Engaging Interaction

Designing for Immersive and Sustained User Experiences

Dissertation

by

Dmitry Alexandrovsky

Submitted in partial fulfillment of the requirements for the degree of:

Doctor of Engineering (Dr.-Ing.)

Faculty 3: Mathematics / Computer Science
University of Bremen
Bremen, Germany
Submission: April 16, 2021
Defense: August 3, 2021

1st Supervisor: Prof. Dr. Rainer Malaka (University of Bremen, Germany)
2nd Supervisor: Prof. Dr. Regan L. Mandryk (University of Saskatchewan, Canada)
External Reviewer: Assistant Prof. Dr. Kathrin Gerling (KU Leuven, Belgium)
AFFIRMATION

I hereby affirm, that I have

1. produced the doctoral thesis without unauthorized outside help,

2. that I have not used any sources or aids other than those indicated by me,

   and

3. that I have marked the places taken from the works used, either literally or in terms of content, as such.

__________________________________  ______________________
(Town, Date)                        (Sign)
Abstract

Engagement has become a fundamental research topic in Human-Computer Interaction (HCI). Extending the notion of traditional usability, HCI has focused on the hedonic properties of interactive systems. In particular, gameful design and multimodal interaction received much attention in the literature. However, despite a significant and growing body of research on engaging design, many interactive systems for learning, training and, health intervention suffer from low participation and massive attrition. The present work tackles this gap and investigates how interaction design can support a sustainable engagement with interactive systems. As engagement is a manifold construct that involves affective, cognitive and behavioral components. In this thesis, it is conceptualized from the perspectives of experience intensity and in terms of user behavior. These two perspectives are addressed in three threads of research: (i) game design for user engagement: effects of game elements on engagement, (ii) haptic interaction in engaging environments: effects of interaction modality on engagement, and (iii) assessment methods of user engagement in immersive environments: effects of embedded assessment methods on engagement. The thread on game design is twofold. This work presents the snacking framework, which consists of five game mechanics that facilitate a regular but brief play pattern. The snacking framework was first developed and evaluated using a casual game and then transferred onto a serious game. Adjacent to the snacking game mechanics, for a special case of serious games in the context of exposure therapy in Virtual Reality (VR), this work discusses an alternative approach to game design which employs an approach of playful user-generated content. The interaction design investigates the effects of haptic interaction on user engagement. This thread of research examines how static passive props both in VR as well as in the physical reality and interaction with sand – as a form of passive shape-changing props – in VR affect the user engagement. The meta-research on measurement methods developed and evaluated an approach that allows administering subjective self-reports in the form of questionnaires directly in the virtual environments. On a macro-level, these lines of research conceptualize the design for user engagement holistically and afford prescriptive design elements. On the micro-level, this dissertation extends existing theories of engagement and reveals how different design elements affect user behavior and the intensity of experiences with interactive systems.
Zusammenfassung

Acknowledgments

First, I thank Prof. Rainer Malaka, my doctoral supervisor, boss, and colleague, for his support and trust, the freedom and independence in my research, and great advice. I thank Prof. Regan L. Mandryk for being my secondary supervisor. I also feel gratefully honored for getting the chance to visit Prof. Mandryk’s lab and learning from her tremendous lore in research and her incredible personality. The time in Saskatoon will take a special place in my heart. Thank you, Prof. Kathrin Gerling, for taking the role of the external reviewer.

I thank my incredible colleagues, co-authors, and graduate students I was honored to supervise. I am thankful for the fruitful collaborations, the inspiring discussions, and your support. Thank you for taking part in the projects that contributed to this dissertation. The collaboration with you was always inspiring and insightful. In alphabetical order: Katya Alejandra Alvarez Molina, Angelika Bükner-Ahsbahs, Mehrdad Bahrini, Ioannis Bikas, Max Birk, Stefan Bollen, Michael Bonfert, Bastian Dänkas, Tanja Döring, Stefan Finke, Thomas Fröhlich, Max Friehs, Jendrik Grittner, Supara Grudpan, Marc Herrlich, Sebastian Höflner, Yenchin Lai, Pitt Michelmann, Thomas Mildner, Thomas Mündner, Denis Obrezkov, Johannes Pfau, Frederic Pollmann, Robert Porzel, Felix Putze, Susanne Putze, Anke Reinschlüssel, Nat Sararit, Johannes Schöning, Peter Schulz, Jan Smeddinck, Max Spliethöfer, Timo Stabbert, Georg Volkmar, Benjamin Walther-Franks, Dirk Wenig, Nina Wenig, Rowan Yates, Ambreen Zaman, Nima Zargham. A special thanks goes to the DMLab staff members Irmgard Laumann, Gerald Volkmann, Svenja Voss, Evgenia Sazonkina, and Insa Warms. You always backed me up and resolved barriers in the best imaginative way.

Я особенно благодарен моей семье, моим родителям Татьяне Стороженко и Леониду Александровскому, моим братьям Михаилу и Владимиру Александровским, а также, Юлии Эйдиновой и моей бабушке Софье Александровской. Спасибо за ваше терпение, за ваш интерес и вашу поддержку.

Dear friends, Alissa, David, Ivana, Jule, Johann, Konrad, Lina, Rosa, I’m thankful to have such a lovely family of choice where I can always find warmth and deep friendship. Thank you for always keeping my spirit up and for your patience when work had priority. My inmost gratitude is owed to my partner Yvonne Leite Remondes. Thank you for your love, patience, and support during all the years. You were always supportive when things didn’t go as expected and understanding when deadlines or papers had put other things else aside.

You all made this episode of my life unforgettable!
## Contents

List of Figures ix  
List of Tables xi  

1 Introduction 1  
1.1 Contributions ........................................ 2  
1.2 Thesis Outline ........................................ 5  

2 Background 7  
2.1 User Experience ........................................ 7  
2.2 Experience in Virtual Environments ...................... 8  
2.3 Fun, Pleasure, and Enjoyment .......................... 10  
2.4 Player Experience ....................................... 11  
2.5 Engaging Interaction Design ......................... 14  
2.6 User Engagement ........................................ 16  
2.7 Measuring Experiences .................................. 18  
2.8 Summary .................................................. 19  

3 Game Design for User Engagement 21  
3.1 Designing for Snacking Behavior ...................... 21  
3.2 Player Attrition: Snacking in Serious Games ............ 26  
3.3 Playful User-Generated Treatment ...................... 33  
3.4 Discussion ............................................... 39  

4 Haptic Interaction in Engaging Environments 41  
4.1 Passive Physical Props for VR Climbing .................. 41  
4.2 Passive Physical Props for Math-learning ................ 44  
4.3 Creative Interaction Sand in VR ...................... 46  
4.4 Discussion ............................................... 47  

5 Assessing User Engagement in Immersive Environments 49  
5.1 In-VR Questionnaire Design ............................ 49  
5.2 Effects of Question-asking in VR ...................... 53  
5.3 Discussion ............................................... 57  

6 Discussion and Conclusion 59  

7 Limitations and Future Work 63  

Bibliography 65  

Acronyms 87  

Fundamental Publications 91  
F₁ Game Dynamics That Support Snacking, Not Feasting .................. 91  
F₂ Serious Snacking: A Survival Analysis of How Snacking Mechanics Affect Attrition in a Mobile Serious Game .................. 91  
F₃ Playful User-Generated Treatment: A Novel Game Design Approach for VR Exposure Therapy .......................... 92
## Contents

F.4 The Role of Physical Props in VR Climbing Environments ................................................. 92  
F.5 VRBox: A Virtual Reality Augmented Sandbox for Immersive Playfulness, Creativity  
and Exploration .................................................................................................................. 92  
F.6 Examining Design Choices of Questionnaires in VR User Studies .................................. 93  
F.7 Breaking The Experience: Effects of Questionnaires in VR User Studies ..................... 93  

### Supportive Publications

S.1 Exploring Interactive Systems for Algebra Learning in School ........................................ 95  
S.2 Demonstrating VRBox: A Virtual Reality Augmented Sandbox ....................................... 95  
S.3 Playful User-Generated Treatment: Expert Perspectives on Opportunities and Design  
Challenges ......................................................................................................................... 95  
S.4 Evaluating User Experiences in Mixed Reality ............................................................... 96
# List of Figures

