that offer a higher degree of freedom, immersion, and presence (sense of being there) is just starting to be discovered as a research tool in these domains. The use of virtual environments can reduce costs associated with traditional laboratory experiments, as participants no longer need to physically travel to a research facility and the research environment can be altered at the touch of a button. For example, immersive virtual

reality (IVR) has the potential to enable remote participation, to reduce the need for physical resources and dedicated lab space, while facilitating reproducibility and replicability [5]. Despite the potential advantages of IVR, the discrepancies between the stimuli and response relationships in IVR and the physical environment remain unclear and require further investigation to ensure the results can be generalized beyond the virtual context [5], [6].

The objective of this study is to address the aforementioned gap in knowledge by examining whether a robust effect in the field of social science can be confirmed in a three-dimensional immersive environment. If the effect can be confirmed in IVR, it will be demonstrated that this effect fulfills the criterion of ecological validity [7]. To perform our experiment we decided to investigate the mere exposure effect, as it has demonstrated robust results in traditional media [8], [9]. The *mere exposure effect*, initially proposed by social psychologist Robert Zajonc in the 1960s, is a psychological phenomenon whereby individuals tend to develop a preference for stimuli with which they are familiar. This effect indicates that repeated exposure to a specific stimulus, whether an object, another person, or an idea, increases the individual's positive affect toward it. The effect is fundamental to psychological research because it highlights basic psychological principles of learning and familiarity. By investigating this effect within IVR environments, we can ascertain whether the underlying psychological processes remain consistent or if they differ in a virtual context. Understanding these differences, if any, will provide crucial insights into the cognitive aspects of IVR. Thereby informing the broader implications of using IVR as a tool to perform social science or psychology research. This will support the potential value of IVR as a research tool,

demonstrating its capability to replicate established psychological phenomena and thereby enhancing its credibility and utility in experimental research.

8.2 Related Work

This section presents a review of the two primary themes of our investigation: the current research on the use of IVR as a tool for conducting research and the mere exposure effect. It is crucial to differentiate between studies that *aim to gain insight into IVR*, such as examining how individuals respond to the virtual environment or to avatars, and those that *utilize IVR as a research instrument* to gain an understanding of how individuals would behave in the real-world. An illustrative example of the latter is investigating the bystander effect using embodied agents in a virtual setting [10].

8.2.1 Immersive Virtual Reality as a Research Tool

IVR is proving to be a valuable and flexible tool across various scientific disciplines, offering interactive and realistic environments that provide new opportunities for controlled experimentation, data collection, and analysis. Unlike traditional research methodologies, IVR allows researchers to simulate complex scenarios, manipulate variables with precision, and observe participant behaviors in real-time within a fully immersive context. The integration of IVR into research paradigms not only enhances the ecological validity of experimental setups but also provides new avenues for participant engagement and data richness.

While there are theoretical advantages to creating one's own IVR research environment, it can also present a significant challenge. The configuration necessitates not only considerable technical proficiency but also substantial resources. From a technical standpoint, researchers must integrate a variety of hardware components, including HMDs and different input devices. Regarding the visual aspects, developing the software to create interactive and responsive virtual characters and environments demands specialized knowledge in programming and 3D modeling. The complexity of these tasks can present a significant obstacle to researchers who are new to the field or lack access to extensive technical support. To address these challenges, specific toolkits have been developed to facilitate the creation of IVR research scenarios. These toolkits provide pre-built components and frameworks that simplify the development process, allowing researchers to focus on their experimental design rather than the technical intricacies of IVR implementation. An overview of such toolkits can be found in [11].

In form of a scoping review Hepperle and Wölfel [6] list 56 research works, that compare IVR either to the real-world or to 2D monitors. Within this work, specific differences such as difficulties in distance estimation and depth judgments [12], higher cognitive load [13], reaction time [14], and anxiety levels [15] are mentioned. Yet, most people compare their results with their own experiments and not with robust, and proven effects. So it is not necessarily given, that these effects are replicable and are not affected by bias.

The Trier Stress Test, known for its robust effects, has been shown to induce similar responses in IVR compared to in vivo scenarios. However, only a few studies have directly compared these responses in their own

experiments. Among studies that explicitly compared in vivo and IVR conditions, researchers found that IVR induces similar stress responses [16], [17] and a comparable increase in heart rate [17], [18] in both scenarios.

8.2.2 Mere Exposure Effect

The *mere exposure effect* refers to the observation that an individual's liking for a particular stimulus tends to increase with repeated exposure to that stimulus. After Zanjonc's initial investigation [19], a multitude of studies have been conducted, further elucidating the mere exposure effect and providing additional insight. With more than 3000 citations, Bornstein's meta-analysis [8] is among the most influential. Through the investigation of 208 independent experiments, Bornstein determined a combined effect size of .260 and a fail-safe N of 33,047 for all stimuli types, with the largest effect sizes for meaningful words (.486), followed by polygons (.413) and photographs (.367). The effect sizes are larger for studies employing a heterogeneous design (.301), which incorporates multiple stimuli than for homogeneous design (-.020) which incorporates a single stimuli type. Approximately 30 years later, Montoya et al. [9] conducted a reinvestigation of the mere exposure effect by examining 81 articles. As a primary finding, the authors report that the relationship between exposure and liking is not linear, but rather exhibits an inverted-U shape for visual but not auditory stimuli, exposure durations less than 10 seconds and longer than one minute, both presentation types (homogeneous and heterogeneous) and ratings that were taken after stimuli were presented.

8.3 Methodology

In our study, we aim to investigate the extent to which findings and conclusions from experiments conducted in IVR can be generalized to real-world scenarios. The replication crisis in psychology highlights the difficulty of replicating many published results due to various factors such as lack of documentation and small effect sizes with low participant numbers [20], [21]. Consequently, it is essential to utilize an established, validated, and robust effect proven in the real-world as a basis for comparison, ensuring that results obtained in an IVR environment are reliable and applicable to real-world scenarios. To address the challenges posed by the replication crisis and to ensure robust findings, we chose the mere exposure effect for its well-documented and replicable nature. To avoid any potential discrepancies between our IVR study and the original one, we have decided to investigate two different settings. These settings will allow us to determine whether any differences can be attributed to the way our studies are conducted or to any misinterpretation of the original findings:

- a direct replication of the mere exposure effect using similar tools such as Amazon Mechanical Turk (MTurk) and modalities (2D monitor) with a comparable population.

- a replication in IVR using the same procedures and stimuli in an immersive VR setting.

This dual approach aims to fully demonstrate the effect-response relationship conducted by the same research team. While the first experiment is a *direct replication,* which is also proposed as a countermeasure to the replication crisis [20], different theoretical concepts describe the second

approach. One is understood as *conceptual replication* as defined by [22] and the other as more than a *generalization* [23]. Neither term fully describes the rather unusual situation we face when conducting a study within another modality[1]. Also note, that previous research suggests that effects measured within the IVR may be more pronounced compared to real-world settings [15], [24] which adds another level of difficulty in replicating the results.

For direct replication, an effect that has the potential to be replicated must meet the following conditions: First, it must have been successfully replicated before. Second, there should be meta-studies examining the effect. Third, the effect in question should be sufficiently strong to yield meaningful results. Fourth, the priming study must be sufficiently well documented to provide a solid basis for replication. Fifth, the experimental design must be feasible in an IVR environment to ensure that the study can be effectively conducted using this technology. Finally, the stimuli used in the experiment must be available to facilitate accurate replication and comparison.

The study by Mrkva and van Boven 2020 [25] met these necessary requirements for replication. As a primary objective, Mrkva and van Boven aimed to test the salience theory of the *mere exposure effect*, but—as the first part of their study—they conducted a replication of mere exposure with the most common stimuli used in previous work. Given the focus on the mere exposure effect itself, we will concentrate on this specific part of their study in our work.

[1] For clarity throughout this paper, we refer to the MTurk experiment as the direct replication and the immersive VR experiment as the IVR replication.

The initial phase of the study involved replicating the effect in a 2D monitor setting using MTurk and a questionnaire tool. This was followed by the conduct of the study in an IVR environment. This methodology was designed to replicate the effect as precisely as possible in a 2D environment and subsequently assess the differences when the exposure occurs in an immersive VR environment.

8.4 Study Procedure

To facilitate a convincingly close replication attempt, we have oriented ourselves for this work on the replication recipe proposed by Brandt et al. [26]. The recipe has been filled out and is preregistered on OSF accordingly[2]. Additionally, analysis files, datasets, and the questionnaire including all stimuli can be found on OSF as well[3]. The supplementary material of the original study provided by Mrkva and van Boven can be found on osf.io as well[4].

8.4.1 Study Procedure by Mrkva and van Boven

Given the supplemental material as well as the documentation within the original paper, it wasn't possible to derive the correct stimuli order and the number of exposures per stimulus. The provided materials lacked specific details that were crucial for accurately replicating the experimental conditions of the original study. This presented a significant challenge,

[2] https://osf.io/ny3ak?view_only=56e79aa3e5e34b368656340a33fddee2

[3] https://osf.io/crakb/?view_only=bd7170826ab448dd8d7c8ebf92db940e

[4] https://osf.io/q6a3b/?view_only=dca757b6691b47ad935191882aa3de09

as precise replication is essential for validating the findings and ensuring the reliability of the results.

Contacting the authors via ResearchGate, they generously provided additional documentation that contained the essential details necessary to replicate the stimuli order precisely as it was in the original study.

In addition to the supplementary documents, the authors also shared the raw Qualtrics questionnaire through the Qualtrics platform. This was particularly helpful as it allowed us to take a deeper look into the survey process and understand the exact flow and structure of the questions. Having access to the raw questionnaire enabled us to ensure that all aspects of the experimental procedure, including the presentation and timing of stimuli, matched those of the original study. With the additional information and resources provided by the authors, we were able to reconstruct the exact setup of the original experiment as follows[5]:

1. A consent form was presented, and demographic data were collected.

2. The study was carried out under the pretense of *investigating memory*. Specifically, participants were told that the test was investigating whether people remember words after being exposed to them only a few times and whether more exposures are required to remember artworks, shapes, and symbols.

3. After clicking the consent/start button, participants viewed slideshows consisting of three different sets of stimuli: Chinese characters, Turkish words, and segments of an abstract art painting.

[5] To remain as close as possible to the original study setup, we posed the exact same questions, despite not using all items in the evaluation.

The stimuli are shown in Fig. 8.1. These exact stimuli were used because they are the most common in previous mere exposure research [9]. The exposure was manipulated by presenting some stimuli more frequently than others.

4. Each participant viewed a slideshow consisting of four stimuli from the same set of stimuli.

5. Two stimuli were presented nine times and two were presented three times. In total 24 slides per participant. Slide order was pseudo-randomly determined.

6. Stimuli were presented for 1.0 seconds each. A fixation cross was shown for 1.0 seconds between each stimulus.

7. The following questionnaire items were presented after the first set of slides[6]:

    a) How much do you like this image? [7 point Likert scale, dislike – like]

    b) For each image, indicate the extent to which you think it means something good or bad. [7 point Likert scale, very bad – bad].

    c) How emotionally intense is each image? [9 point Likert scale, not at all intense – extremely intense]

    d) How intense was your emotional reaction to each image? [9 point Likert scale, not at all intense – extremely intense]

[6] Since Study 2 (IVR) was conducted with a German population, the questions were translated accordingly.

8. After these four questions were answered, participants started the next set of stimuli by clicking the "next" button. These steps were repeated until all three stimuli sets were shown to the participants.

9. At the end of the experiment, participants were given a memory test. Within the memory test, all 12 stimuli that participants saw across the three slideshows were presented along with 18 foil stimuli from the same stimulus sets in the original study. Studies 1 and 2 had fewer foil stimuli because they weren't available online. Participants were asked, "Which of the below images, words, and characters did you see in the slideshows earlier?" Participants could select multiple images.

10. Finally, participants answered questions assessing their awareness of the research question. "Which of the following do you think the researchers were studying?" [multiple choice]

    a) Whether people are better at remembering words and word-like letter strings compared to shapes and symbols.

    b) Whether people are better at remembering pictures presented a couple of times compared to pictures presented more times.

    c) Whether people like words with familiar letters more than words with unfamiliar (e.g., Chinese) characters.

    d) Whether people remember pictures or words presented only one or a few times, but require more presentations to remember symbols and shapes.

e) Whether people like pictures presented more times than pictures presented fewer times.

f) Whether features of symbols and shapes (e.g., sharp edges) influence whether it is perceived to be good or bad.