1.1 Threads of research affecting the UE by O’Brien and Toms (2008) .......................... 3
2.1 Reality-Virtuality Continuum (Milgram and Jr, 1999) ........................................ 8
2.2 Components of PX by Wiemeyer et al. (2016) ...................................................... 12
2.3 Flow Model and PX Model .......................................................... 13
2.4 Space of Directness (Hutchins et al., 1985) .................................................. 14
2.5 RBI (Jacob et al., 2008) .......................................................... 16
2.6 Stages of UE (O’Brien and Toms, 2008) .................................................. 19
3.1 GOC game view (Alexandrovsky et al., 2019b, F1) .............................................. 23
3.2 GOC performance (Alexandrovsky et al., 2019b, F1) ........................................... 25
3.3 GOC performance (Alexandrovsky et al., 2019b, F1) ........................................... 26
3.4 GOC clusters (Alexandrovsky et al., 2019b, F1) ................................................... 27
3.5 Infinitus Algebraica gameplay (Alexandrovsky et al., 2021a, F2) ............................. 28
3.6 Infinitus Algebraica long-term play times (Alexandrovsky et al., 2021a, F2) .......... 32
3.7 Infinitus Algebraica survival functions (Alexandrovsky et al., 2021a, F2) .............. 34
3.8 PUT screenshots (Alexandrovsky et al., 2020b, F3) ............................................ 36
3.9 PUT results (Alexandrovsky et al., 2020b, F3) ..................................................... 38
4.1 Climbing in VR setup (Schulz et al., 2019, F4) ................................................... 42
4.2 Climbing in VR subjective responses (Schulz et al., 2019, F4) ............................... 43
4.3 Climbing in VR physiological responses (Schulz et al., 2019, F4) ........................... 44
4.4 Algebra Tiles study setup and UMUX results (Alexandrovsky et al., 2018, S1) ....... 45
4.5 VRBox v1 setup (Fröhlich et al., 2018, F5) ....................................................... 46
4.6 VRBox v2 setup (Alexandrovsky et al., 2019a, S2) ............................................. 48
5.1 inVRQs subjective responses (Alexandrovsky et al., 2020a, F6) ............................. 50
5.2 Archery game and in-VR questionnaire (inVRQ) (Alexandrovsky et al., 2020a, F6) .. 51
5.3 Shooting game and inVRQ (Putze et al., 2020, F7) ............................................ 53
5.4 Study flow (Putze et al., 2020, F7) .......................................................... 54
5.5 PENS and IPQ results (Putze et al., 2020, F7) ................................................... 55
5.6 BIP and EDA responses (Putze et al., 2020, F7) ................................................... 58
6.1 Triggers of Engagement .......................................................... 59

ix
List of Tables

1.1 Contributions of this Thesis .................................................... 4

3.1 GOC pre-study IMI and PENS (Alexandrovsky et al., 2019b, F1) ........... 24
3.2 GOC long-term study demographics (Alexandrovsky et al., 2019b, F1) ...... 24
3.3 GOC performance and behavior correlations (Alexandrovsky et al., 2019b, F1) 25
3.4 GOC cluster analysis (Alexandrovsky et al., 2019b, F1) .................... 26
3.5 Infinitus Algebraica pre-study demographics (Alexandrovsky et al., 2021a, F2) 29
3.6 IMI pre-study descriptives (Alexandrovsky et al., 2021a, F2) ............... 29
3.7 IMI pre-study ANOVA (Alexandrovsky et al., 2021a, F2) .................... 30
3.8 Infinitus Algebraica pre-study performance (Alexandrovsky et al., 2021a, F2) 30
3.9 Infinitus Algebraica long-term demographics (Alexandrovsky et al., 2021a, F2) 31
3.10 Infinitus Algebraica long-term IMI and SIMS (Alexandrovsky et al., 2021a, F2) 33
3.11 Infinitus Algebraica median survival times (Alexandrovsky et al., 2021a, F2) 33
3.12 Infinitus Algebraica statistics and estimates of the Cox-regression (Alexandrovsky et al., 2021a, F2) ........................................ 35
3.13 PUT anxiety measures (Alexandrovsky et al., 2020b, F3) ................. 37
3.14 IMI results (Alexandrovsky et al., 2020b, F3) ................................ 38

5.1 inVRQs design study statistics (Alexandrovsky et al., 2020a, F6) .......... 51
5.2 inVRQs user study results (Alexandrovsky et al., 2020a, F6) ............... 52
5.3 inVRQs physiological responses (Putze et al., 2020, F7) .................... 56

6.1 Classification of the contribution according to the dimensions of engagement ..... 61
Over the past decades, Human-Computer Interaction (HCI) has turned from functional-oriented usability goals such as effectiveness and efficiency (ISO, 2006; Preece et al., 2015) towards non-utilitarian hedonic facets of interaction (Attfield et al., 2011; Bargas-Avila and Hornbæk, 2011; Hassenzahl et al., 2003; Hassenzahl and Tractinsky, 2006; Olson and Olson, 2003). Traditionally, user satisfaction – which is a central component of usability (Nielsen, 1994; Nielsen, 2012) – is considered an indicator of high productivity in using an interactive system (Jegers, 2007). However, as Hassenzahl and Tractinsky (2006) constitute, a central objective of HCI has become the design of interactive systems for pleasure and to improve the quality of life rather than to design for the “absence of pain”. Koivisto and Hamari (2019) describe this trend as a convergence of utilitarian and hedonic interactive systems. Along with utilitarian systems developed solely to improve the productivity of work, always emerged applications for sole enjoyment, i.e., games. Koivisto and Hamari (2019) point out that first hedonic video-games were a “playful re-appropriation of instrumental information technology”. Games are interactive media that are designed solely for enjoyment (Mekler et al., 2014; Nabi and Krcmar, 2004; Sweetser and Wyeth, 2005). Blythe and Hassenzahl constitute that “[b]esides computer games, enjoyment, fun, pleasure was absent when it came to technology” (Blythe and Hassenzahl, 2018). As researchers and designers recognized that hedonic qualities could promote the productivity and enjoyment in using interactive systems (Abeele et al., 2020; Blythe and Hassenzahl, 2018; Nacke et al., 2014), HCI witnesses the completion of the “convergence cycle” where research on playful experience has been re-appropriated and now also informs the design of systems for production (Koivisto and Hamari, 2019) or behavioral change (Mandryk and Birk, 2017). Particularly for behavior change applications in the domain of health, literature frequently reports high drop-off rates from the interventions (Eysenbach, 2005; Mandryk and Birk, 2017; Perski et al., 2017). A similar phenomenon can be observed in the domain of learning applications (Gütl et al., 2014). Frequently, the literature argues for gameful design to promote adherence with the interventions (Mandryk and Birk, 2017; Orji et al., 2013). To date, literature established gameful design as a warranted strategy to improve sustained engagement. However, how different interactive and gameful elements facilitate user engagement is not well-understood (Boyle et al., 2012; Landers et al., 2019; Mekler et al., 2014). For example, wrongly applied game design or rewarding mechanisms can lead to a reversed effect and undermine the students’ intrinsic motivation in (gamified) learning environments (Deci et al., 2001; Hanus and Fox, 2015; Johnson-Glenberg and Megowan-Romanowicz, 2017). Moreover, while a growing body of research showed evidence for the motivational pull of gameful design (cf. Hamari et al., 2014), literature reports that in many instances the employed playful/gameful interactive systems it is not effective in fighting long-term attrition (Perski et al., 2017). Although multiple theories of engaging design exist, the research is fragmented (Attfield et al., 2011; Perski et al., 2017) and a unifying theory of engaging design is missing (Landers et al., 2019). To address this gap, this dissertation investigates and identifies design patterns for engaging interaction with games and Virtual Environments (VEs). It aims to understand the challenges of – and provide guidance on – engaging interaction design for immersive interactive systems with a purpose. This work promotes a perspective shift in existing approaches on engagement research and contributes to the groundwork for a future unifying theory of User Engagement (UE) in VEs. This body of work is guided by the overarching research theme:

**How should interactive systems be designed to facilitate sustained user engagement?**

As a theoretical foundation of engagement, this dissertation builds on O’Brien and Toms’ (2008) en-
gagement cycle, which describes how interactive systems are used and what qualities of the experiences are addressed. The engagement cycle consists of four stages: point of engagement, period of engagement, disengagement, and re-engagement. These four stages conceptualize UE holistically and provide anchor points for the design of sustainable UE. More details on O’Brien and Toms’s UE are discussed in Chapter 2.6. To address all stages of the UE cycle (O’Brien and Toms, 2008), engaging interaction design requires a consideration of multiple key aspects. Attfield et al. (2011) identified three main directions for research agenda on UE: (i) “interaction patterns and development of engagement measures” (ii) “adaptation of immersion concepts from gaming” (iii) “designing for user engagement”. According to Attfield et al., these lines of research should not be considered independent, but rather, they cover most of the relevant aspects to define and measure UE. Therefore, following Attfield et al.’s (2011) threads of research, this work takes multiple perspectives both from interaction design as well as Games User Research (GUR) and investigates engaging game design elements, haptic interaction design in engaging environments, and measurement methods in immersive environments to instruct the design for engaging experiences.

RQ1: From the game design perspective, GUR provides a solid foundation of knowledge on how game design affects short-term motivation (Hamari et al., 2014). However, how individual game elements affect player engagement or how to design for the long-term adherence of games with a purpose is rarely discussed in the literature (Jung et al., 2020; Landers et al., 2019; Mekler et al., 2014; Nacke and Drachen, 2011). Therefore, RQ1 investigates: How can game design strengthen the emotional/psychological bonding between players and interactive system to improve the long-term adherence?

RQ2: The rise of novel reality-altering technology, such as VR and Augmented Reality (AR) opens up research and a design landscape of previously impossible training and intervention methods. Simultaneously, it also faces HCI with a multitude of challenges for engaging interaction design (Mekler et al., 2014). Considering novel, embodied, and multimodal interaction techniques such as gestures, body tracking, or voice interaction, a growing body of work guides such interaction design. However, the effects of embodied interaction (e.g., haptics) on performance and subjective experience are not completely understood. To better understand the impact of direct interaction, RQ2 examines: How does haptic interaction affect user engagement in VEs?

RQ3: Evaluation of experiences with interactive systems is central in HCI, as it allows to infer how the design affects the users, and usually, the evaluation results guide the design direction of subsequent iterations of the interactive systems. Most commonly, the evaluations rely on standardized measurement tools. However, regarding the measurement methods of UE and related constructs that describe the experiential characteristics of an interactive system, it is not well-known if established standardized measurement methods of UE from the physical reality transfer well to VR (Nacke and Drachen, 2011) as (a) traditional metrics of usability lose their relevance and (b) with a stronger immersion, a break of the experience would have a stronger negative impact on the users’ affect (cf. Brown and Cairns, 2004; Slater et al., 2003; Slater and Steed, 2000). To improve the research methods of user studies in VEs, RQ3 investigates: How should existing evaluation methods of user engagement be adapted for the measurement of experiences in VEs?