11. The following questionnaire items "What do you think was the research question that the experimenters were testing?" and "Why were some words, characters, or shapes presented more times than others?" [free text answer].

12. Participants were asked if they speak Chinese / Mandarin or Turkish or if the comprehended any of the words / characters and images seen as single choice.

13. Participants were asked if they had previously seen any of the stimuli, as the exposure effect may be more pronounced for novel stimuli [27].

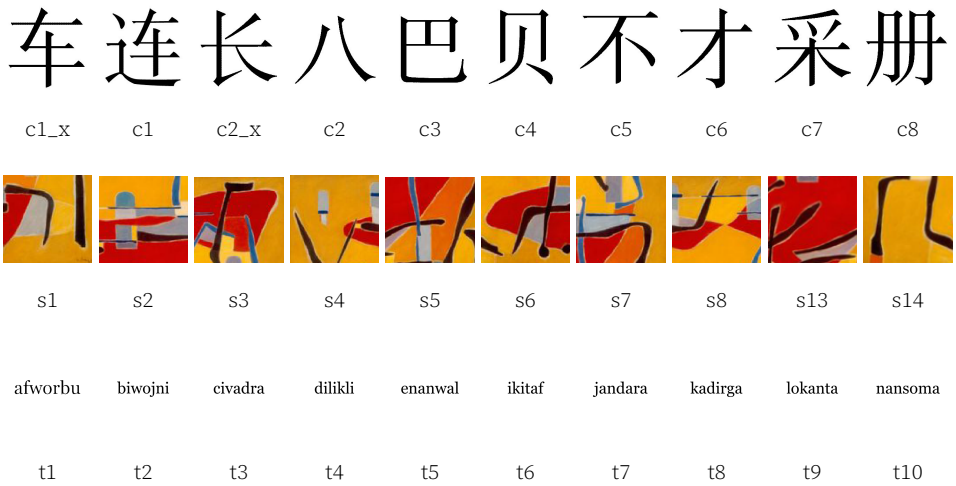

Figure 8.1: Stimuli used in all studies. The text beneath the images indicates filenames as they are listed on OSF. Note: Not all Turkish words are authentic; some are pseudoword.

8.5 Differences Between the Experiments

Despite the different setups, the procedure, as well as the stimuli used within the experiments conducted were the same. In the following, you find an overview of all three studies:

Table 8.1: Differences between the Original Study, Direct Replication (MTurk), and Replication in IVR

Step	Original Study	Direct Replication (MTurk)	Replication in IVR
Platform	Amazon MTurk, Qualtrics	Amazon MTurk, SoSci Survey	Self-developed social VR platform, VR Questionnaire Toolkit
Environment	2D-Monitor, Qualtrics	2D-Monitor, SoSciSurvey	Immersive VR Meta Quest 2
Memory Test	18 foil stimuli	No foil stimuli due to availability	No foil stimuli due to availability
Interaction	Mouse and keyboard	Mouse and keyboard	Meta Quest 2 controller
Onboarding	Online text-based instructions	Online text-based instructions	Onboarding session in IVR
Mean Age	36	32	26
Language	English	English	German
Population	50m; 50f	70m; 49f (5 removed for missing data)	39m; 24f
Compensation	1.25 USD	1.25 USD	10 Euro (Voucher)

From previous studies, we anticipated challenges in recruiting a sufficient number of participants, particularly for the IVR study. Kelly et al. [28] show that offering incentives increases participants' motivation to participate compared to no incentive at all with higher incentives further enhancing this motivation. For the Amazon Mechanical Turk replication, our pretest demonstrated rapid engagement, assuring us that we would gain enough participants. We maintained the same compensation of $1.25

as in the original study to establish a baseline for comparison to rule out the experimenter as an influencing factor in the IVR study. Despite offering a higher incentive for the IVR study, we were still unable to recruit the same number of participants, which may be due to the fact that participating in an IVR experiment requires more effort than taking part from home on a regular PC. Also, the stimuli used in the current experiment are neither English nor German, and therefore we preferred to conduct the study in Germany in German rather than in a non-native language. This approach aligns with previous findings that the mere exposure effect is consistent across cultures, as noted by Ishii [29], and it helped us maintain the study's focus on exposure frequency without introducing language comprehension as a confounding factor.

8.5.1 Direct Replication using MTurk (Experiment 1)

For the direct replication using MTurk, we first had to create stimuli exposure pages included in the survey. Contrary to the authors of the original study who used Qualtrics as a platform, we used a self-hosted SoSci Survey[7] instance.

8.5.2 Immersive VR (Experiment 2)

The immersive part of the study was carried out within a self-developed social VR platform [30]. The platform was originally developed to carry out lectures within immersive VR and was modified to also serve as a research environment. It is GDPR-compliant, self-hosted, and can manage

[7] https://www.soscisurvey.de/

up to 60 participants simultaneously using Meta Quest 2 HMDs. Participants and the experimenter are both represented by custom Ready Player Me[8], avatars and can talk to each other via speech. Upon joining the VR Campus, participants enter a virtual lobby where they are welcomed and briefed on the upcoming experiment. The experimenter conducts a live onboarding session using presentation slides to ensure that all participants understand the procedures and objectives of the study.

Following the onboarding, participants are teleported into a designated separate room. This room is specifically tailored for the study, featuring a screen that displays the stimuli and the questionnaires. At the end of the experiment, participants are teleported back to the initial area. This transition helps in maintaining the integrity of the study by clearly delineating different phases of the participant's experience. As a framework for the virtual questionnaire, the VR Questionnaire Toolkit was used [31]. This toolkit allows for the collection and management of participant responses in IVR.

8.6 Results

Exposure increased liking[9] (see Fig. 8.2 and Table 8.2). As in the original study, participants reported that they liked stimuli they were exposed to nine times ($M_{Original}$ = 0.951 | M_{MTurk} = 0.777 | M_{IVR} = 0.692) more than stimuli they were exposed to three times ($M_{Original}$ = 0.092 | M_{MTurk} = 0.351 | M_{IVR} = 0.0454) or zero times ($M_{Original}$ = –1.043 | M_{MTurk} =

[8] `https://readyplayer.me`

[9] The two items "liking" and "meaning something good" were averaged for all analyses across experiments, as done in the original study by Mrkva and Van Boven [25] and will be referred to as "liking."

–1.121 | M_{IVR} = –1.147) with $t_{MTurk}(50.25) = 3.06$, b = 0.17, p < .001 and $t_{IVR}(33.58) = 4.29$, b = 0.23, p < .001. This replicates previous research on mere exposure [8], [19].

To provide a more insightful comparison between the experiments, we also include the means from Montoya et al.'s meta-analysis [9] (M_9 = 0.801 | M_3 = 0.320 | M_0 = –1.121) in Fig. 8.2, and therefore had to z-standardize the values.

In [9], they averaged 30 mere exposure effects for 9 exposures, 49 effects for 3 exposures, 110 for 0 exposures. While the bar plot shows the expected trend that, as the number of exposures increases, liking ratings also increase, it also offers an interesting comparison between the studies. For 0 exposures, the "liking" ratings are relatively consistent across all studies. However, it is noteworthy that the original study demonstrates the largest overall difference when compared to the other experiments. For 3 exposures, it is similar. Again, the original study is the furthest away from the other three bars. For 9 exposures, the pattern is consistent. The original study remains the most distinct, showing the greatest difference from the other three studies. Carrying out a Tukey HSD Post-Hoc Test returns the following results. For Experiment 1 (MTurk), the comparison of exposure levels between 0 and 3 showed a difference of 0.34, with a 95% confidence interval (CI) of [0.19, 0.50] and p < 0.001. The comparison between exposure levels 0 and 9 revealed a difference of 0.33, with a 95% CI of [0.18, 0.49] and p < 0.001. Lastly, for the exposure level comparison between 3 and 9, the difference was –0.01, with a 95% CI of [–0.19, 0.17] and p = 0.99.

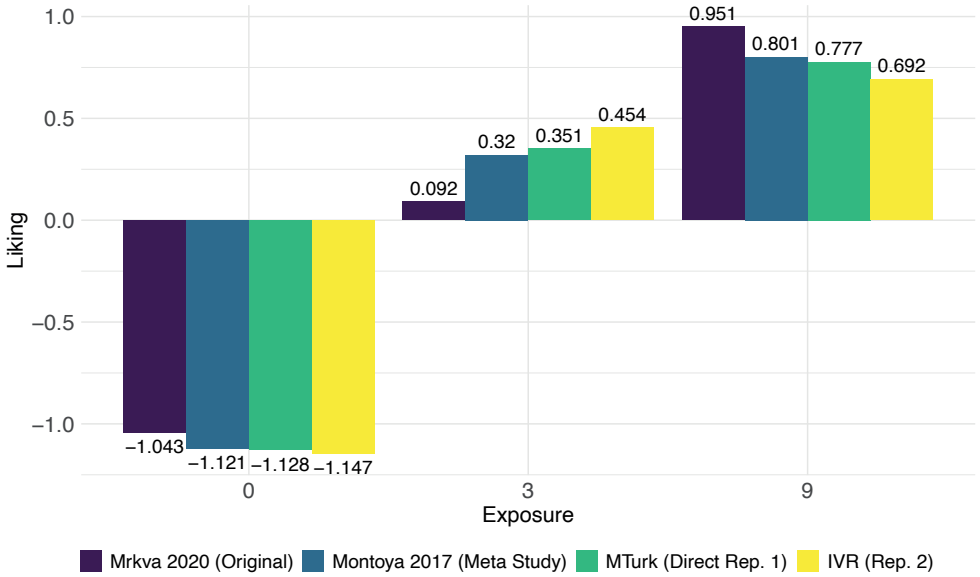

Figure 8.2: Comparison of liking ratings across different exposure levels and experiments.

Table 8.2: Fixed Effects of the Linear Mixed-Effects Model for Direct Replication (MTurk) and Replication in IVR. Contrast 1: -1 = 0 exposures, 0 = 3 exposures, 1 = 9 exposures. Contrast 2: $-1/3$ = 0, $2/3$ = 3, $-1/3$ = 9 exposures.

	Direct Replication (MTurk)				Replication in IVR			
	b	**SE(b)**	**t**	**p**	**b**	**SE(b)**	**t**	**p**
(Intercept)	1.84	0.09	20.04	<0.001	-0.05	0.11	-0.48	0.63
Exposure Contrast 1	0.17	0.06	3.06	<0.001	0.23	0.05	4.29	<0.001
Exposure Contrast 2	0.18	0.06	2.98	<0.001	0.17	0.08	2.25	<0.05

For Experiment 2 (IVR), the exposure level comparison between 0 and 3 indicated a difference of 0.39, with a 95% confidence interval (CI) of [0.20, 0.59] and $p < 0.001$. Between 0 and 9, the difference was 0.45, with a 95% CI of [0.26, 0.65] and $p < 0.001$. Lastly, the comparison between levels 3 and 9 yielded a difference of 0.06, with a 95% CI of [–0.17, 0.28] and $p = 0.82$. The decline in slope goes in line with the findings from Montoya et al. [9] in which they argue that boredom after repeated exposure might lead to a decline in positive liking, resulting in an inverted-U shape distribution for liking.

8.7 Discussion and Outlook

The presented study demonstrates the efficacy and practicality of employing IVR to induce effects analogous to those observed in a real-world context. While there is a substantial body of literature delineating both dissimilarities and similarities between the effects observed in the real-world, on a 2D screen, and in IVR, the methodologies utilized to infer these effects were frequently conducted within the same study by the same researchers, thereby limiting the generalizability of the findings [6].

The process presented in this work involved replicating a study within the same technological framework and then conducting it again in an immersive VR application. This approach reduces the possibility of human-induced error and demonstrates the robustness and generalizability of the effects across IVR, the real-world, and 2D screens. The following key findings are reported within the scope of this work:

- The replication study yields similar results to those obtained in IVR: Exposure increases liking in both the replication study and IVR (see Fig. 8.2 and Table 8.2). The findings are further supported by the results of the meta-study by Montoya et al. [9].

- The study's success in replicating the mere exposure effect in both MTurk and IVR environments demonstrates the potential for VR to induce effects similar to those observed in real-world scenarios.

- The study represents a significant advancement in establishing the validity of IVR as a research tool.

To enable others to build on our findings, we provide our datasets and calculations for both experiments presented in this study via `osf.io`, similar to the materials provided by the authors of the original study [25].

With the current state-of-the-art technology, neither photorealistic graphics nor natural interaction, locomotion, and haptics have been achieved. This is particularly affecting social interactions with other humans (depicted by avatars) as well as with embodied agents within IVR. Also, due to side effects such as the weight of HMDs and cybersickness long-duration studies cannot be executed. We chose the mere exposure effect for our study because it is minimally impacted by technical constraints. Nevertheless, other effects of course might be limited due to technical constraints, and therefore further investigation is needed. It was only possible to carry out our experiment as we were able to use self-developed software to perform IVR experiments remotely and because our institute owns 60 Meta Quest 2 headsets, which are distributed to students at the beginning of the semester. Without these resources recruiting participants would be even more challenging than for traditional experiments.

Additionally, although the IVR experiment was conducted within a self-developed immersive social VR environment, the study setup for remote execution cannot be readily used for replication without a particular technical skillset and comprehensive documentation of the setup process.

After all, with this work, we contribute to the growing understanding of how closely IVR can replicate real-world sensations to carry out valid research experiments. Nevertheless, we acknowledge the necessity for further research to gain a comprehensive understanding of the potential

applications, for instance, in the context of dyadic and multi-user experiments in which participants and confederates are represented by virtual avatars. While text-based assistants driven by large language models already might pass a Turing test, avatars still face the challenge to stay out of the uncanny-valley especially regarding natural co-gesticulation and non-verbal communication in cases where these behaviors are generated procedurally (artificial intelligence driven) rather than manually. But, as IVR technology continues to evolve, the goal is to approach ever more realistic simulations, aligning with Ivan Sutherland's vision of an 'ultimate display' that fully immerses users [32].

References

[1] J. R. Evans and A. Mathur, "The value of online surveys: A look back and a look ahead," *Internet Research*, vol. 28, no. 4, pp. 854–887, Aug. 2018.

[2] J. R. Evans and A. Mathur, "The value of online surveys," *Internet Research*, vol. 15, no. 2, pp. 195–219, Apr. 2005. DOI: 10.1108/10662240510590360.

[3] A. J. Moss, C. Rosenzweig, J. Robinson, S. N. Jaffe, and L. Litman, "Is it ethical to use mechanical turk for behavioral research? relevant data from a representative survey of mturk participants and wages," *Behavior Research Methods*, vol. 55, no. 8, 2023.

[4] X. Pan and A. F. d. C. Hamilton, "Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape," *British Journal of Psychology*, vol. 109, no. 3, 2018. DOI: 10.1111/bjop.12290.

[5] D. Hepperle, T. Dienlin, and M. Wölfel, "Reducing the human factor in virtual reality research to increase reproducibility and replicability," in *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, IEEE, Oct. 2021. DOI: 10.1109/ismar-adjunct54149.2021.00030.

[6] D. Hepperle and M. Wölfel, "Similarities and differences between immersive virtual reality, real world, and computer screens: A systematic scoping review in human behavior studies," *Multimodal Technologies and Interaction*, vol. 7, no. 6, p. 56, May 2023. DOI: 10.3390/mti7060056.

[7] E. Brunswik, *Perception and the Representative Design of Psychological Experiments*. University of California Press, Dec. 1956.

[8] R. F. Bornstein, "Exposure and affect: Overview and meta-analysis of research, 1968–1987.," *Psychological Bulletin*, vol. 106, no. 2, pp. 265–289, Sep. 1989. DOI: 10.1037/0033-2909.106.2.265.

[9] R. M. Montoya, R. S. Horton, J. L. Vevea, M. Citkowicz, and E. A. Lauber, "A re-examination of the mere exposure effect: The influence of repeated exposure on recognition, familiarity, and liking.," *Psychological Bulletin*, vol. 143, no. 5, pp. 459–498, 2017. DOI: 10.1037/bul0000085.

[10] A. Rovira, R. Southern, D. Swapp, *et al.*, "Bystander affiliation influences intervention behavior: A virtual reality study," *SAGE Open*, vol. 11, no. 3, Jul. 2021. DOI: 10.1177/21582440211040076.

[11] M. Wölfel, D. Hepperle, C. F. Purps, J. Deuchler, and W. Hettmann, "Entering a new dimension in virtual reality research: An overview of existing toolkits, their features and challenges," in *2021 International Conference on Cyberworlds (CW)*, IEEE, Sep. 2021. DOI: 10.1109/cw52790.2021.00038.

[12] E. Ebrahimi, S. V. Babu, C. C. Pagano, and S. Jörg, "An empirical evaluation of visuo-haptic feedback on physical reaching behaviors during 3d interaction in real and immersive virtual environments," *ACM Transactions on Applied Perception*, vol. 13, no. 4, 2016.

[13] G. Makransky, T. S. Terkildsen, and R. E. Mayer, "Adding immersive virtual reality to a science lab simulation causes more presence but less learning," *Learning and Instruction*, vol. 60, pp. 225–236, Apr. 2019.

[14] Y.-C. Chen and H.-W. Liang, "Reliability and validity of a virtual reality-based measurement of simple reaction time: A cross-sectional study (preprint)," 2023.

[15] D. Liao, W. Zhang, G. Liang, *et al.*, "Arousal evaluation of vr affective scenes based on hr and sam," *2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)*, vol. 1, pp. 1–4, 2019.

[16] P. Zimmer, B. Buttlar, G. Halbeisen, E. Walther, and G. Domes, "Virtually stressed? a refined virtual reality adaptation of the trier social stress test (tsst) induces robust endocrine responses," *Psychoneuroendocrinology*, vol. 101, pp. 186–192, 2019.

[17] Y. Shiban, J. Diemer, S. Brandl, R. Zack, A. Mühlberger, and S. Wüst, "Trier social stress test in vivo and in virtual reality: Dissociation of response domains," *International Journal of Psychophysiology*, vol. 110, pp. 47–55, Dec. 2016. DOI: 10.1016/j.ijpsycho.2016.10.008.

[18] M. Kotlyar, C. Donahue, P. Thuras, *et al.*, "Physiological response to a speech stressor presented in a virtual reality environment," *Psychophysiology*, vol. 45, no. 6, 2008.

[19] R. B. Zajonc, "Attitudinal effects of mere exposure.," *Journal of Personality and Social Psychology*, vol. 9, no. 2, Pt.2, pp. 1–27, 1968. DOI: 10.1037/h0025848.

[20] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, 2015.

[21] M. Lanier, T. F. Waddell, M. Elson, D. J. Tamul, J. D. Ivory, and A. Przybylski, "Virtual reality check: Statistical power, reported results, and the validity of research on the psychology of virtual reality and immersive environments," *Computers in Human Behavior*, vol. 100, pp. 70–78, Nov. 2019. DOI: 10.1016/j.chb.2019.06.015.

[22] C. S. Crandall and J. W. Sherman, "On the scientific superiority of conceptual replications for scientific progress," *Journal of Experimental Social Psychology*, vol. 66, pp. 93–99, Sep. 2016. DOI: 10.1016/j.jesp.2015.10.002.

[23] B. A. Nosek and T. M. Errington, "What is replication?" *PLOS Biology*, vol. 18, no. 3, e3000691, Mar. 2020. DOI: 10.1371/journal.pbio.3000691.

[24] D. Hepperle, C. F. Purps, J. Deuchler, and M. Wölfel, "Aspects of visual avatar appearance: Self-representation, display type, and uncanny valley," *The Visual Computer*, vol. 38, no. 4, pp. 1227–1244, Jun. 2021. DOI: 10.1007/s00371-021-02151-0.

[25] K. Mrkva and L. V. Boven, "Salience theory of mere exposure: Relative exposure increases liking, extremity, and emotional intensity.," *Journal of Personality and Social Psychology*, vol. 118, no. 6, pp. 1118–1145, Jun. 2020. DOI: 10.1037/pspa0000184.

[26] M. J. Brandt, H. IJzerman, A. Dijksterhuis, *et al.*, "The replication recipe: What makes for a convincing replication?" *Journal of Experimental Social Psychology*, vol. 50, pp. 217–224, 2014. DOI: https://doi.org/10.1016/j.jesp.2013.10.005.

[27] A. A. Harrison, "Mere exposure," in ser. Advances in Experimental Social Psychology, L. Berkowitz, Ed., vol. 10, Academic Press, 1977, pp. 39–83. DOI: https://doi.org/10.1016/S0065-2601(08)60354-8.

[28] B. Kelly, M. Margolis, L. McCormack, P. A. LeBaron, and D. Chowdhury, "What affects people's willingness to participate in qualitative research? an experimental comparison of five incentives," *Field*

Methods, vol. 29, no. 4, pp. 333–350, Jun. 2017. DOI: 10.1177/1525822x17698958.

[29] K. ISHII, "Does mere exposure enhance positive evaluation, independent of stimulus recognition? a replication study in japan and the usa1," *Japanese Psychological Research*, vol. 47, no. 4, pp. 280–285, 2005.

[30] J. Deuchler and M. Wölfel, "Lessons learned in transferring a lecture on virtual reality into immersive virtual reality," 2022.

[31] M. Feick, N. Kleer, A. Tang, and A. Krüger, "The virtual reality questionnaire toolkit," ser. UIST '20 Adjunct, Virtual Event, USA: Association for Computing Machinery, 2020. DOI: 10.1145/3379350.3416188.

[32] I. E. Sutherland, *The ultimate display*, 1965.

9 Asymmetric Normalization in Social Virtual Reality Studies

Publication Note: This chapter is based on the following published work. The content of the chapter is identical to the published article, with only the formatting and numbering being adapted for this dissertation.

Deuchler, J., Hepperle, D. and Wölfel, M. (2022)
IEEE Conference on Virtual Reality and 3D User Interfaces
Abstracts and Workshops (VRW), pp. 51-53
DOI: 10.1109/VRW55335.2022.00019

Abstract

We introduce the concept of *asymmetric normalization*, which refers to decoupling sensory self-perception from the perception of others in a shared virtual environment to present each user with a normalized version of the other users. This concept can be applied to various avatar-related elements such as appearance, location, or non-verbal communication. For example, each participant in a polyadic virtual reality study can see other participants at an average height of the respective test population, while individual participants continue to see themselves embodied according to their actual height. We demonstrate in a pilot experiment how asymmetric normalization enables the acquisition of new information about

social interactions and promises to reduce bias to promote replicability and external validity.

Keywords Human-centered computing – Human computer interaction (HCI) – HCI design and evaluation methods User – studies

9.1 Introduction

The ability of *virtual reality* (VR) to liberate users and researchers from real-world constraints and enable entirely new experimental setups is an opportunity not fully explored. Studies in VR allow the virtual environment and the embodiment of the participant to be altered according to the defined hypothesis. This could mean, for example, altering the participant's phenotype to reduce their bias against a specific group (see [1]).

In virtual multiplayer environments, it is possible to tailor the shared environmental data to each of the cooperating users' needs (i.e. text labels translated accordingly), a mechanism which has been introduced as *subjective views* [2]. For virtual interactions, a basic idea of decoupling interactant behavior from other users' view has been described by Loomis et al. more than 20 years ago [3]. Later, this has been further advanced with the *transformed social interaction* paradigm which describes the possibility of interactants to break many constraints inherent in face-to-face interaction by strategically altering their information stream differently for each user simultaneously [4]. What is new is that the design possibilities of VR can also be used to reduce the degree of freedom in social studies by normalizing virtual experiences to improve the scientific process in polyadic studies.

The *Virtual Reality Scientific Toolkit*[1] (VRSTK) is an open-source Unity3D package being developed by us. Our goal is to facilitate VR studies by providing specific templates with various data acquisition and analysis

[1] The latest version of the VRSTK is available for download via GitHub as a Unity3D package and in source code from https://github.com/ixperience-lab/VRSTK.

functionalities [5], while simultaneously reducing human-induced errors in user studies [6]. Here we present *asymmetric normalization* as a work-in-progress feature that extends multi-user VR functionality provided with the toolkit.