1.1 Contributions

This work investigates how VEs and gameful systems (Landers et al., 2019) for training or intervention can be improved to facilitate User Engagement (UE). The present dissertation is a thesis by published work. It consists of seven fundamental peer-reviewed publications (F1-F7) which build the conceptual as well as practical contributions of this work and four supportive publications (S1-S4) that extend the foundational contributions. These contributions aim to provide tangible and prescriptive design elements to improve UE with interactive systems. This dissertation builds on Attfield et al.’s (2011) directions for engaging design which encompass the adaption of engaging characteristics from game
1.1 Contributions

Figure 1.1: Threads of research affecting O’Brien and Toms’ (2008) stages of engagement. Each thread contributes target points to facilitate User Engagement (UE).

design, engaging interaction patterns, and measurement methods for engagement and decomposes engagement in accordance to O’Brien and Toms’ (2008) four stages of engagement. This approach allows to investigate how individual design elements and patterns affect components of UE. Figure 1.1 illustrates on what stages of UE the investigated design patterns operate. Table 1.1 summarizes the contributions of this thesis and depicts the applied methods to investigate the research questions.

**RQ1:** An established method to facilitate engagement is gameful/playful design. It has been shown that gamification (Deterding et al., 2011) and serious games (Abt, 1970) are warranted methods that can improve at least short-term engagement (Dale, 2014; Hamari et al., 2014; Mora et al., 2015). However, how individual game elements affect user engagement is not clear (Landers et al., 2019; Tondello and Nacke, 2019; Tondello et al., 2017). Moreover, although promising mechanics exist, it is unclear how they should be implemented to facilitate intended player behavior (Boyle et al., 2012); for example, to adhere to a health intervention. Three publications from this line of research investigate what game design elements can strengthen the emotional/psychological bonding between players and interactive systems. First, based on established game elements in casual games, we developed the snacking framework consisting of five game mechanics that facilitate a brief but regular player behavior (the snacking pattern). Next, we transferred the snacking framework into a serious game. Finally, this work investigated game mechanics beyond functional challenges (Cole et al., 2015) – game elements that build around the motor or cognitive skills – and how playfulness can be embedded into Virtual Reality Exposure Therapy (VRET) – a prototypical mental health intervention in VR by addressing creativity and self-empowerment.

**RQ2:** A pivotal metric of engagement in virtual environments is the sense of presence facilitated through immersive characteristics of the interactive systems such as degree of realism or “naturalness” of the interaction. Haptic interaction is an effective method to facilitate immersion in VR (Johnson-Glenberg and Megowan-Romanowicz, 2017). However, literature has no complete picture of which characteristics of User Experience (UX) are affected by haptic interaction. Therefore, although the literature shows strong evidence that haptic interaction improves the overall experience, it is rarely found in end-user products. If researchers and designers could better understand what aspects of UX are affected by haptic interaction, this would give them more clear guidance for the design of a particular experience, allowing them to design for sustained UE more effectively. This line of research investigated in three exemplary scenarios of full-body interaction with passive props in VR, passive
Table 1.1: Table of Contributions.

<table>
<thead>
<tr>
<th>Thread of Research</th>
<th>Methods</th>
<th>Affected Stages of Engagement</th>
<th>Contributions</th>
<th>Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1: Engaging Game Design</td>
<td>Empirical user studies including subjective and objective measures of experience and behavior, content analysis, iterative design, expert interviews.</td>
<td>All stages</td>
<td>• Development and evaluation of a game design framework with 5 mechanics that address player behavior at different points of the engagement cycle.  • Extension of game design strategies for emotional challenges with creativity mechanics that address Attention, Control, Creativity, and Interest empower players and strengthen the relationship with the VE.</td>
<td>“Casual Snacking” (Alexandrovsky et al., 2019b, F1), “Serious Snacking” (Alexandrovsky et al., 2021a, F2), “Playful User-Generated Treatment” (Alexandrovsky et al., 2020b, F3), “Playful User-Generated Treatment: Interviews” (Volkmar et al., 2020, S3)</td>
</tr>
<tr>
<td>RQ2: Haptic interaction in engaging environments</td>
<td>Empirical user studies including subjective and physiological measures, development of artifacts.</td>
<td>Point of Engagement</td>
<td>• Validation of theories on UE from the non-VR context: Haptic interaction facilitates Aesthetics, Novelty, Sensory Appeal, Motivation, Sense of Realism, Positive and Negative Affect, and Creativity.  • Artifacts that exemplify the interaction space of passive haptics in VR and physical reality.</td>
<td>“Physical Props in VR” (Schulz et al., 2019, F4), “Math Tangibles” (Alexandrovsky et al., 2018, S1), “VRBox” (Fröhlich et al., 2018, F5), “VRBox V2” (Alexandrovsky et al., 2019a, S2)</td>
</tr>
<tr>
<td>RQ3: Measuring UX in VR</td>
<td>Empirical user studies including subjective and physiological measurements, expert online survey, literature review.</td>
<td>Disengagement, Re-engagement</td>
<td>• Identification of reality-change in user studies as a potential bias.  • Development of usability guidelines for inVRQs.  • Improvement of the flow for VR user studies through minimizing disruption and thus, bringing subjective measurements closer to the experience.  • Artifacts that exemplify the approach.</td>
<td>“InVRQ Design” (Alexandrovsky et al., 2020a, F6), “InVRQ BIP” (Putze et al., 2020, F7), “Evaluating User Experiences in Mixed Reality” (Alexandrovsky et al., 2021b, S1)</td>
</tr>
</tbody>
</table>

haptic props in a mobile learning environment, and passive shape-changing props in VR for tabletop construction. Findings from these publications reveal how haptic interaction affects the UX giving researchers and practitioners of interactive systems guidance on how to modulate the intensity of an experience concerning UE.

RQ3: Throughout this research in VR, we noticed that the administration of post-experience questionnaires outside VR is often tedious and can break the study flow; especially in cases of repeated assessments. Therefore, we aimed at conducting the questionnaires in VR. However, not knowing if this assessment method is warranted for VR user studies since possible biases and lack of comparability with other studies might be serious concerns of this assessment method. This leads to the third research thread, which investigates if question-asking in VR is a warranted method and, if yes, how it should be designed? This research question is addressed in two publications: First, investigating the interaction and usability design of in-VR questionnaires (inVRQs) and secondly examining the effect of question-asking in VR on the sense of presence.

On a macro-level, these lines of research, which manifest in the three research questions, are targeting O’Brien and Toms’ four stages of engagement and aim to provide prescriptive design choices that researchers and designers can implement in their interactive systems to overcome attrition and therefore
improve the effectiveness of the interventions (cf. Fig. 1.1). On the micro-level, each of these research threads extends existing theories around experience research and improves the understanding of how facets of the experience are affected by individual design elements.

While acknowledging that external factors such as environmental context or personality traits play an important role in how people engage with technology (Hassenzahl, 2008; Mekler et al., 2014; Nacke and Drachen, 2011; Olson and Olson, 2003; Perski et al., 2017), they are beyond the scope of this dissertation. External factors are often very specific in different scenarios and might not necessarily generalize well to provide prescriptive guidelines for the design of engaging interactive systems, as they require individual considerations that might not be directly related to distinct designs of the interactive systems. However, the presented contributions in this thesis offer several anchor points to adjust the design of the interactive systems to specific needs.

### 1.2 Thesis Outline

The remaining thesis is outlined as follows: Chapter 2 establishes some central vocabulary around the design of experiences with interactive systems and briefly discusses different types of experiences in HCI. The following 3 Chapters constitute the main contributions of the thesis. Chapter 3 discusses game design that aims to facilitate player engagement. It presents a framework of game mechanics that support player adherence and introduces a game design approach for mental health games. Chapter 4 presents a series of contributions in the domain of haptic interaction in VEs and shows how haptic interaction affects the intensity of the experiences. Chapter 5 discusses evaluation methods in VR and presents the design as well as the evaluation of in-VR questionnaires (inVRQs) as a design method that minimizes the disruption from the experience during VR user studies. Next, Chapter 6 aggregates the results from the presented contributions and draws a conclusion in light of the overarching research theme on UE. Finally, Chapter 7 discusses the limitations of this work and outlines directions for future research.
HCI research around experiences with interactive systems developed a broad vocabulary to describe different aspects of the users’ cognitive, affective, and behavioral responses to interaction. However, often the terms are used interchangeably, which manifests a vague understanding in this field. To shed light on the language around hedonic experiences with interactive technology, this chapter provides an overview of some foundational research in this field, which clarifies the terminology used throughout this thesis. First, Section 2.1 establishes factors of User Experience (UX) as it is a core concept in HCI to describe how users perceive the interaction with technology. Next, Section 2.2 discusses presence and immersion as key constructs related to experiences in VEs. To discriminate different types of joyful experiences, Section 2.3 looks at different terminology around enjoyment. Since games differ drastically from utilitarian interactive systems, Section 2.4 defines Player Experience (PX) and discusses its differences between the experiences with pragmatic and hedonic interactive systems. Building on the established terms, Section 2.6 defines User Engagement (UE) and differentiates it from UX and PX. Further, Section 2.7 discusses measurement methods of experiences with interactive systems. Finally, 2.8 summarizes the related work and discusses the visited concepts concerning the objectives of the thesis.