9.2 Asymmetric Normalization

Asymmetric normalization describes a technique that presents an individualized state of the world to each participant in a shared multi-user VR environment by asymmetrically altering the virtual environment and decoupling sensory self-perception from others' perception of one's self. This means that participant *A* does not necessarily see participant *B* the same way as participant *B* sees themselves, but an *altered* version of their virtual representation. By normalizing individual characteristics, which refers to either adjusting them to the mean of a target population or according to the individual's social or cultural backgrou nd, we aim to reduce bias and foster replicability and (external) validity. Additionally, asymmetric normalization can break apart bidirectional interaction elements like the interpersonal distance and thus, enables the gathering of additional information on social interactions that can not be measured outside the virtual world.

This concept of presenting everybody with a normalized view of the world can be applied to many characteristics of avatar interactions, such as appearance (e.g. phenotype, gender, apparel, prosthetics), location, and *non-verbal communication* (NVC) which are planned to be implemented in the VRSTK.

9.2.1 Appearance Normalization

The embodiment of participants through avatars already represents an abstraction compared to the representation of a real person. It is not perfect or detailed enough even in those cases where photogrammetry is involved. Matching the appearance of the avatar to the real person is by today's standards too time-consuming and thus—in most cases—cartoonish characters are used in social VR tools [7]. Besides the truthful representation of facial features, other individual features like jewelery, tattoos, or prosthetics are rarely considered [8]. If participants can only be partially represented in VR, it is reasonable to normalize the complete embodiment from the perspective of other participants.

Participants' appearance may have undesirable effects on other participants as well as on the perception of one's self. While, for example, it is critical to adjust the height of the avatar to the user controlling it to maximize immersion and to prevent irritation and other influence such as the *Proteus effect*[2], it can also be irritating if the conversation partner is presented much taller or smaller than the average. Human height is positively related to interpersonal dominance [10] and thus could influence social interaction and study outcomes. Therefore, from a participant's perspective, scaling the virtual representations of fellow participants to an average height and keeping one's own representation to reflect the true height could reduce phenotypic bias.

We refer to a process where particular appearance features are normalized from the perspective of everybody in a shared environment except oneself

[2] The *Proteus effect* describes the phenomenon that people conform to stereotypes associated with their avatar's visual characteristics [9].

as *appearance normalization*. While changing superficial aspects such as skin color or clothing is fairly trivial, changing other aspects such as body morphology is error-prone and entails potential adjustments to tracking data. Altering the height, for example, requires adjusting the eye angles in order to maintain natural gaze interaction.

9.2.2 Location Normalization

Typically, in a dyadic conversation, both participants must agree on a common distance between each other and the interplay between them is part of the interaction.[3] Due to this, measuring individual preferences for interpersonal distance is not feasible in polyadic real-world scenarios. VR offers the possibility to change the position of the participants individually. Breaking the symmetry, every participant can now move from their own perspective but stand still from the other's perspective, and vice versa. We refer to this process as *location normalization*. It enables to measure and analyze information about social interactions—individual interpersonal distance preferences—that otherwise cannot be extracted.

Note that particular care has to be taken in the case of full-body avatars, as leg and feet movement cannot simply be inherited from tracking data but needs to be replaced by an inverse kinematic assumption from the perspective of others.

[3] Research on this topic is usually referred to under the term *proxemics*.

9.2.3 Non-Verbal Communication Normalization

The concept of *non-verbal communication normalization* refers to the asymmetric alteration of all tracked features that might influence interaction in a certain direction such as facial expressions, gaze, gestures, or posture. Similar to location normalization, the visualization of the tracking can be suppressed or altered so that other participants only see the normalized state of the face, arms, or posture, while one's self does not see the alternation. For example, posture affects empathy and dominance [11], which in some cases might be an unwanted side effect in a polyadic study and therefore can be normalized across participants. Another aspect in this regard is misunderstandings based on intercultural differences. For example, when thinking about eye contact, which signals confidence in the west and can be seen rude in parts of Asia, asymmetrically normalizing the gaze direction in accordance to the participant's socialization can help overcome misunderstandings and thus reduce unwanted biases.

9.3 Proof of Concept

To prove the proposed concept of asymmetric normalization, we conducted a preliminary study with 40 participants applying appearance, location, and NVC normalization. Two different forms of tracking fidelities were compared during a dyadic avatar-avatar conversation, one with *high tracking fidelity* (including full-body, hand, and facial tracking) and one with *low tracking fidelity* (without full-body, hand, and facial tracking).

Figure 9.1 shows a visualization of appearance (height), location, and NVC normalization in a dyadic conversation as implemented in the study. Two

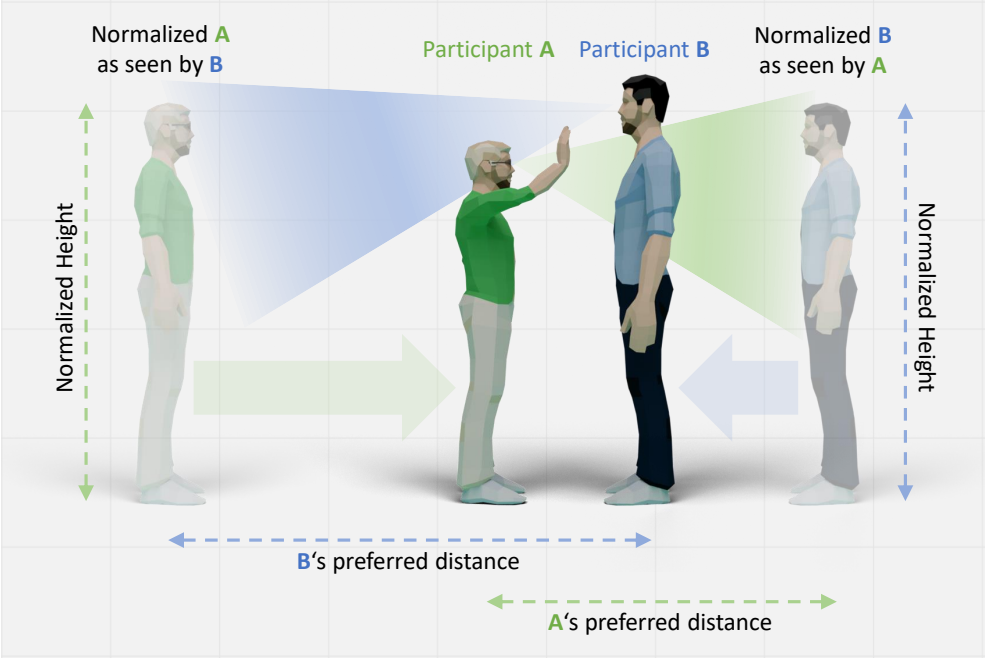

Figure 9.1: Visualization of a conversation between a shorter than average participant *A* (green, wearing glasses) and a taller than average participant *B* (blue, no glasses) with appearance (height), location, and non-verbal communication normalized. Both participants only see the normalized version (visualized transparent) of their opposing participant.

participants enter a shared virtual environment where they are embodied with high tracking fidelity from their own perspective and are tasked with getting to know each other. Each participant sees the other as a *normalized* version (transparent in Figure 9.1).

During the conversation, both participants take their preferred interpersonal distance, while the interlocutor seems to stay in place for them (location normalization). In Figure 9.1, participant *A* is moving closer compared to participant *B*. This also explains how asymmetric normalization allows for the collection of information that would otherwise be unobservable in a regular, real-world conversation.

In addition, the interlocutor appears to be of average male height for both participants, while in reality, both deviate from the average (appearance

normalization). In Figure 9.1, participant *A* is actually shorter than average, while participant *B* is taller than average. Both participants believe to interact with an average-sized person which is a better fit for a real-world scenario.

Finally, participant *A* waves, but participant *B* does not see this because the virtual environment of the normalized version of *A* withholds any tracking related movements to satisfy the desired low tracking fidelity embodiment condition for testing the hypothesis (non-verbal communication normalization).

An independent-samples t-test was conducted to compare the individually taken interpersonal distance at the end of a two minute conversation between the participants, one represented with high and the other with low tracking-fidelity. At the beginning of the experiment, both avatars were placed approximately two meters away from each other. The interpersonal distance after two-minutes was significantly closer for participants whose opposing avatars possessed *low tracking fidelity* (distance after 120s: μ = 1.68m, σ = 0.34m) compared to the distance participants took towards avatars with a *high tracking fidelity* (distance after 120s: μ = 1.89m, σ = 0.22m) with t(38) = –2.44, p = 0.02, d = –0.77.

9.4 Conclusion and Outlook

Social interactions in VR are increasingly becoming the subject of research due to the emergence of social VR platforms. As a result, research methods for polyadic studies in VR need to be explored. While the advantages of VR for studying human behavior, such as data acquisition and

analysis, have been exploited since the first applications of VR, asymmetric normalization represents a novel approach that appears to be useful for gaining additional insights within VR, the findings of which could also be transferable to the real world.

Asymmetric normalization provides several options and methods for standardization that help researchers in polyadic studies to reduce bias and thus to promote replicability. Furthermore, the technique enables the exploration of entirely new forms of social interactions in VR that were previously difficult or impossible to observe in polyadic studies. This is demonstrated in the pilot study presented, where location normalization is successfully used to measure significant differences in preferred interpersonal distance associated with tracking fidelity.

It is important to note that these techniques have the potential to influence the interaction between participants and can therefore lead to different effects. On one hand, this could be used, for example, by preventing unintentional intruders into the personal space without the other even noticing. On the other hand, the interpersonal distance could be used as a means of communication, or the appearance of the avatar could be used as a form of expression and statement, which might be obscured by this technique. Therefore, the use of asymmetric normalization must be carefully considered to not cause irritations and odd social interactions.

Furthermore, it is a topic of debate whether participants should be informed about the use of asymmetric normalization and the accompanying concealment of their appearance and behavior before conducting a user study. The knowledge about asymmetric normalization might also influence participants' behavior.

Applying asymmetric normalization to compensate for gender- or ethnicity-specific effects only to visual features is technically less challenging than on visual and acoustic features simultaneously. However, recent advances in deep learning, such as voice transfer, are promising developments in this regard that further increase the possibilities and applicability of the proposed approach of asymmetric normalization.

References

[1] B. S. Hasler, B. Spanlang, and M. Slater, "Virtual race transformation reverses racial in-group bias," *PLOS ONE*, vol. 12, no. 4, pp. 1–20, Apr. 2017. DOI: 10.1371/journal.pone.0174965.

[2] G. Smith, "Cooperative virtual environments: Lessons from 2d multi user interfaces," in *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work*, ser. CSCW '96, Boston, Massachusetts, USA: Association for Computing Machinery, 1996, pp. 390–398. DOI: 10.1145/240080.240350.

[3] J. M. Loomis, J. Blascovich, and A. C. Beall, "Immersive virtual environment technology as a basic research tool in psychology," *Behavior Research Methods, Instruments, & Computers*, vol. 31, pp. 557–564, 1999.

[4] J. N. Bailenson, A. C. Beall, J. Loomis, J. Blascovich, and M. Turk, "Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments," *Presence: Teleoper. Virtual Environ.*, vol. 13, no. 4, pp. 428–441, Aug. 2004. DOI: 10.1162/1054746041944803.

[5] M. Wölfel, D. Hepperle, C. Purps, J. Deuchler, and W. Hettmann, "Entering a new dimension in virtual reality research: An overview of existing toolkits, their features and challenges," Sep. 2021, pp. 180–187. DOI: 10.1109/CW52790.2021.00038.

[6] D. Hepperle, T. Dienlin, and M. Wolfel, "Reducing the human factor in virtual reality research to increase reproducibility and replicability," in *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, IEEE, Oct. 2021. DOI: 10.1109/ismar-adjunct54149.2021.00030.

[7] B. Yoon, H.-i. Kim, G. A. Lee, M. Billinghurst, and W. 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)*, 2019, pp. 547–556. DOI: 10.1109/VR.2019.8797719.

[8] D. Hepperle, C. F. Purps, J. Deuchler, and M. Wölfel, "Aspects of visual avatar appearance: Self-representation, display type, and uncanny valley," *The Visual Computer*, Jun. 2021. DOI: 10.1007/s00371-021-02151-0.

[9] N. Yee, J. Bailenson, and N. Ducheneaut, "The proteus effect implications of transformed digital self-representation on online and offline behavior," *Communication Research*, vol. 36, Apr. 2009. DOI: 10.1177/0093650208330254.

[10] G. Stulp, A. P. Buunk, S. Verhulst, and T. V. Pollet, "Human height is positively related to interpersonal dominance in dyadic interactions," *PLOS ONE*, vol. 10, no. 2, pp. 1–18, Feb. 2015. DOI: 10.1371/journal.pone.0117860.

[11] D. Küster, E. G. Krumhuber, and U. Hess, "You are what you wear: Unless you moved—effects of attire and posture on person perception," *Journal of Nonverbal Behavior*, vol. 43, no. 1, pp. 23–38, Oct. 2018. DOI: 10.1007/s10919-018-0286-3.

10 Summary

This dissertation consists of 5 different papers that all contribute to the central question on how IVR can contribute to research in the social sciences. The research investigates the transferability and ecological validity of findings from IVR to the real world, and confirms the potential of IVR as a research tool.

The scoping review provides a comprehensive overview of the current state of research by highlighting similarities and differences between IVR, 2D monitors, and physical reality [77]. Since some results are influenced by technical constraints, the mere exposure effect was chosen as a fundamental psychological phenomenon, as it is robust against such limitations and therefore suitable for evaluating ecological validity [78]. The findings show that the mere exposure effect could be consistently replicated within an IVR environment. This represents a first step toward examining the transferability of such effects and underscores the potential of IVR as a scientific methodology. Furthermore, the results demonstrate that automation in IVR studies can improve the reproducibility and generalizability of findings [75].

New concepts such as asymmetric normalization help to reduce bias in social interactions in ways that are only partially achievable in physical reality [79].

To assess the feasibility of such studies, various IVR toolkits were examined [76]. The work reveals the diversity of available software solutions and identifies challenges such as the lack of benchmarks and user-friendly interfaces. At the same time, it emphasizes the potential chances of integrating AI in future developments.

These findings underscore the potential of IVR to address the replication crisis and expand the methodological capabilities of the social sciences.

10.1 Synthesis of Findings

In Chapter 3 Theoretical Background: Immersive Virtual Reality as a Research Tool, four research questions (RQ 1 — RQ 4) were defined. For each research question, the associated publications and the underlying study design are shown below, along with a summary of the key findings. Table 4.1 shows an overview of all questions and the related papers. These findings are then synthesized to provide a direct answer to the corresponding research question.

10.1.1 Research Question 1

[RQ 1] In which parts of the research process can IVR be the most helpful and how do toolkits contribute to replicability and reproducibility?

Paper 1

Paper 1 follows a theoretical-exploratory design. It addresses the replication crisis in Human-Computer Interaction (HCI) and IVR by systematically adapting solutions from other disciplines and mapping them onto the specific challenges faced in HCI and IVR. The objective is to critically evaluate these approaches and tailor them to the field, ultimately improving the reproducibility and replicability of research in IVR.

Main Findings Paper 1

- The publication focuses on demonstrating how the *human factor*—specifically, human-induced errors—can be reduced through studies conducted in IVR. Parallels are drawn from disciplines such as computer science to illustrate established practices that enhance traceability (e.g. version-control software) and distribution (the study exists as program code and can therefore be executed anywhere) for use in social science research.

- A central theme is the *automation* of study preparation within IVR environments. By automating key processes, the human component can be minimized, which in turn increases the generalizability of results.

- The so-called "*Areas of Concern*" in a standard scientific study workflow are mapped onto an abstracted procedure for IVR studies. These areas are then broken down in detail, and specific recommendations are provided for each corresponding Area of Concern.

Paper 2

A comparative evaluation study was conducted in this work to examine how well seven existing frameworks are suited for conducting studies in IVR. To this end, five key categories—each with at least three subcategories—were identified as relevant for IVR-based research.

The analysis considers the extent to which the frameworks support components such as "Setup & Control" (e.g., process planning, scene customization), "Sensing Participants" (e.g., tracking head movements and biosignals), "Representation" (e.g., depiction of participants and objects within the virtual environment), and "Data Handling" (e.g., importing, analyzing, and sharing data).

Main Findings Paper 2:

- The publication provides an overview of a wide range of toolkits available for research and development in the field of IVR. These range from commercial software to open-source platforms, each tailored to specific needs and functionalities.

- The paper describes the features of various IVR toolkits, including support for multi-user interactions, integration with external devices, and compatibility with different IVR hardware. Some toolkits are specialized for specific tasks, while others offer a broad range of functionalities.

- The publication discusses several challenges associated with the development and use of IVR toolkits. These include issues such as

the steep learning curve for beginners and the need for standardized tests and benchmarks.

- The research concludes with a perspective on the potential future development of such toolkits. This includes the integration of artificial intelligence and machine learning, the development of more user-friendly interfaces, and the need for interdisciplinary collaboration to fully leverage the potential of IVR-based studies.

Synthesis for RQ 1

In Paper 1 the different areas of concerns in the research process are defined and adapted to potential solutions using IVR and IVR toolkits. Thos toolkits are then evaluated in Paper 2. A clear pattern emerges across both papers: IVR yields the greatest scientific benefits precisely where traditional behavioral studies struggle to maintain consistency: during setup, participant interaction, data capture, and post-study documentation. The core contribution of current toolkits is turning those steps into scripted, shareable software modules.

Conclusion RQ 1

IVR can support research where conventional studies rely on manual setup, subjective observation as well as extensive documentation. By embedding those steps in toolkits—templates for design, automatic high-resolution logging, standard export routines, and shareable code repositories—researchers turn experiments into portable software artifacts.

This software-centred workflow directly addresses the replication crisis: another team can 1) rerun the original code on the original data (reproducibility) and 2) rerun the same code on a new participant sample (replicability). This of course is not valid only for IVR enviornments but also could—at least partially—be applied when carrying out experiements on a 2D screen. Yet, options such as (re)-visiting the scene to feel the sense of being as a researcher or a participant there might yield greater insights compared to 2D screens. These factors are discussed in RQ 2.

10.1.2 Research Question 2

[RQ 2] What are the similiarities and differences between IVR, the real world and computer screens in studies from current literature?

Paper 3

In Paper 3 the following study design was applied:

A literature review in the form of a systematic scoping review was conducted in this work. A total of 1083 publications were identified, of which 56 publications were included in the qualitative synthesis following a multi-stage screening process in accordance with the PRISMA guidelines by Moher, Liberati, Tetzlaff, *et al.* [81] (See Fig. 7.2 for details). After screening, 56 articles remained and were compared for a qualitative synthesis that provides the reader with a summary of current research on the differences between HMDs, computer screens, and the real world.

Main Findings Paper 3:

- The goal of the systematic scoping review was to investigate similarities and differences between IVR, physical reality, and 2D monitors.
- The findings from the reviewed studies were categorized into three main areas: 1. Interaction, 2. Perception, and 3. Sensing and Reconstructing Reality, and further subdivided into subcategories to ensure comparability across the studies.
- The results are broken down in a table over all results (See Tab. 7.4), highlighting whether the result is from the perspective of IVR:
  - ▲ *advantageous* in relation to the screen or real world,
  - ▼ *disadvantageous* in relation to the screen or real world,
  - ► *similar* in relation to the screen or real world if there is no significant difference, and
  - ■ *undecided*, if no clear tendency can be inferred, but there is a significant difference.
- The research methodology and study designs (e.g., between-group vs. within-group Sec. 7.5.4, demographic data in Tab. 7.2, number of participants, type of questionnaire in Tab. 7.3) of the included studies were analyzed, compared, and detailed.
- Across the studies, more effects were found that show similarities between IVR and the real world than effects that indicate differences. For the comparision between IVR and screens it's the opposite.

- Most of the differences between IVR and physical reality identified in the studies are due to technical limitations. Many of these differences are likely to become negligible in the future as technology continues to advance.

The detailed results are shown in Tab. 7.5 and 7.6.

Conclusion RQ 2

The systematic scoping review conducted in Paper 3 provides a comprehensive overview of the similarities and differences between IVR, real world, and 2D screen-based environments. The analysis across 56 studies reveals that, in many domains of human experience—particularly interaction and perception—IVR demonstrates a high degree of similarity to real-world conditions. Notably, most of the identified differences stem from current technological constraints rather than fundamental conceptual limitations. These discrepancies are expected to diminish as hardware and software improve. The structured comparison shows that IVR is already capable of approximating real-world experiences in meaningful ways, often outperforming traditional 2D screens in providing immersive and naturalistic interactions. This supports the idea that IVR can serve as a viable and potentially alternative to both screen-based and physical setups in certain research contexts.

10.1.3 Research Question 3

[RQ 3] Are findings obtained in virtual reality transferable to real-world contexts?

Paper 4

In order to evalute if results from IVR are similar to those from the real world, two replication attempts were carried out in Paper 4. For the study design, we implemented two approaches: a direct replication of the work by Mrkva and Boven [80] conducted via Amazon Mechanical Turk (MTurk) using 2D monitors, and second, a replication in an IVR environment, carried out using the same procedures and stimuli in an IVR setting (See Fig 10.1). A total of 119 participants were recruited for the direct replication, and 63 participants for the replication in the IVR environment.

Main Findings Paper 4:

- The study replicated the mere exposure effect and demonstrated that increased exposure consistently enhances liking of the presented stimuli, both in the original setting and in IVR.
- The results confirm that the mere exposure effect remains robust and occurs with statistically similar strength in both traditional and IVR environments, supporting the ecological validity of IVR as a research tool.

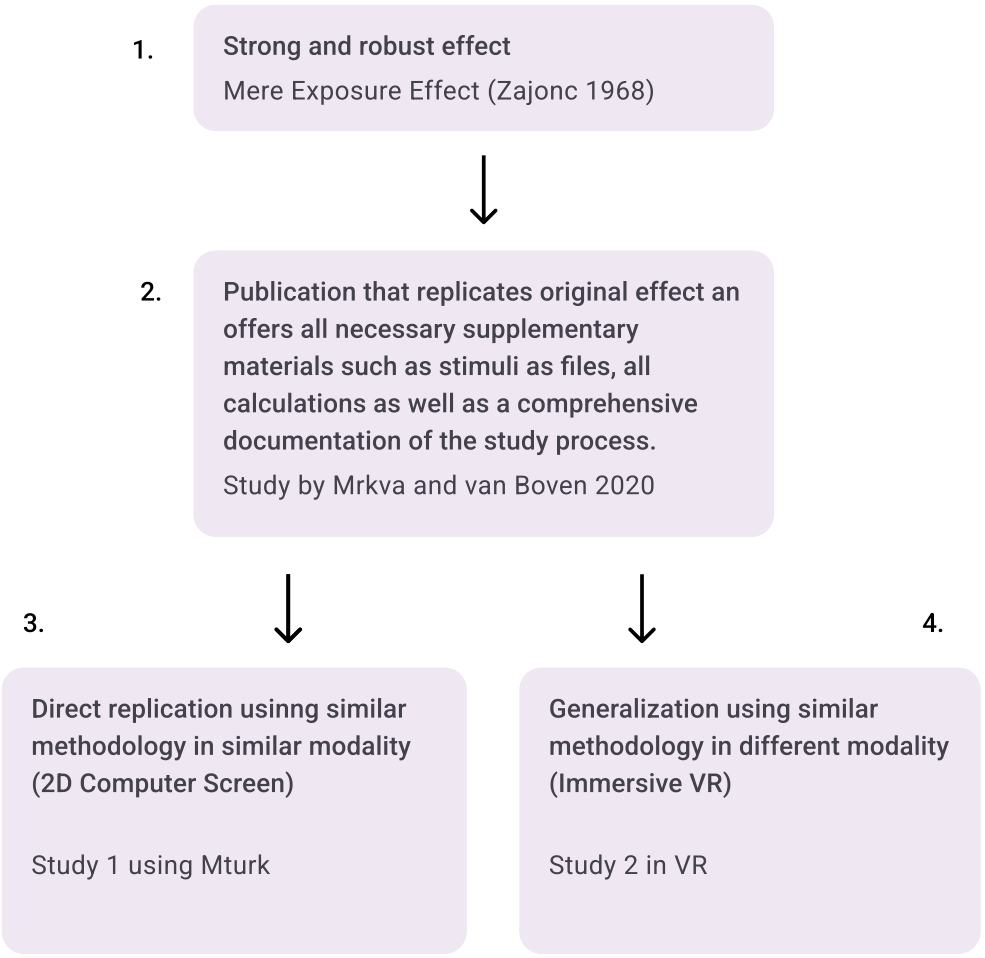

Figure 10.1: Study Design for the two replications carried out in Paper 4

- The findings closely align with earlier studies [80], including a meta-analysis by Montoya, Horton, Vevea, *et al.* [82].
- IVR proves suitable for replicating insights from traditional settings and offers the potential for cost-efficient and reproducible experiments in realistic and controlled environments.
- Despite its advantages, IVR faces technical challenges such as the lack of photorealistic graphics, cybersickness, and limited natural interaction capabilities, which may restrict the generalizability of findings to other effects and scenarios.