2.1 User Experience

User Experience (UX) is a central concept in HCI (Hassenzahl and Tractinsky, 2006; Law et al., 2009). It is referred to as the consequence of users interacting with digital artifacts (i.e., interactive systems), which encompass the internal state of the user (psychological, cognitive), the characteristics of the system, and the context in which the interaction takes place (Hassenzahl and Tractinsky, 2006; ISO, 2019; Preece et al., 2015). All these outcomes describe different facets of a satisfactory and sustained experience (Dix, 2009). In contrast to usability (Nielsen, 1994), UX highlights the non-utilitarian facets of contact with interactive technology (Law et al., 2009). Hassenzahl (2008) “define[s] UX as a momentary, primarily evaluative feeling (good-bad) while interacting with a product or service” (Hassenzahl, 2008) with an emphasis on the dynamic nature of the experience, which remains present-oriented but variable over time. In their earlier work, Hassenzahl et al. (2003) developed the AttrakDiff questionnaire as a measurement tool of attractiveness for interactive systems. Interestingly, Hassenzahl et al. identified two types of hedonic qualities (identity and stimulation) and one subscale of pragmatic qualities. In the questionnaire, each item is represented as a pair of two opposing qualities. The hedonic qualities of identity describe aspects of what users identify as attractive. Some example pairs of this scale are tracky – stylish, cheap – premium, or alienating – integrative. In contrast, the stimulation hedonic qualities characterize attributes that trigger attraction. Example item pairs for this scale are isolating – connecting, dull – captivating, or ordinary – novel. The pragmatic qualities are closely related to the traditional notion of usability (ISO, 2006) with item pairs such as complicated – simple, impractical – practical, or unpredictable – predictable. Similarly, to quantify UX, Laugwitz et al. (2008) developed the User Experience Questionnaire (UXQ), a 26-item questionnaire with the subscales Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty. These scales combine both pragmatic and hedonic qualities of the experience and are adjacent to Hassenzahl’s comprehension of UX. Hassenzahl (2008) attribute the pragmatic qualities of a system to the do-goals, which describe the objectives
2 Background

Mixed Reality (MR)

Augmented Reality (AR)

Augmented Virtuality (VR)

Real Environment (RE)

Virtual Environment (VE)

Reality-Virtuality (RV) Continuum

Figure 2.1: Reality-Virtuality Continuum (Milgram and Jr, 1999).

related to the functionality of the interactive system. The hedonic qualities, Hassenzahl summarizes as properties of a interactive system that facilitate the be-goals, which describe how the system supports the fulfillment of human needs beyond utilitarian outcomes such as novelty, challenges, personal growth, or self-expression. Hassenzahl (2008) concludes that hedonic qualities are at the core of positive experiences while pragmatic qualities indirectly affect the experience through lowering the hindrance of fulfilling meaningful user needs. To determine how human needs of satisfaction relate to the hedonic properties, Hassenzahl (2008) correlated the needs of satisfaction Autonomy, Competence, and Relatedness from Self-Determination Theory (SDT) (Ryan and Deci, 2000b) – a commonly applied theory to describe the experiences with games (Tyack and Mekler, 2020) – with Positive and Negative Affect Scale (PANAS) (Crawford and Henry, 2004) as a source of positive experience with interactive systems. Autonomy and Competence were mostly linked to positive affect but not Relatedness. Relatedness was found to be associated with the negative affect of distress. However, the authors reckon that the negative effect of Relatedness is most likely related to the limitations of the existing technology. Interestingly, the analysis reveals that the different dimensions of motivation facilitate diverse affective qualities of UX. Further, Hassenzahl highlight that none of the dimensions of SDT is correlated with excitement and enthusiasm, which are central hedonic qualities of UX. These results might be a consequence of the fact that UX was mainly designed to assess the hedonic and pragmatic qualities of interaction with utilitarian systems or web products. Further, these results underline that playful experiences and forms of interaction beyond the desktop such as VR are not well covered by the traditional notion of UX. Therefore, other aspects of psychometric constructs are required to better understand the experiences with these types of interactive systems.

2.2 Experience in Virtual Environments

Reality-altering technology poses a new perspective on what is considered “real” as the two extremes of real reality and virtual reality start to blend into a so called Mixed Reality (MR). Milgram and Jr (1999) conceptualize MR on a reality-virtuality continuum which arranges the forms of reality alteration dependent on the degree of augmentation (cf. Fig. 2.1). One extreme is the Real Environment where no digital augmentation takes place. On the other end of the spectrum is the Virtual Environment (VE) with no content from the physical world in it. Moving on the spectrum from the real environment pole towards the virtual pole augments more physical world with virtual content. Moving in the opposite direction from the virtual environment toward the real environment extends the computer-generated world with real content.

Two constructs that are commonly used to describe the quality characteristics of MR (i.e., VR and AR) experiences are the sense of presence and immersion (Dinh et al., 1999; Hodges et al., 1994; Sanchez-Vives and Slater, 2005; Wiemeyer et al., 2016; Witmer and Singer, 1998). Presence is commonly defined as the subjective experience of being in the virtual environment rather than in the physical space (Draper et al., 1998; Minsky, 1980; Skarbez et al., 2017; Slater, 2007; Witmer and Singer, 1998).
Similarly, Lombard and Ditton identify presence as the disappearance of the communication medium – a “perceptual illusion of nonmediation” (Lombard and Ditton, 1997), which is consistent with Bystrom et al.’s suspension of disbeliefs (Bystrom et al., 1999). Slater et al. extend this definition and describe presence as “how well a person’s behavior in the VE matches their behavior in similar circumstances in real life” (Slater et al., 1996). They argue that how data is displayed – and how the participants can interact in VR – is more important for presence than the level of realism (Sanchez-Vives and Slater, 2005; Slater, 2007; Zahorik and Jenison, 1998). IJsselsteijn et al. (2000) adapt Lombard and Ditton’s definition of and distinguish two subcategories of physical and social presence. Physical presence refers to the sense of a physical location, and social presence describes the feeling of “being together”. Further, the authors provide four determinants of presence: extent and fidelity of sensory information, the match between sensors and display, content factors, and user characteristics. Likewise, Sheridan (1992) proposed three determinants of presence: the extent of sensory information, control of relation of sensors to the environment, and ability to modify the physical space and Witmer and Singer (1998) defined four factors that contribute to the sense of presence: control, sensory, distraction, and realism. The immersion, presence, performance model (Bystrom et al., 1999) follows Slater et al.’s definition (Slater et al., 1996) and proposes a feedback loop between attention, engagement, and presence which also underlies the flow theory (Csikszentmihalyi, 1990). While there are diverse definitions of presence across the literature (cf. Heeter, 1992; Lombard and Ditton, 1997; Slater, 2003; Witmer and Singer, 1998), there is a consensus that presence is a multidimensional construct that is driven by media characteristics, such as technological factors (Lombard et al., 2000; Sas and O’Hare, 2003; Skarbez et al., 2017; Welch et al., 1996), and personal characteristics (Baños et al., 2004). Skarbez et al. (2017) provide a comprehensive review of literature on presence, the authors identify similarities between the definitions and aggregate common variables and related constructs that contribute to the sense of presence into a conceptual model of Place Illusion, Plausibility Illusion, and Social Presence Illusion. Predominantly, presence is assessed using post-experience questionnaires (Schwind et al., 2019). However, the act of assessing it breaks the illusion and directs the users’ consciousness back to the physical reality (Slater and Steed, 2000) and thus, compromises the psychological state the questionnaire tries to assess (Sanchez-Vives and Slater, 2005; Slater, 2004; Slater and Steed, 2000). Therefore, Slater (2004) and Slater and Steed (2000) argue to move away from assessing presence through questionnaires, but rather by looking at events when the illusion generated by the VE collapses. These events are referred to as Breaks in Presence (BIPs) (Slater et al., 2003; Slater and Steed, 2000). A BIP occurs when the users experience inconsistencies between their mental model and the VE (Liebold et al., 2017). BIPs have been associated with disorientation and negative emotions (Knibbe et al., 2018; Scherer and Ellgring, 2007; Schwind et al., 2019; Slater et al., 2003). Examples for BIPs are loss of tracking, glitches, or noises outside the VE (Jerald, 2016).

In contrast to presence, immersion is recurrently described as the objective properties of the VE and the applied technology that induces the sense of presence (Bowman and McMahan, 2007; Slater et al., 1996; Slater and Wilbur, 1997). Immersion and presence are two logically separated constructs; however, they are directly related (Slater, 2007) as presence is the outcome of immersion (Schubert et al., 2001). Immersion includes the software and the hardware components that produce stimuli to the user’s senses and affect how they perceive the VE (Regenbrecht et al., 1998). Stereoscopic rendering, resolution, frame rate, the field of view, levels of user tracking, and fidelity of sensory input are considered as driving factors that enable immersion and thus, facilitate the sense of presence (Bowman and McMahan, 2007; Cummings and Bailenson, 2016; Gruchalla, 2004; Hendrix and Barfield, 1996; Meehan, 2001; Slater et al., 1994; Welch et al., 1996). From a content analysis of 83 studies, Cummings and Bailenson (2016) infer that technological immersion has a medium-sized effect on presence. Brown and Cairns (2004) identified engagement, engrossment, and total immersion as three levels of immersion in games on a scale of involvement. Witmer and Singer (1998) conclude that both involvement and immersion are vital to invoke the sense of presence.

Literature has argued that besides the positive impact of presence on the experience, it is also affecting the users’ performance in task execution (Barfield et al., 1995; Slater et al., 1996; Youngblut and Huie, 2003). Therefore, it can be argued that by improving the immersive characteristics of an interactive system (which facilitate the sense of presence), the application would be more effective. However, the relationship between presence and performance is ambiguous. Indeed, a substantial body of research
shows evidence that the immersive characteristics of a VE can promote task performance (Narayan et al., 2005; Schulz et al., 2019; Sousa et al., 2017), learning outcomes (Cheng et al., 2017; Chowdhury et al., 2017; Ai-Lim Lee et al., 2010; Mania and Chalmers, 2000) and therapy effects (Carlin et al., 1997; Dinh et al., 1999; Freeman et al., 1999; Hodges et al., 1994). However, while some literature showed a positive correlation between presence and performance (Singer et al., 1995; Slater et al., 1996), others could not find a significant relationship (Kjær et al., 2017; Modjeska and Waterworth, 2000; Zimmons and Panter, 2003). Welch (1999) states that presence does not facilitate performance, arguing that positive effects are most likely due to the increased immersive properties of the VE (e.g., frame rate, latency, resolution). Similarly, Bystrom et al. (1999) contend that the relationship between presence and performance is task-specific, and performance only increases with greater presence if the latter is relevant for the task. Moreover, there is an open debate in the literature if presence is a suitable construct to describe the quality of a VR experience since (a) it is difficult to measure and (b) its relationship with user performance (Kjær et al., 2017; Modjeska and Waterworth, 2000; Zimmons and Panter, 2003) or the fidelity (Bowman and McMahan, 2007; Slater et al., 2009; Zimmons and Panter, 2003) of the environment is ambiguous. Particularly for MR applications, using a construct such as presence requires critical discussion since non-fully immersive technology such as AR does not affect the same sense of presence as VR does. Nevertheless, the construct of presence summarizes a degree of involvement not captured by traditional UX metrics. Therefore, it is an invaluable metric for the quality of experiences with reality-altering technology.

2.3 Fun, Pleasure, and Enjoyment

Frequently, HCI literature uses the terms *fun*, *pleasure*, and *enjoyment* interchangeably to describe positive experiences (i.e., UX) with interactive systems (Blythe and Hassenzahl, 2018). However, this ambiguity in terminology hazards a clear understanding of how interaction with technology affects people. To disambiguate the qualities of experiences, this section establishes some key terms that characterize the positive outcomes of interaction.

Blythe and Hassenzahl (2018) differentiate between *fun* and *pleasure* attributed to experiential and cultural differences between the two terms. *Fun* is about feeling good (Koster, 2013). It is related to distraction and is associated with triviality, repetition, spectacle, and transgression (Blythe and Hassenzahl, 2018). Accordingly, fun satisfies psychological needs by offering a satisfactory stimulating distraction from mundane activities. In contrast, *pleasure* is attributed to be a deeper type of positive experiences which is related to the feeling of absorption with an activity and is closely related to flow (Csikszentmihalyi, 1990), which is described as a highly intensive experience of an optimal balance between an individual’s felt skills and the difficulty in mastering a challenge. Referring to Aristotle, Blythe and Hassenzahl describe pleasure as “stimulation through action” (Blythe and Hassenzahl, 2018). Pleasure differs from fun in its intensity and connection to the activity. Connotations of pleasure are relevance, progression, aesthetics, and commitment (Blythe and Hassenzahl, 2018). The terms pleasure and fun are not a polar dichotomy as enjoyable experiences are fluidly translating between both (Blythe and Hassenzahl, 2018).

Investigating how individuals are entertained by media consumption, Vorderer et al. (2004) present a conceptual model of entertainment with multimedia and its prerequisites of enjoyment. Vorderer et al. argue that positive experiences should not be considered as monolithic processes but as “set of differentiated subcomponents” (Vorderer et al., 2004). Enjoyment is referred to as “positive cognitive and affective appraisal of the game experience” (Mekler et al., 2014). Similarly, Nabi and Krmer (2004) conceptualize media enjoyment as a three-dimensional construct that encompasses affective, cognitive, and behavioral responses that mutually interfere. Accordingly, *enjoyment* is defined as an experiential state which includes physiological, cognitive, and affective components (Vorderer et al., 2004). The experience of enjoyment is assumed to have at least one of the following prerequisites to be present: (i) a suspension of disbelief, (ii) empathy with the media content, (iii) desire to relate with the characters, (iv) sense of presence, and (v) interest in a specific topic (Vorderer et al., 2004). However, enjoyment is context-dependent. “Activities associated with enjoyment offer potentials for enjoyment rather than
enjoyment itself. [...] In this sense, enjoyment doesn’t exist in and of itself. It’s a relationship between ongoing activities and states of mind” (Blythe and Hassenzahl, 2018). Yet still, although various kinds of enjoyment exist and positive experiences are culturally and individually dependent, there is certainly shared agreement on what is expected to be enjoyable (Blythe and Hassenzahl, 2018; Vorderer et al., 2004). For example, Csikszentmihalyi (1985) considered curiosity, creativity, challenge, and progress as inherently enjoyable experiences. Referring to the design of enjoyable experiences with interactive systems, utilizing these terms, Blythe and Hassenzahl (2018) argue that the enjoyment of repetitive routine tasks could be achieved through fun – distracting an individual from the unpleasing activity. In contrast, non-routine or creative tasks should be absorbing (pleasure), allowing one to concentrate on the activity to be perceived as enjoyable. A significant and growing body of work has studied the enjoyment of digital media on games as their primary objective is to be both pleasurable and fun (Landers et al., 2019; Mekler et al., 2014; Sweetser and Wyeth, 2005). To further clarify game enjoyment, Mekler et al. (2014) conducted a literature review of quantitative studies on game enjoyment and “provide a clearer outline of how enjoyment, flow, immersion, and presence differ and interrelate, and suggest that the player experience may be studied in terms of its valence, that is, how enjoyable it is” (Mekler et al., 2014). In line with Nacke and Drachen (2011), who suggested that immersion does not necessarily need to be enjoyable, Mekler et al.’s analysis show that presence affects game enjoyment only indirectly through facilitated flow experience (Mekler et al., 2014). Mekler et al. summarize that enjoyment is strongly related to flow. However, although these two experiences share similarities, these two constructs differ as flow requires enjoyment and involvement to persist. In contrast, players may experience enjoyment without being in a state of flow (Mekler et al., 2014). Reflecting on definitions of fun, pleasure, and enjoyment by Blythe and Hassenzahl (2018) and Vorderer et al. (2004), Mekler et al. (2014) constitute that games can facilitate both absorbing and distracting experiences. Hence, the more general term enjoyment, which encompasses both fun and pleasure, is more appropriate to describe the experiences with interactive systems.

The design of positive experiences with interactive systems requires careful considerations, which depend on the type of application and its tasks. The presented distinctions of fun (positive distraction), pleasure (absorbing stimulation through an activity), and enjoyment (a general experiential state) consider positive experiences with interactive system on a more differentiated level and help to guide the design of engaging experiences.

2.4 Player Experience

There is an active debate in the literature on what a game is, and many scholars attempted to find a suitable definition for games. The literature distinguishes between game and play. Play is mostly referred to as a voluntary, informal activity. Huizinga defined play as a “free activity standing quite consciously outside ‘ordinary’ line as being ‘not serious’, but at the same time absorbing the player intensely and utterly.” (Huizinga, 2020) In contrast, games are frequently defined as closed formal systems that follow a set of rules which also determine the winning or losing outcome (Crawford, 1984; Salen et al., 2004; Schell, 2008). For an in-depth discussion about the definitions of games, the reader is referred to Stenros’ (2017) recent review of 63 definitions of games which are summarized in ten overarching themes.

Games are considered to have the inherent potential for being enjoyable and, thus, intrinsically motivating (Deci and Ryan, 2000; Gee, 2007; Rigby and Ryan, 2011; Vorderer et al., 2004; Yee, 2006). Therefore, in the past years HCI research turned towards playful experiences and approaches like gamification (Deterding et al., 2011) – embedding game elements into non-game context – and serious games (Abt, 1970) – implementing objectives outside the game content into the game design to facilitate the motivation (i.e., engagement) with interactive systems. Nacke and Drachen (2011) argues that although several models of UX exist, none of them is suitable to explain gameplay experience. The key difference is that games are primarily designed for enjoyment compared to utilitarian systems whose primary goals are pragmatic (i.e., functionality) (Mekler et al., 2014; Nacke and Drachen, 2011; Sweetser and Wyeth, 2005). Similar to UX research, Games User Research (GUR) attempts to capture
and formalize the nature of playing games (Huotari and Hamari, 2017; Koivisto and Hamari, 2019; Malone, 1981; Tyack and Mekler, 2020) by applying psychological, physiological, and behavioral metrics to understand game design and Player Experience (PX) which can be applied into recommendations, guidelines, or principles for the design of more appealing games (Abeele et al., 2020) and engaging interactive systems with a purpose (Boyle et al., 2012; Deterding et al., 2011). Player Experience (PX) is defined as “the qualities of the player-game interactions and is typically investigated during and after the interaction with games” (Wiemeyer et al., 2016). PX derives from UX as an experience that takes motivation (Birk et al., 2016b; Koivisto and Hamari, 2019), emotion (Bateman and Boon, 2006; Bouvier et al., 2014b; Hallifax et al., 2019), and personality (Nacke et al., 2014; Teng, 2008; Zammitto, 2010) into account. Furthermore, to evoke PX, good usability is a prerequisite (Nacke and Drachen, 2011; Wiemeyer et al., 2016). The theoretical foundation of research on PX builds mainly on Self-Determination Theory (SDT) (Ryan and Deci, 2000a) which states that individuals are intrinsically motivated to engage with activities that fulfill the psychological needs of autonomy, competence, and relatedness (Hallifax et al., 2019; Koivisto and Hamari, 2019; Tyack and Mekler, 2020). Tyack and Mekler (2020) provide an extensive overview of SDT and adjacent theories in the field of GUR. Wiemeyer et al. (2016) distinguish three levels of PX: (socio-)psychological level (individual experience), behavioral level, and physiological level. The (socio-)psychological aspects encompass the personal level of experience such as “intrinsically motivated actions (free of external determination), performing symbolic or fictional actions in a quasi-real context constrained by the rules of a game, ambivalence, and openness to both procedure and outcomes, presence and immersion” (Wiemeyer et al., 2016). The behavioral and physiological levels refer to the respective responses such as ducking from a flying to object (Sheridan, 1992) or raised heart rate (Meehan, 2001; Slater et al., 2003). Wiemeyer et al.’s model of PX is depicted in Figure 2.2.