Conclusion for RQ 3

Using the mere exposure paradigm of Mrkva and Boven [80], Paper 4 executes two parallel replications: one with 119 participants on Amazon Mechanical Turk using 2D monitors and one with 63 participants in an IVR setting, both employing identical stimuli and procedures. The mere exposure effect emerges with comparable strength in both modalities, thus offering strong empirical support for **RQ3**: at least for this psychological phenomenon, IVR findings translate to laboratory based contexts and, by extension, to real-world interpretation (see also the alignment with the meta-analysis of [82] in Fig. 8.2).

We, however, acknowledge boundary conditions: graphics realism, cybersickness and the naturalness of interaction that may moderate transfer claims for less robust effects.

10.1.4 Research Question 4

[RQ 4] How can IVR factors be manipulated and varied to maximize knowledge gain, and how can potential biases be reduced?

Paper 5 introduces a new concept to equalize certain factors in IVR is described in order to analyze chances to maxmize knowledge gain and lower potential biases in social science studies.

The research design is experimental and exploratory. A new theoretical model is developed that describes how a user's *sensory self-perception*

can be decoupled from the perception of others within a shared virtual environment.

To validate the concept empirically, a pilot experiment with 40 participants demonstrates its practical applicability and initial effects.

Main Findings Paper 5

- Asymmetric normalization can decouple a user's sensory self-perception from how other users are perceived, so that each participant sees a *normalized* version of the others.
- The concept can be applied to features such as appearance, position or non-verbal communication. For example, in a multi-user VR study, every participant could see the other users at an average population height while still experiencing their own avatar at their actual height.
- The pilot study shows that asymmetric normalization enables novel ways to measure social interaction, which would not have been possible in a regular setting in the real world (See. Fig. 9.1). Thus, IVR in some cases, might have the chance to gain findings about phenomenon in a more direct way than the real world.

Conclusion for RQ 4

The paper outline an interesting idea on how to maximize knowledge acquisition and minimizing bias in IVR research.

- Paper 5 provides a concrete, theory-driven manipulation—"Asymmetric Normalization"—that expands the range of social variables that can be explored, thereby enriching research possibilities and knowledge gain.

- Perceptual and social biases are reduced by adjusting how participants perceive one another in real time.

11 Contribution to Research

This dissertation situates IVR as part of a common research process. By showing how different stages of the research pipeline can be supported, the work demonstrates that IVR can add to experimental control without sacrificing ecological validity. The findings contribute to an ongoing approach in defining if and how to use IVR as a research tool. In doing so, it advances ongoing efforts to improve replicability and reproducibility in the social sciences while simultaneously broadening the kinds of phenomena that can be studied.

11.1 Extending the Experimental Design Space

Figure 5.1 maps the typical pitfalls in the research workflow and gives readers a concise overview of the stages where IVR toolkits can provide the most effective interventions.

Advancements in the aforementioned interventions are facilitated by an expanding array of specialized toolkits, a concept that has been identified as a pivotal component of research involving IVR in research. These tools are necessary to approach the "gold standard", as mentioned by Kothgassner and Felnhofer [69]. Those toolkits are evaluated in five categories

essential for any IVR study: Setup & Control (i.e. does the toolkit help set up remote testing?), Sensing Participants (i.e. does the toolkit offer an option to record eye and gaze movements?), Representation (i.e. does the toolkit offer a solution to represent participants?), Data Handling (i.e. does it offer a way to import and export data?), and Integration (i.e. is it open source?). The importance of such toolkits, also for non-experts is mentioned by Wang and Bailenson [83] in their conclusion as an essential part for progressing the use of IVR in research practice.

By applying the concept of asymmetric normalization traditional control variables can be manipulated to an extent that cannot be applied in common real world situations. Wang, Miller, and Bailenson [84] likewise explore how IVR can enable scenarios impossible in the physical world. Among other innovations, they examine a "rewind function" that lets users step back in time while other people remain present in the scene—a capability that is similar to the revisiting feature described in Papers 1 and 2.

11.2 Strengthening Ecological Validity

Gaps in current research regarding the special differences in IVR compared to the real world are identified and summarized in the systmatic scoping review (Paper 3). Most observed differences (IVR vs. real world) were traced to temporary technical limitations (e.g. graphics fidelity, input hardware), suggesting these gaps will narrow with technological advances. This contribution presents a first overview where IVR results might hold true in the real world improving confidence in IVR as a tool for social science research.

Next to this, within the dissertation we extend the IVR toolbox by an individual *tool* inside that toolbox. Using procedures such as asymmetric normalization—a technique that decouples a user's self-perception from how others see them in the virtual world—offers new possibilities to draw inferences in experiements. We are currently implementing this manipulation in a triadic IVR discussion taks: three participants debate a topic, and for eachobsever one of the interlocutor's avatars includes full facial-expression tracking while for the other avatar shows gaze and facial movements generated by a pseudo-random, rule-based model that derives its dynamics from cues such as speech rhythm and head orientation. Because all three participants share the same interaction while perceiving facial-cue fidelity at two distinct levels, the design yields precise within-conversation contrasts of how accurate versus algorithmically approximated facial signals shape social perception and behaviour.

Another contribution of this work is the successful replication of the mere exposure effect in IVR.

This replication thus joins a small but growing set of replications that 1) prove that they are able to replicate the effect in the same setting as the original study and 2) can replicat the effect in IVR. The replication of the mere exposure effect demonstrate not only that effects observed in physical environments persist in VR, but that their underlying causal mechanisms might as well.

Given the concerns raised by the Open Science Collaboration [71] about the reproducibility of classic effects, such evidence is vital. Our findings suggest that IVR can reliably elicit even subtle psychological phenomena, despite the shift in medium.

Furthermore, the study adheres to open science practices: data, materials, and analysis scripts are publicly accessible on OSF[1], ensuring transparency and reproducibility. This underscores IVR's potential as a rigorous and ecologically valid research platform.

11.3 Chances and Ethical Challenges

Finally, research in IVR has the potential to enable experiments on human behavior in scenarios that are either too expensive or too dangerous in real life. Of course, given the idea that a IVR experience is comparable to a real world experience, this might be ethically questionable. This moral dilemma is discussed by Ramirez [85]. In their work they formulate *The Equivalence Principle (TEP)*, stating that "If it would be wrong to subject a person to an experience, then it would be wrong to subject a person to a virtually-real analog of that experience. As a simulation's likelihood of inducing virtually-real experiences in its subject increases, so too should the justification for the experimental protocol." Afterwards they discuss possibilites how virtual realism can be traded against a possible traumatization of participants. McIntosh [86] addresses this the problem and discusses ways to mitigate the risk of traumatizing or re-traumatizing participants within a IVR because of the "perceptual proximity" to the experience.

[1] See: `https://osf.io`

12 Implications for Research Practice

As consumer-grade HMDs become more affordable, high-end systems are delivering greater fidelity and a diverse set of sensors, such as Electroencephalography (EEG), or eye- and face-tracking cameras, IVR has a potential to become widely available in research facilities. In short, the main practical implications drawn from this work are:

- **Toolkit benchmarking** — An assessment of state-of-the-art IVR toolkits that support researchers in designing, implementing, and executing immersive-virtual-reality studies.

- **Cross-reality scoping review** — The first comprehensive synthesis of empirical work comparing IVR with both real world and 2D screen settings, showing where findings converge and where gaps persist. This overview equips researchers to spot eventual pitfalls and pinpoints priority areas for future IVR research.

- **Replication study** — A preregistered replication confirms that the mere-exposure effect generalizes to IVR, demonstrating the platform's suitability for theory-driven, confirmatory research and establishing a benchmark protocol.

- **Asymmetric normalization** — Introduces an IVR technique that decouples how participants perceive themselves from how they perceive others by normalizing co-participant avatars (e.g., mapping everyone else to average height while preserving one's own perspective.)

While this is encouraging, important challenges associated with immersive virtual reality still need to be considered.

Those challenges are summarized and categorized in Tab. 7.5 as well as in Tab. 7.6. The table offer a clear, summary of the main advantages and disadvantages of IVR. At a brief overview, you can see where IVR, the real world, or standard 2D screens align—or diverge—in three key areas: perception, interaction, and sensing and rebuilding the surrounding world.

Another important aspect is the growing ecosystem of user-friendly toolkits that potentially allows scholars without a computer science and 3D artist / game developedment background to build custom experimental environments with minimal coding. Here this dissertation contributes to research practice in a way that it presents a clear research pipeline for IVR in Fig. 6.1 which is essential for researchers that are new to using IVR as a research tool. Afterwards, the different available toolkits are presented and compared against each other in Tab. 6.1. This helps researchers to get an idea of the potentials of such toolkits as well as the chance to identify which package might meet their experimental needs the most.

The concept of asymetric normalization offers an idea of how to make extended use of IVR. Fig. 9.1 shows how we can normalize participants height for example. The concept also has the potential to be applicable

to normalize other factors such as appearance or position normalization which eventually might help increase generalizability of results.

Finally, with the replication of the mere exposure effect, at least to the best of my knowledge, we show a first of its kind procedure that is based on a robust effect which was replicated by the authors in the same environment as well as within the IVR environment.

Even though it still takes effort to set up such a research environment, in the end it might be cheaper than creating a complex experiment in a real world setting.

This offers several new possibilities:

- Students as well as senior researchers now have the change to create highly complex research environments with almost limitless, interactive stimulis while maintaining high ecological validity that would not have been possible in the real world.

- The research environment can be set up all around the world like a smartphone app basically with the click of a button.

- Researchers can participant in remote settings from whererever they are.

- With the concept of asymetric normalization we see a tendency and allowing those experiements for a higher degree of geralizibility and reduced bias in social interactions.

In detail, the adaption of the common research process allows future researchers to identify potential pitfalls and shows them how to lower

the risk of human-induced error. For these "Areas of Concern" in a generalized research process, features from virtual reality research toolkits are presented to cope with the problems.

Also, the general idea of the dissertation shows high potential that it might work in reverse as well: Adapting procedures from the real world into IVR. For example, a meta-analysis by Carl, Stein, Levihn-Coon, *et al.* [87] shows that virtual environments can elicit the same fear structures as in-vivo exposure and often achieve comparable—or even superior—clinical outcomes at lower cost and risk. IVR therefore becomes not only a measurement tool but a therapeutic platform.

13 Conclusion, Limitations and Outlook

This dissertation set out to evaluate the potential of using IVR as a tool for research in the social sciences. Taken together, the five studies show that IVR

- has the potential to reduce human-induced error and thus improve replicability and reproducibilty.
- research toolkits already integrate the core functionality needed to design, run, and log experiments, and many are released under open-source licences that make them readily extensible to project-specific requirements.
- does cover certain aspects of the real-world but still differs in some aspects in comparison to the real-world and 2d screens. These challenges are mainly limited by technological innovation.
- can faithfully reproduce the well-established mere exposure effect, laying a solid foundation for further investigations.

While the findings advance IVR research, they are still bouded by several limitations. Of course there are technical limitations which we already can see some of them are solved just recently. For example Lin, Gao, Tang, *et al.* [88], presented an Artificial Intelligence (AI) approach to translate

text and image prompts into 3D objects and scenes which was one of the major challenges for researcher to create virtual scenes. Another important aspect is the depicition of high-fidelity social avatars [69], [89]. A challenge that typcially involves experts from different domains has just been made less complex by the use of LLMs. Unreal Engine for example ships with add-on SDKs such as the Convai[1] plug-ins built on the MetaHuman framework—that let researchers spawn virtual humans able to perceive the scene, navigate through it, multilingually converse via a LLM, and automatically synchronize facial expressions and body language with the generated speech. These non-verbal behaviours such as eye gaze, posture, gesture, facial micro-expressions, carry much of the affective and relational content in human exchange. It directly influences trust and rapport, yet the validity of conclusions drawn from this needs further investigation.