Another influential theory that has informed game design and enjoyable experiences, in general, is Csikszentmihalyi’s flow theory. The flow theory argues that the quality of an experience is determined by a perceived difficulty of a challenge and an individual’s felt skills in mastering the challenge. The flow experience emerges when the challenges and skills are in balance. A too challenging task will lead to anxiety; a too low challenge is perceived as boring. A detailed overview of the affective outcomes is depicted in Figure 2.3a. Contrasting how game enjoyment relates to flow, Mekler et al. (2014) categorize PX on two dimensions with enjoyment as the valence of PX and immersion as the intensity of the experience. As depicted in Figure 2.3b, this conceptualization of game enjoyment produces a plane with four extreme affective outcomes: boredom, anxiety, relief, and flow. Building upon the flow theory in the context of games, research proposes seven major components that contribute to the enjoyment of an activity: (i) chance of completion, (ii) ability to concentrate on a task, (iii) clear goals, (iv) immediate feedback, (v) effortless involvement, (vi) sense of control, (vii) disappearance of concerns for the self, and (viii) alteration of the perception of time (Csikszentmihalyi, 1990; Sweetser and Wyeth, 2005).
2.4 Player Experience

(a) Affective outcomes of the Flow model (Csikszentmihalyi, 1997).

(b) Valence and Intensity of PX by Mekler et al. (2014).

In support of the flow theory as an important factor of PX, research on intrinsically motivating instructional environments identified an inverted-U relationship between intrinsic motivation and difficulty (Abuhamdeh and Csikszentmihalyi, 2012; Lomas et al., 2017). Recurrently literature has shown that flow allows modeling optimal challenges such as Dynamic Difficulty Adjustment (DDA) (Constant and Levieux, 2019; Hunicke, 2005) to facilitate PX (Tyack and Mekler, 2020). However, gameful experiences can evoke enjoyment during gameplay without being in the state of flow (Landers et al., 2019). Therefore, flow theory is only adjacent to PX but cannot describe it solely (Landers et al., 2019; Mekler et al., 2014).

Individual differences have a significant impact on what people might enjoy (Abeele et al., 2020; Nacke et al., 2014; Orji et al., 2013; Tondello et al., 2017). Therefore, a significant and growing body of work on PX investigated how demographic factors and personality traits are responsible for individual preferences in games. Consequently, a considerable body of work links different typologies of players and motivational game elements (Hallifax et al., 2019). For personality traits, the Five Factor Model (FFM) (Goldberg, 1992) – Big Five – with the factors Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Emotional stability is acknowledged as one of the most reliable models to explain personality structures (Digman, 1990). Digman (1990) describes the origins of the FFM and gives an in-depth discussion on the five factors. The FFM model has been frequently applied to evaluate PX (Jia et al., 2016; Orji et al., 2017; Teng, 2008; Zammitt, 2010) or to predict play styles (Bean and Groth-Marnat, 2016; Delhove and Greitemeyer, 2020; McCcreery et al., 2012). However, the results are inconsistent. While it is invaluable as a starting point, the literature results are mixed and limited when explaining what facilitates PX (Chris et al., 2011; Hallifax et al., 2019). Further, Wiemeyer et al. (2016) constitutes that PX and player types are distinguishable. According to Wiemeyer et al. PX “denotes a transient and dynamic construct (state)”, whereas player type is a “more or less stable and static construct or trait” (Wiemeyer et al., 2016); cf. Nacke and Drachen, 2011. Tondello et al. (2017) argue that many models of PX are often too abstract to provide prescriptive guidance on how to design for specific experiences. Therefore, they propose a conceptual framework of player preferences at an intermediate level of abstraction, which maps game elements and playing styles to specific player preferences. Further, Tondello et al. identified several demographic traits that are affecting the individual preferences of play. An in-depth discussion of personality in games is beyond the scope of this thesis. For further reading, Angelides and Agius (2014) provides an in-depth overview and discussion on modeling PX. The authors discuss each model’s strengths and weaknesses and outline how GUR evolved and moved from static player types (cf. Bartle, 1996) towards more dynamic trait-based theories (Abeele et al., 2020). Building on the common ground of enjoyment (cf. Section 2.3), this work primarily focuses on generalizable personality-independent design elements to facilitate a sustainable engagement with the interactive systems.

Figure 2.3: The Flow Model and Player Experience.
Seemingly, most research on PX tends to focus on the experiences at the very instance of play. Nacke and Drachen (2011) emphasize the temporal and contextual dimensions PX which links flow, immersion, and presence as underlying psychological constructs of PX. The temporal aspect of PX is frequently conceptualized as player retention (Nacke and Drachen, 2011; Viljanen et al., 2016), which builds on behavioral psychology to draw players into extended play sessions and afford them to return playing. However, Nacke and Drachen argue that a game with high retention must not necessarily provide a high PX (Nacke and Drachen, 2011). Moreover, the authors point out that “[t]here are as yet no universally accepted guidelines for how to test PX in relation to the temporal dimension of gameplay” (Nacke and Drachen, 2011). Furthermore, Landers et al. (2019) point at the existing gap in the literature that, despite a significant and growing body of research on PX, practical design advice on how to design for sustained enjoyable experiences is still missing.

2.5 Engaging Interaction Design

Interactivity – the degree of how users are able to manipulate virtual content in real time (Speicher et al., 2018) – is a crucial part of what makes interactive systems (both games but also utilitarian systems) enjoyable and what distinguishes them from other types of (digital) media such as television that are passively consumed (Jensen, 1998; Klimmt et al., 2007; Prensky and Thangarajan, 2007). Moreover, it has been shown that raised interactivity facilitates user absorption (Oh and Sundar, 2015). Therefore, when designing interactive systems for engaging experiences, the interaction part should be considered (Stromer-Galley, 2004). Hutchins et al. (1985) coined the term direct engagement as a feeling that occurs “when a user experiences direct interaction with the object in a domain”. Direct engagement facilitates the “feeling of involvement directly with the world of objects rather than of communication with an intermediary” (Hutchins et al., 1985). “Direct, multi-sensory representations have the capacity to engage people intellectually as well as emotionally, to enhance the contextual aspects of information, and to encourage integrated, holistic responses” (Laurel, 2014). Following Laurel (1986), Hutchins et al. (1985) derived four criteria for the feeling of direct engagement: (i) “Execution and evaluation should exhibit both semantic and articulatory directness” (ii) “Input and output languages of the interface should be inter-referential” (iii) “The system should be responsive, with no delays between execution and the results” (iv) “The interface should be unobtrusive, not interfering or intruding” As responsible dimensions for Shneiderman’s (1982) concept of direct manipulation, the authors propose two dimensions of distance and direct engagement; see Figure 2.4. The dimension of distance describes
2.5 Engaging Interaction Design

how one’s thoughts translate into physical actions that the system can interpret. The distance of directness relies on the *gulf of execution* – the match between the users’ thoughts and the commands – and the *gulf of evaluation* – the representation of the output system. Further, Hutchins et al. distinguish between *semantic* and *articulated* forms of distance. The semantic distance relates the meaning of an expression in the interface language to the users’ intention. The articulatory distance is referred to as “the relationship between the meanings of expressions and their physical form” (Hutchins et al., 1985). The authors found that the cognitive effort of interacting with systems is inversely proportional to the feeling of directness. These findings are also in line with theories on embodied cognition (Wilson and Foglia, 2017), which assume that part of our knowledge and world experiences with the world is subconsciously woven into our bodies (Dourish, 2004; Kilteni et al., 2012). According to Johnson-Glenberg and Megowan-Romanowicz (2017), embodiment is not as a binary state, but rather a continuum where the degree of embodiment depends on the amount of sensor-motor engagement, congruency of gestures and amount of immersion. Research on playful physical interaction constitutes that embodied interaction can facilitate learning (Johnson-Glenberg and Megowan-Romanowicz, 2017; Lyons et al., 2012; Moreno and Mayer, 2007) and cognition (Kim and Maher, 2008). Moreover, the embodiment in game design has been shown to facilitate persuasion (Gerling et al., 2014; Whitson et al., 2008). Likewise, Bianchi-Berthouze et al. (2007) investigated how the usage of controllers affected the experience and identified that full-body interaction facilitates the sense of presence, communication, and affect. Therefore, “the goal of designing for articulatory directness is to couple the perceived form of action and meaning so naturally that the relationships between intentions and actions and between actions and output seem straightforward and obvious” (Hutchins et al., 1985).