Therefore, future research is needed to benchmark which non-verbal channels must be rendered at what level of fidelity for different study goals. Interestingly, for visual depiction, Slater, Pertaub, Barker, *et al.* [67] found, that virtual avatars with low visual fidelity can induce the same level of anxiety as a more realistic version. Mori, MacDorman, and Kageki [90] introduced the uncanny-valley hypothesis, arguing that near-human but imperfect representations evoke a sense of eeriness. Recent work in IVR confirms and amplifies this pattern: Hepperle, Purps, Deuchler, *et al.* [91] show that users' sensitivity to such imperfections is even stronger when avatars are experienced in a HMD. Addressing these topics are necessary to ensure that future IVR experiments using artificial participants capture

[1] `https://convai.com/` (Also available for Unity).

not only what people say but how they silently negotiate meaning, status, and emotion (as far as this is even possible in the real world).

Another interesting approach regarding the documentation of the research process using LLM was recently introduced by Kim, Aamir, Singh, *et al.* [92] for Extended Reality (XR). With their toolkit, they provided a streamlined pipeline to analyze user behavior in XR environments by leveraging LLMs. If and how this analysis will contribute to ecological validity is currently a topic of debate because LLMs are not error-prone.

Also, with recent HMDs such as the Meta Quest 3 and Apple Vision Pro deliver high-quality video-passthrough often branded as MR or spatial computing. Unlike optical see-through AR headsets, video-passthrough presents a wide, IVR-typical field of view (≈ 90-110°). The HMD front cameras stream a live, stereoscopic view of the real environment to the internal displays. Spatial-tracking data then anchors virtual objects with high precision to the physical surroundings (i.e. walls, desks among others). Because every pixel of the video feed can be blurred, recoloured, or occluded in real time, researchers can both preserve full peripheral context and manipulate selected elements of the scene. An intersting combination for social science experiments since nowadays, the virtual surrounding can be annotaded and analyzed in almost real time with tools such as "Locate 3D: Real-World Object Localization via Self-Supervised Learning in 3D" by Meta[2].

This leads us to another major challenge: The heavy dependence on proprietary hardware, software platforms, and cloud infrastructures of which many are controlled by companies and data centers mostly located

[2] `https://locate3d.atmeta.com/`

outside the European Union. This raises data-privacy and sovereignty concerns and creates the risk that key research capabilities will disappear if a vendor discontinues a product line. A recent example is the Meta Quest Pro Headset: its advanced face-tracking features have proven valuable for social science experiments, yet future support remains uncertain. Future investigations should therefore not only focus on theoretical impact of IVR but also on hard- and softwarespecific concerns related to the topic.

The results of this dissertation certainly can also be transferred into other domains. If the ecological validity of IVR is high, learning outcomes might be similar or even higher than in the real world. This opens up another interesting opportunity: immersive learning. IVR allows educators to construct experiences that are impossible or too expensive in the real world. In light of the findings presented in this dissertation, it is advisable to exercise caution. Virtual environments still differ from the physical world, so certain high-stakes training scenarios—for example, a hostage-rescue exercise[3] or a trauma-bay simulation[4]—should be examined critically. Although IVR can deliver valuable procedural learning, the lack of real world consequences may diminish the sense of responsibility and pressure that practitioners ultimately need. Whether that "consequence gap" undermines transfer to the real setting remains an open question and deserves further discussion.

Also, current research suggests, wearing a HMD for longer periods of time might have impact on the physicial comfort and visual fatigue of the participants [93] especially in learning situations. This should also

---

[3] https://mixed.de/new-yorker-polizei-trainiert-amoklauf-szenarien-in-virtual-reality/

[4] https://tricat.net/neuigkeiten/universitaet-ulm-medizinstudierende-trainieren-in-vr/

be addressed when thinking about the “wild-west” practices using IVR as described by Vasser and Aru [94] when further working on the (still missing [95]) “gold standard” as proposed by Kothgassner and Felnhofer [69].

Ultimately, the dissertation points toward an exciting possibility: that IVR could become a credible research instrument for the social sciences even though further investigation is necessary. Given the astonishing pace at which the underlying technology is advancing, two futures seem plausible: either IVR will reach a level of fidelity that mirrors the physical world so precisely that findings transfer seamlessly, or IVR will evolve into such a rich domain in its own right that studies need only be valid inside virtual environments to remain scientifically valuable. Let’s hope for the better one!

References

[1] E. Brunswik, *Perception and the Representative Design of Psychological Experiments*. University of California Press, Dec. 1956. DOI: 10.1525/9780520350519.

[2] J. Fahrenberg, *Wilhelm Wundt (1832–1920): Introduction, Quotations, Reception, Commentaries, Attempts at Reconstruction*. Lengerich, Germany: Pabst Science Publishers, 2020, Abridged English translation of the 2018 German edition. DOI: 10.23668/psycharchives.10325.

[3] J. M. Converse, *Survey Research in the United States: Roots and Emergence 1890–1960*. Berkeley, CA: University of California Press, 1987.

[4] L. L. Thurstone, "Attitudes can be measured," *American Journal of Sociology*, vol. 33, no. 4, pp. 529–554, Jan. 1928. DOI: 10.1086/214483.

[5] R. Likert, *A Technique for the Measurement of Attitudes* (Archives of Psychology). New York: Columbia University Press, 1932, vol. 140, pp. 1–55.

[6] K. Koffka, *Principles of Gestalt Psychology*, English. New York: Harcourt, Brace & Company, 1935, APA PsycNet record 1935-03991-000.

[7] M. Sherif, O. J. Harvey, B. J. White, W. R. Hood, and C. W. Sherif, "Intergroup conflict and cooperation: The robbers cave experiment," *University of Oklahoma Book Exchange*, 1954/1961.

[8] S. E. Asch, "Studies of independence and conformity: I. a minority of one against a unanimous majority.," *Psychological Monographs: General and Applied*, vol. 70, no. 9, pp. 1–70, 1956. DOI: 10.1037/h0093718.

[9] E. J. Webb, D. T. Campbell, R. D. Schwartz, and L. Sechrest, *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally, 1966.

[10] J. McCambridge, J. Witton, and D. R. Elbourne, "Systematic review of the hawthorne effect: New concepts are needed to study research participation effects," *Journal of Clinical Epidemiology*, vol. 67, no. 3, pp. 267–277, Mar. 2014. DOI: 10.1016/j.jclinepi.2013.08.015.

[11] F. J. Roethlisberger and W. J. Dickson, *Management and the Worker: An Account of a Research Program Conducted by the Western Electric Company, Hawthorne Works, Chicago*. Cambridge, MA: Harvard University Press, 1939.

[12] J. R. P. J. French, "Experiments in field settings," in *Research Methods in the Behavioral Sciences*, L. Festinger and D. Katz, Eds., New York: Holt, Rinehart and Winston, 1953, pp. 98–135.

[13] H. A. Landsberger, *Hawthorne Revisited*. Ithaca, NY: Cornell University, School of Industrial and Labor Relations, 1958.

[14] R. Larson and M. Csikszentmihalyi, "The experience sampling method," in *New Directions for Methodology of Social and Behavioral Science: Applications of Time Sampling Methods*, 1983, pp. 41–56.

[15] R. Larson and M. Csikszentmihalyi, "Experiential correlates of time alone in adolescence," *Journal of Personality*, vol. 46, no. 4, pp. 677–693, Dec. 1978. DOI: 10.1111/j.1467-6494.1978.tb00191.x.

[16] M. Csikszentmihalyi and R. Larson, "Validity and reliability of the experience-sampling method," *The Journal of Nervous and Mental Disease*, vol. 175, no. 9, pp. 526–536, Sep. 1987. DOI: 10.1097/00005053-198709000-00004.

[17] G. Miller, "The smartphone psychology manifesto," *Perspectives on Psychological Science*, vol. 7, no. 3, pp. 221–237, May 2012. DOI: 10.1177/1745691612441215.

[18] J. Fahrenberg, "Ambulantes Assessment. Computerunterstützte Datenerfassung unter Alltagsbedingungen," 1994. DOI: 10.23668/psycharchives.10951.

[19] M. Raento, A. Oulasvirta, and N. Eagle, "Smartphones: An emerging tool for social scientists," *Sociological Methods & Research*, vol. 37, no. 3, pp. 426–454, 2009.

[20] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay, "Myexperience: A system for in situ tracing and capturing of user feedback on mobile phones," in *Proceedings of the 5th International Conference on Mobile Systems, Applications, and Services (MobiSys)*, San Juan, PR: ACM, 2007, pp. 57–70. DOI: 10.1145/1247660.1247670.

[21] M. Mun, S. Reddy, K. Shilton, *et al.*, "Peir: The personal environmental impact report, as a platform for participatory sensing systems research," in *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services (MobiSys)*, Kraków, Poland: ACM, 2009, pp. 55–68. DOI: 10.1145/1555816.1555823.

[22] J. Kukkonen, E. Lagerspetz, P. Nurmi, and M. Andersson, "Betelgeuse: A platform for gathering and processing situational data," *IEEE Pervasive Computing*, vol. 8, no. 2, pp. 49–56, 2009. DOI: 10.1109/MPRV.2009.23.

[23] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, "Soundsense: Scalable sound sensing for people-centric applications on mobile phones," in *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services (MobiSys)*, Kraków, Poland: ACM, 2009, pp. 165–178. DOI: 10.1145/1555816.1555834.

[24] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, "Emotionsense: A mobile phones based adaptive platform for experimental social psychology research," in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp)*, Copenhagen, Denmark: ACM, 2010, pp. 281–290. DOI: 10.1145/1864349.1864393.

[25] J. Fahrenberg, R. Leonhart, and F. Foerster, *Alltagsnahe Psychologie mit hand-held PC und physiologischem Mess-System.* 2002. DOI: 10.23668/psycharchives.10415.

[26] D. C. Mohr, M. Zhang, and S. M. Schueller, "Personal sensing: Understanding mental health using ubiquitous sensors and machine learning," *Annual Review of Clinical Psychology*, vol. 13, no. 1, pp. 23–47, May 2017. DOI: 10.1146/annurev-clinpsy-032816-044949.

[27] N. Terzi, "The impact of e-commerce on international trade and employment," *Procedia - Social and Behavioral Sciences*, vol. 24, pp. 745–753, 2011. DOI: 10.1016/j.sbspro.2011.09.010.

[28] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, Jan. 2011. DOI: 10.1177/1745691610393980.

[29] S. Palan and C. Schitter, "Prolific.ac—a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, Mar. 2018. DOI: 10.1016/j.jbef.2017.12.004.

[30] T. S. Behrend, D. J. Sharek, A. W. Meade, and E. N. Wiebe, "The viability of crowdsourcing for survey research," *Behavior Research Methods*, vol. 43, no. 3, pp. 800–813, Mar. 2011. DOI: 10.3758/s13428-011-0081-0.

[31] J. Chandler, P. Mueller, and G. Paolacci, "Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers," *Behavior Research Methods*, vol. 46, no. 1, pp. 112–130, Jul. 2014. DOI: 10.3758/s13428-013-0365-7.

[32] M. G. Keith, L. Tay, and P. D. Harms, "Systems perspective of amazon mechanical turk for organizational research: Review and recommendations," *Frontiers in Psychology*, vol. 8, Aug. 2017. DOI: 10.3389/fpsyg.2017.01359.

[33] M. A. Webb and J. P. Tangney, "Too good to be true: Bots and bad data from mechanical turk," *Perspectives on Psychological Science*, p. 174 569 162 211 200, Nov. 2022. DOI: 10.1177/17456916221120027.

[34] M. G. Keith and A. S. McKay, "Too anecdotal to be true? mechanical turk is not all bots and bad data: Response to webb and tangney (2022)," *Perspectives on Psychological Science*, Mar. 2024. DOI: 10.1177/17456916241234328.

[35] D. Lazer, A. Pentland, L. Adamic, *et al.*, "Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009. DOI: 10.1126/science.1167742.

[36] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts,"

Political Analysis, vol. 21, no. 3, pp. 267–297, 2013. DOI: 10.1093/pan/mps028.

[37] M. Molina and F. Garip, "Machine learning for sociology," *Annual Review of Sociology*, vol. 45, no. 1, pp. 27–45, Jul. 2019. DOI: 10.1146/annurev-soc-073117-041106.