A more recent framework that helps to conceptualize direct engagement within a larger framework of post-WIMP interfaces is Reality-based Interaction (RBI) by Jacob et al. (2008) (cf. Fig. 2.5). From a literature analysis, Jacob et al. identified RBI as a unifying thread in research, which they categorized into four themes: (i) *Naïve Physics*: The informal understanding of the basic physical world and its principles. (ii) *Body Awareness & Skills*: The embodied understanding of the space relative to their bodies. (iii) *Environment Awareness & Skills*: The sense of presence in the physical environment. (iv) *Social Awareness & Skills*: The verbal and non-verbal abilities to communicate with other individuals in the environment. Jacob et al. point out that while realism allows to build the interaction on the users’ existing knowledge, it may stand in stark contrast to the usefulness/practicality of the interface. Jacob et al. (2008) formulated the five tradeoffs *expressive power, efficiency, versatility, ergonomics, accessibility, and practicality*, which may occur if designs try inconsiderately mimicking the reality. As Figure 2.5 shows, these five tradeoffs impair all four dimensions of RBI. Therefore, designers require careful considerations on when and how to apply RBI in order to overcome these tradeoffs. Yet, referring to Slater et al.’s (1995) and Usoh et al.’s (1999) comparison studies of locomotion methods in VR, Jacob et al. noted, that in certain instances, users may prefer a more realistic interaction paradigm over a more efficient one (Jacob et al., 2008). Therefore, to achieve the best of both worlds, interfaces should both consist of RBI elements where realistic tasks are mapped on realistic actions and non-realistic “computer-only functionality” (Jacob et al., 2008) which deviate from realism. Still, designs should try “to use analogies for these commands whenever possible” (Jacob et al., 2008).

A significant branch of HCI is investigating and developing Tangible User Interfaces (TUIs) (Ishii, 2008; Stone, 2001). This research works at the intersection of physical and digital interfaces that build on haptic and embodied interaction (Shaer and Hornecker, 2009). TUIs are physical objects – physical props – which act as an input and/or output device for the interactive systems (Shaer and Hornecker, 2009). There is a consensus in the literature that TUIs facilitate the enjoyment with the interface (van Huysduynen et al., 2016; Xie et al., 2008), can support learners in the understanding of complex or abstract concepts (Antle and Wise, 2013; Avila-Soto et al., 2017; Engelkamp and Zimmer, 1984; Papert and Harel, 1991; Price et al., 2008), improve the users’ performance with task execution (Antle et al., 2009) or task planning (Underkoffler and Ishii, 1999). Therefore, various physical interfaces, ranging from purely passive props (Hinckley et al., 1994) over passive (Villar et al., 2018; Voelker et al., 2015) and actuated (Le Goc et al., 2016) shape-changing (Rasmussen et al., 2012) interfaces, have been developed and refined. Similarly, a growing body of work proposed shape-changing VR-controller, which extends the physical illusion in the virtual space (Zenner and Kruger, 2017). In line with Ishii and Ullmer’s (1997) vision of “tangible bits” – self-organizing TUIs which dynamically adapt the physical
Background

2 Background

naïve physics → expressive power
body awareness & skills → efficiency
environmental awareness & skills → plasticity
social awareness & skills → ergonomics
accessibility → practicality

Figure 2.5: Dimensions of Reality-based Interaction (RBI) and their connection the design tradeoffs by Jacob et al. (2008).

For haptic interaction, Kim and Schneider (2020) coined the term Haptic Experience (HX) which describes a theoretical model that links and structures design parameters, usability factors, and experiential dimensions of haptic interaction. Kim and Schneider discriminate HX from UX due to the entanglement between the physical feedback and the reaction of the interactive system. Moreover, the timeframe of HX differs. While UX embodies the aspects of the experience before (point of engagement) during (period of engagement) and after (point of disengagement), “HX depends highly on focused interaction in the moment” (Kim and Schneider, 2020); i.e., the period of engagement. This framework helps to conceptualize how users engage with haptic interfaces as it unifies the pragmatic and hedonic outcomes of HX with the design of tangible interfaces.

Most VR experiences rely on visual and auditory stimuli (Melo et al., 2020). Despite the visual dominance (Posner et al., 1976), the more senses a simulation addresses, the more realistic the experience is perceived (Dinh et al., 1999; Slater and Wilbur, 1997). Melo et al. (2020) reviewed the literature on multi-sensory stimuli in VR and identified haptic interaction as the dominant modality in VR research. However, other senses such as smell or taste have been frequently addressed with interfaces as well (Harley et al., 2018). Melo et al. (2020) highlight the importance of multi-sensory stimulation and its appropriate (plausible) integration into the VE to facilitate a Plausibility Illusion (Slater et al., 1994) and to achieve better performance in the VE (Dinh et al., 1999; Slater and Wilbur, 1997). It is well known that congruent multi-sensory cues improve the reaction time to simple stimuli (Hughes et al., 1994; Laurienti et al., 2004; Miller, 1982). However, multi-sensory stimuli do not necessarily raise the sense of presence (Fröhlich and Wachsmuth, 2013) since unaimed adding cross-modal stimuli to the experience may negatively affect the experience (cf. Skarbez et al., 2020; Zenner and Kruger, 2019) or can lead to a raised cognitive load (Melo et al., 2020). These concerns call for careful considerations of multimodal interaction in VR.

2.6 User Engagement

A different attempt to conceptualize the experiences with interactive systems is User Engagement (UE) (Doherty and Doherty, 2019). Similar to the active discussion on the definitions of UX and
2.6 User Engagement

PX, there are multiple definitions for UE (cf. Attfield et al., 2011; Lehmann et al., 2012; O’Brien and Toms, 2008; O’Brien, 2016; Peters et al., 2009). Generally speaking, it refers to the “quality of user experience characterized by the depth of an actor’s investment when interacting with a digital system” (O’Brien, 2016) and “emphasises the positive aspects of the interaction and, in particular, the phenomena associated with being captivated by the technology” (Attfield et al., 2011). UE is a concept that embodies the “health of the system” and helps to design useful and not just usable systems (Doherty and Doherty, 2019). As Attfield et al. (2011) stated, “successful technologies are not just used, they are engaged with; users invest time, attention, and emotion into the equation” (Attfield et al., 2011). Therefore, in most cases, it is desirable to have an engaging interactive system as it facilitates retention, productivity, and overall satisfaction. For instance, many computer-aided health interventions rely on recurrently repeated exercises over a prologues period (Mandryk and Birk, 2017; Perski et al., 2017). Similarly, learning and skill acquisition are much more effective and sustainable if the practice sessions are repeated regularly (Criscimagna-Hemminger and Shadmehr, 2008; Johanson et al., 2019; Kim et al., 2013) or follow a pattern of spaced repetition – repetitions with exponentially growing intervals between the sessions (Schimanke et al., 2017). Whereby it is irrelevant if the learner takes a break or switches to a different task (Piller et al., 2020). Research on UE is also embedded in empirically grounded theories such as flow (Csikszentmihalyi, 1990) or SDT (Deci and Ryan, 1985) and investigates how to design technology that users build relationships with and integrate into their daily lives (Attfield et al., 2011). UE is an abstract concept that may manifest differently depending on the context the technology is embedded in (O’Brien et al., 2018; Perski et al., 2017). Therefore, the research considers engagement as a multidimensional and flexible meta construct depending on the application, where the perspective on UE is adopted to reflect a desired experience best (Bouvier et al., 2014a; Doherty and Doherty, 2019; Perski et al., 2017). But along the theories, there are some common threads of describing UE. From a literature review on engagement, Attfield et al. (2011) summarized Focused Attention, Positive Affect, Aesthetics, Endurability, Novelty, Richness of Control, Reputation, Trust and Expectation, and User Context as emotional, cognitive, and behavioral characteristics of UE that frequently appear in the literature. Notably, the construct of PX shares several similarities with UE (cf. Figures 2.2 and 2.6). For example, in both cases, flow, motivation, and the enjoyment of the overall experience are frequently cited as determinants. Adjacent to PX, the sense of engagement is inferred by other cognitive, affective, and behavioral constructs such as flow (Csikszentmihalyi, 1990), motivation (Rigby and Ryan, 2011), or attention (Attfield et al., 2011; O’Brien and Toms, 2008). The cognitive aspect of engagement frequently relies on conscious components such as attention, interest, or effort (Doherty and Doherty, 2019; Islas Sedano et al., 2013; Sun, 2014). The affective component of engagement encompass the subjective emotional responses including enjoyment, aesthetics, endurability, and novelty (Doherty and Doherty, 2019; O’Brien and Toms, 2008). Behavioral component describes the action and participation with the activity (Doherty and Doherty, 2019; Islas Sedano et al., 2013; Silpasuwanchai et al., 2016). Similar to Mekler et al.’s (2014) dimensions of PX, Perski et al. (2017) identified two general themes of engagement in the literature: engagement as subjective experience and engagement as behavior.