[38] I. Rahwan, M. Cebrian, N. Obradovich, *et al.*, "Machine behaviour," *Nature*, vol. 568, no. 7753, pp. 477–486, Apr. 2019. DOI: 10.1038/s41586-019-1138-y.

[39] M. Wölfel, *Immersive Virtuelle Realität: Grundlagen, Technologien, Anwendungen*. Springer Berlin Heidelberg, 2023. DOI: 10.1007/978-3-662-66908-2.

[40] O. Grau, *Virtual art* (Leonardo), en. London, England: MIT Press, Jan. 2003.

[41] C. Wheatstone, "Contributions to the Physiology of Vision.—Part the First. On Some Remarkable, and Hitherto Unobserved, Phenomena of Binocular Vision," *Philosophical Transactions of the Royal Society of London*, vol. 128, pp. 371–394, 1838. DOI: 10.1098/rstl.1838.0019.

[42] C. P. Comeau and J. S. Bryan, "Headsight television system provides remote surveillance," *Electronics*, vol. 34, no. 45, pp. 86–90, Nov. 1961.

[43] M. L. Heilig, "Sensorama simulator," US3050870A, https://patents.google.com/patent/US3050870A/en, Aug. 1962.

[44] M. L. Heilig, "Stereoscopic-television apparatus for individual use," *US Patent*, no. 2955156, Oct. 1960.

[45] I. E. Sutherland, "A head-mounted three dimensional display," in *Proceedings of the Fall Joint Computer Conference*, ACM, 1968, pp. 757–764. DOI: 10.1145/1476589.1476686.

[46] F. Steinicke, *Being Really Virtual*. Springer International Publishing, 2016. DOI: 10.1007/978-3-319-43078-2.

[47] I. E. Sutherland, "The ultimate display," in *Proceedings of the IFIP Congress*, vol. 2, New York, 1965, pp. 506–508.

[48] J. Lanier, *Dawn of the New Everything: Encounters with Reality and Virtual Reality*. New York: Henry Holt and Company, 2017.

[49] B. J. Frederick P., *Is there any real virtue in virtual reality?* Public lecture co-sponsored by the Royal Academy of Engineering and the British Computer Society, London, UK, Nov. 1998.

[50] M. Slater, M. Usoh, and A. Steed, "Taking steps: The influence of a walking metaphor on presence in virtual reality," in *Proceedings of ACM VRST*, 1999.

[51] S. Hayden. "Quest 3 s vs quest 3 vs quest 2 compared with detailed specs." (2024), [Online]. Available: https://www.roadtovr.com/quest-3s-quest-3-quest-2-specs-compared/.

[52] R. N. McDonnell, "Real-time image generation for a helicopter flight simulator," AIAA Flight Simulation Technologies Conference, Tech. Rep., 1997.

[53] M. Developers. "Unity performance recommendations for quest 3." (2024), [Online]. Available: https://developers.meta.com/horizon/documentation/unity/unity-perf/.

[54] D. Begault, E. Wenzel, M. Godfroy-Cooper, J. Miller, and M. Anderson, "Applying spatial audio to human interfaces: 25 years of nasa experience," *Proceedings of the AES International Conference*, Jan. 2011.

[55] B. Lang. "Vision pro and quest 3 hand-tracking latency compared." Accessed 15 July 2025, Road to VR. (2024), [Online]. Available: http

s://www.roadtovr.com/apple-vision-pro-meta-quest-3-hand-tracking-latency-comparison/.

[56] M. H. O. Developers. "Build believable mixed reality with depth api (quest 3)." (2023), [Online]. Available: https://developers.meta.com/horizon/blog/mesh-depth-api-meta-quest-3-developers-mixed-reality/.

[57] VR-Compare. "Virtual research vr4 vs meta quest pro—specification sheet." (2025), [Online]. Available: https://vr-compare.com/compare?h1=xUedlnDGs%5C&h2=-MpSqv-rB.

[58] Meta. "Meta quest 3," Meta. (), [Online]. Available: https://www.meta.com/de/quest/quest-3/.

[59] D. Bartelme, "Hand-built virtual databases for simulated environments," in *IEEE VR Annual International Symposium*, 1995.

[60] S. Pidhorskyi and M. Morehead, "Syglass: Interactive exploration of multidimensional images using vr hmds," *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[61] M. Slater and S. Wilbur, "A framework for immersive virtual environments (five): Speculations on the role of presence in virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 6, no. 6, pp. 603–616, Dec. 1997. DOI: 10.1162/pres.1997.6.6.603.

[62] M. Slater, "Measuring presence: A response to the witmer and singer presence questionnaire," in *Presence: Teleoperators & Virtual Environments*, MIT Press, vol. 8, 1999, pp. 560–565.

[63] M. Slater, "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3549–3557, 2009.

[64] M. Lombard and T. Ditton, "At the Heart of It All: The Concept of Presence," *Journal of Computer-Mediated Communication*, vol. 3, no. 2, JCMC321, Sep. 1997. DOI: 10.1111/j.1083-6101.1997.tb00072.x.

[65] C. J. Wilson and A. Soranzo, "The use of virtual reality in psychology: A case study in visual perception," *Computational and Mathematical Methods in Medicine*, vol. 2015, pp. 1–7, 2015. DOI: 10.1155/2015/151702.

[66] J. Short, E. Williams, and B. Christie, *The Social Psychology of Telecommunications*. London: John Wiley & Sons, 1976.

[67] M. Slater, D.-P. Pertaub, C. Barker, and D. M. Clark, "An experimental study on fear of public speaking using a virtual environment," *CyberPsychology and Behavior*, vol. 9, no. 5, pp. 627–633, Oct. 2006. DOI: 10.1089/cpb.2006.9.627.

[68] X. Pan and A. F. de C. Hamilton, "Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape," *British Journal of Psychology*, vol. 109, no. 3, pp. 395–417, Mar. 2018.

[69] O. D. Kothgassner and A. Felnhofer, "Does virtual reality help to cut the gordian knot between ecological validity and experimental control?" *Annals of the International Communication Association*, vol. 44, no. 3, pp. 210–218, Jul. 2020. DOI: 10.1080/23808985.2020.1792790.

[70] T. Dienlin, N. Johannes, N. D. Bowman, *et al.*, "An agenda for open science in communication," *Journal of Communication*, vol. 71, no. 1, pp. 1–26, Feb. 2020. DOI: 10.1093/joc/jqz052.

[71] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, 2015.

[72] H. T. Hunt, "Why psychology is/is not traditional science: The self-referential bases of psychological research and theory," *Review of General Psychology*, vol. 9, no. 4, pp. 358–374, Dec. 2005. DOI: 10.1037/1089-2680.9.4.358.

[73] M. Wölfel, D. Hepperle, C. F. Purps, J. Deuchler, and W. Hettmann, "Entering a new dimension in virtual reality research: An overview of existing toolkits, their features and challenges," in *2021 International Conference on Cyberworlds (CW)*, IEEE, Sep. 2021, pp. 180–187. DOI: 10.1109/cw52790.2021.00038.

[74] J. Grübel, "The design, experiment, analyse, and reproduce principle for experimentation in virtual reality," *Frontiers in Virtual Reality*, vol. 4, Apr. 2023. DOI: 10.3389/frvir.2023.1069423.

[75] D. Hepperle, T. Dienlin, and M. Wölfel, "Reducing the human factor in virtual reality research to increase reproducibility and replicability," in *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2021.

[76] M. Wölfel, D. Hepperle, C. F. Purps, J. Deuchler, and W. Hettmann, "Entering a new dimension in virtual reality research: An overview of existing toolkits, their features and challenges," in *2021 International Conference on Cyberworlds (CW)*, 2021.

[77] D. Hepperle and M. Wölfel, "Similarities and differences between immersive virtual reality, real world, and computer screens: A systematic scoping review in human behavior studies," *Multimodal Technologies and Interaction*, vol. 7, no. 6, 2023.

[78] D. Hepperle and M. Wölfel, "Exploring ecological validity: A comparative study of the mere exposure effect on screens and in immersive virtual reality," in *Advances in Visual Computing - ISVC 2024,*

LNCS 15047, G. Bebis, V. Patel, J. Gu, *et al.*, Eds., Springer Nature Switzerland AG, 2025.

[79] J. Deuchler, D. Hepperle, and M. Wölfel, "Asymmetric normalization in social virtual reality studies," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2022.

[80] K. Mrkva and L. V. Boven, "Salience theory of mere exposure: Relative exposure increases liking, extremity, and emotional intensity.," *Journal of Personality and Social Psychology*, vol. 118, no. 6, pp. 1118–1145, Jun. 2020. DOI: 10.1037/pspa0000184.

[81] D. Moher, A. Liberati, J. Tetzlaff, and D. G. A. and, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Medicine*, vol. 6, no. 7, e1000097, Jul. 2009. DOI: 10.1371/journal.pmed.1000097.

[82] R. M. Montoya, R. S. Horton, J. L. Vevea, M. Citkowicz, and E. A. Lauber, "A re-examination of the mere exposure effect: The influence of repeated exposure on recognition, familiarity, and liking.," *Psychological Bulletin*, vol. 143, no. 5, pp. 459–498, 2017. DOI: 10.1037/bul0000085.

[83] P. Wang and J. N. Bailenson, "Virtual reality as a research tool," *SSRN Electronic Journal*, 2024. DOI: 10.2139/ssrn.4805041.

[84] P. Wang, M. R. Miller, and J. N. Bailenson, "The belated guest: Exploring the design space for transforming asynchronous social interactions in virtual reality," in *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, IEEE, Mar. 2023, pp. 617–618. DOI: 10.1109/vrw58643.2023.00151.

[85] E. J. Ramirez, "Ecological and ethical issues in virtual reality research: A call for increased scrutiny," *Philosophical Psychology*, vol. 32, no. 2, pp. 211–233, Oct. 2018. DOI: 10.1080/09515089.2018.1532073.

[86] V. McIntosh, "Dialing up the danger: Virtual reality for the simulation of risk," *Frontiers in Virtual Reality*, vol. 3, Aug. 2022. DOI: 10.3389/frvir.2022.909984.

[87] E. Carl, A. T. Stein, A. Levihn-Coon, *et al.*, "Virtual reality exposure therapy for anxiety and related disorders: A meta-analysis of randomized controlled trials," *Journal of Anxiety Disorders*, vol. 61, pp. 27–36, Jan. 2019. DOI: 10.1016/j.janxdis.2018.08.003.

[88] C.-H. Lin, J. Gao, L. Tang, *et al.*, "Magic3d: High-resolution text-to-3d content creation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2023, pp. 300–309. DOI: 10.1109/cvpr52729.2023.00037.

[89] J. M. Loomis, J. J. Blascovich, and A. C. Beall, "Immersive virtual environment technology as a basic research tool in psychology," *Behavior Research Methods, Instruments, & Computers*, vol. 31, no. 4, pp. 557–564, Dec. 1999. DOI: 10.3758/bf03200735.

[90] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012.

[91] D. Hepperle, C. F. Purps, J. Deuchler, and M. Wölfel, "Aspects of visual avatar appearance: Self-representation, display type, and uncanny valley," *The Visual Computer*, vol. 38, no. 4, Jun. 2021.

[92] Y. Kim, Z. Aamir, M. Singh, S. Boorboor, K. Mueller, and A. E. Kaufman, "Explainable xr: Understanding user behaviors of xr environments using llm-assisted analytics framework," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 5, pp. 2756–2766, May 2025. DOI: 10.1109/tvcg.2025.3549537.

[93] A. D. Souchet, S. Philippe, D. Lourdeaux, and L. Leroy, "Measuring visual fatigue and cognitive load via eye tracking while learning with virtual reality head-mounted displays: A review," *International Journal of Human–Computer Interaction*, vol. 38, no. 9, pp. 801–824, Sep. 2021. DOI: 10.1080/10447318.2021.1976509.

[94] M. Vasser and J. Aru, "Guidelines for immersive virtual reality in psychological research," *Current Opinion in Psychology*, vol. 36, pp. 71–76, Dec. 2020. DOI: 10.1016/j.copsyc.2020.04.010.

[95] C. Maymon, Y. C. Wu, and G. Grimshaw, "The promises and pitfalls of virtual reality," in *Virtual Reality in Behavioral Neuroscience: New Insights and Methods*. Springer International Publishing, 2023, pp. 3–23. DOI: 10.1007/7854_2023_440.