Engagement as Subjective Experience

It is argued that engagement is fundamental for dramatic interaction (Laurel, 2014). Laurel (2014) builds on Samuel Taylor Coleridge’s definition of engagement as the “willing suspension of disbelief” (Coleridge, 1817) and states that engagement consists both of cognitive and emotional components. “It is the state of mind that one must attain to enjoy a representation of an action [...] Further, engagement entails a kind of playfulness: the ability to fool around, to spin out ‘what if’ scenarios” (Laurel, 2014). Coleridge noticed that to experience and to enjoy a drama emotionally, an individual must be willing to suspend the awareness that it is a pretended illusion. Therefore, engagement is achieved if the content and the interactive system fulfills the intellectual, interactional, and perceptual expectations (Bouvier et al., 2014a). Laurel (2014) argues that the suspension of disbelief is almost identical to drama in computer games. “Engagement is what happens when one is able to give oneself over to a representational action, comfortably and unambiguously” (Laurel, 2014). A standardized method to measure UE is the User Engagement Scale (UES) (O’Brien and Toms, 2013) which conceptualizes
2 Background

UE on the subscales: Aesthetics, Novelty, Involvement, Focused Attention, Usability, and Endurability. Similarly, Peters et al. (2009) identified attentional and emotional involvement as fundamental underlying factors of engagement. Adjacent to flow, immersion is frequently attributed to the sense of engagement (Bouvier et al., 2014a). Brown and Cairns’ (2004) 3 levels of immersion in games as a scale of involvement refer to (i) engagement, the lowest level of immersion, (ii) engrossment, when players become emotionally affected by the experience and (iii) total immersion where players are detached from reality. This phase is also associated with empathy for the characters in the game and transfer of consciousness (Sanchez-Vives and Slater, 2005) and is linked to the state of flow (Csikszentmihalyi, 1990). However, in contrast to flow or immersion, “engagement does not require a loss of mental contact with the real world (engaged-players can still be aware of their surroundings)” (Bouvier et al., 2014a). These determinants of UE can be summarized as the intensity of the experience.

Engagement as Behavior

The engaging behavior is recurrently attributed to adherence (Couper et al., 2010; Nacke and Drachen, 2011; Perski et al., 2017). For health interventions, adherence has been further described by: (i) “amount” or “breadth” which characterize the total length of the intervention, (ii) “duration” as the period of time for the intervention and (iii) “frequency” expresses how often subjects undergoing the intervention (Perski et al., 2017; Voils et al., 2014). For games, Bouvier et al. (2014a) further specify four types of engaged behavior: (i) Environment-directed engaged-behavior addresses the contemplation and curiosity and elicits the autonomy dimension of SDT. For this type of behavior, the players’ goal is to explore the game and its boundaries. (ii) Social-directed engaged-behavior concerns the social interaction within the game and is linked to relatedness dimension of SDT. (iii) Self-directed engaged-behaviors refers to the players’ self-identification through characters and ownership aspects. This type of engagement is addressed through personalization (Birk et al., 2016a) and is linked to autonomy. (iv) Action-directed engaged-behaviors is associated with the competence and autonomy dimensions of SDT. This type of behavior is closely related to flow and elicits emotions of goal achievement and completion. O’Brien and Toms (2008) conceptualize four stages of UE (cf. Fig. 2.6): point of engagement is the first contact with the interactive system; period of sustained engagement is the actual time span users interacting with the system; disengagement describes the termination point of an engaging period (e.g., end of a session); and re-engagement is referred to when users return to the interactive system. Each of these phases is characterized by different attributes of the UX that the interaction design should emphasize (O’Brien and Toms, 2008). These four stages allow conceptualizing experiences with interactive systems on a timeline with interaction cycles where each stage gives specific target points for the interaction design that researchers and designers could use to orchestrate the experience (e.g., the behavior of the user).

2.7 Measuring Experiences

Essential to the constructs of experience is that they are measurable and can be attributed to specific variables. The quality of an experience is inferred by assessing related or underlying psychometric variables (Attfield et al., 2011; Doherty and Doherty, 2019; O’Brien and Toms, 2008; Skarbez et al., 2017; Wiebe et al., 2014) (e.g., assessing the sense of presence as a metric of engagement). Since UE and related concepts such as UX or PX are multi-dimensional, a multitude of measurement methods has been developed (Attfield et al., 2011). Research distinguishes between subjective and objective metrics (Attfield et al., 2011; Canossa et al., 2011; Skarbez et al., 2017).

Subjective measures are generally self-reported and assess the users’ perception of the experience (Attfield et al., 2011). Subjective self-reports in the form of (standardized) questionnaires are the most commonly applied assessment method of UX in HCI (Skarbez et al., 2017; Wiebe et al., 2014). The major advantage of questionnaires is that they are easy to administer and generally don’t require modifications of the interactive system (Skarbez et al., 2017). However, post-experience questionnaires are not sensitive to state changes during the ongoing experience (Liebold et al., 2017; Slater et al., 2003).
Particularly for VR, post-experience questionnaires may be unfavorable since they disrupt the subjects from the ongoing experience (Freeman et al., 1999; Skarbez et al., 2017; Slater, 2004) and may produce a BIP (Slater and Steed, 2000).

**Objective measures** include performance metrics (Mendes et al., 2017; Sargunam et al., 2017; Singer et al., 1995), physiological responses (Bateman and Nacke, 2010; Mandryk, 2004; Putze, 2019), users’ perception of time\(^1\) (Attfield et al., 2011) and behavioral metrics such as gaze direction (Newn et al., 2017) or body movement (Freeman et al., 2000). Physiological responses provide information about specific episodes of the experience (Liebold et al., 2017) and allow a better interpretation of subjective ratings and task performance (Brogni et al., 2007). Objective measurements of experience overcome some of the subjective measures since they are contemporaneous and non-intrusive (Skarbez et al., 2017). However, often they are cumbersome to administer and often difficult to interpret (Skarbez et al., 2017; Slater, 2007) since different events may cause similar physiological reactions; e.g., raised Heart Rate (HR) can be attributed to anxiety (Croft et al., 2004) or to physical activity (Freedson and Miller, 2000). Further, to trigger specific behavioral responses, the ongoing evaluation procedure requires specific manipulations of the apparatus (e.g., users are required to duck from flying by object), which are not always applicable (Skarbez et al., 2017).

To date, HCI faces a series of challenges and possible pitfalls around measuring UX, especially in immersive environments. Although various toolkits and frameworks exist which address some of those challenges, for example, VR Questionnaire Toolkit (Feick et al., 2020), VRate (Regal et al., 2019), or MRAT (Nebeling et al., 2020) are toolkits designed to conduct questionnaires in VR. However, these tools differ drastically in their interaction as well as presentation. This lack of a shared agreement on assessment methods for immersive environments diminishes the comparability across studies and might produce uncontrolled biases.

### 2.8 Summary

HCI research pursues several approaches to describe the experiences users have when interacting with technology. The discussed constructs of UX, PX, and UE have strong overlaps in terms of their

\(^1\)The subjective ratings of time can be compared to the real-time passed by which provides an objective reference to contrast the subjective rating.
subconstructs (i.e., psychological, behavioral, and cognitive attributes) as well as their prerequisites; i.e., usability. This is not surprising since they aim to describe positive experiences when users interact with technology and suggest a voluntary interaction with ease of use. Similarly, research shares a common agreement that the experience is situated in a context that influences the subjects. Moreover, players or users have their own personalities, making them have different preferences for interactive and playful experiences. In certain instances, these terms could be used interchangeably and with the convergence of hedonic and utilitarian interactive systems (cf. Koivisto and Hamari, 2019) the line between the experience becomes blurred. However, each construct emphasizes a particular aspect of the experience. UX puts focus on the hedonic qualities of interaction with function-oriented interactive systems, discriminating them from pure pragmatic usability goals. Using Mekler et al.’s language to describe positive experiences, the goal of UX design lies on the pleasure (absorbing stimulation) of the interaction rather than the fun (enjoyable distraction) (Hassenzahl and Tractinsky, 2006). In contrast, PX describes the experience of interaction with playful technology (i.e., games or game-like systems such as serious games), putting much attention on the motivational, joyful, and absorbing qualities of the design such as immersion or flow. UE attempts to describe the quality and the depth of an experience. However, the advantage of UE over the other perspectives is the tangible approach to the temporal aspects of the experience through the O’Brien and Toms’ four stages engagement. This lens on experiences with interactive systems allows conceptualizing and designing interactions that account for an intended user or player behavior.

Regarding interaction design, the reviewed literature highlights the importance of embodiment and directness as design paradigms for engaging interaction. Therefore, contemporary HCI research focuses on the design of reality-based and multimodal interaction. However, RBI raises challenges in the design as the direct interaction may limit the expressive power and the practicality of the interfaces.

To investigate the research questions, with the attempt to contribute to the future unifying theory of engaging interaction design, this work follows Perski et al.’s (2017) conceptualization of experiential and behavioral engagement. It synthesizes it with O’Brien and Toms’ (2008) four-stages model of UE. This multi-perspective view on engagement attempts to incorporate both the intensity of the experience at the moment of interaction and re-engaging factors on a bigger timeframe beyond the direct usage of the interactive system.
Chapter 3

Game Design for User Engagement

Following Attfield et al.'s (2011) lines of research on User Engagement (UE) and Koivisto and Hamari's (2019) re-appropriation of playfulness for utilitarian interactive systems, this chapter investigates RQ1 and examines game designs that enact the design for intended player behavior and support player engagement. As many applications, such as health intervention, rely on regular usage, the work presented in this chapter investigates game design that supports adherence. Section 3.1 presents a framework of five game dynamics adopted from mobile games evaluated in a casual game to investigate how different game elements facilitate a regular but brief player engagement – the snacking pattern of play. To validate how these game elements affect PX in other contexts, Section 3.2 investigates how the snacking framework translates from a casual game into a serious game and evaluates its effect on player attrition. Furthermore, this work identified that many traditional game designs are not applicable as a motivational strategy for mental health intervention. Therefore, as a special case of serious games, this work investigates the unexplored area of GUR of emotional game design for the treatment of phobias in VR. Section 3.3 presents Playful User-generated Treatment (PUT) as a two-step game design approach for acrophobia Virtual Reality Exposure Therapy (VRET) which employs creativity and self-empowerment as the driving factors of motivation to engage with the intervention. Finally, Section 3.4 summarizes the results and reflects them with respect to RQ1.

3.1 Designing for Snacking Behavior

Research has shown that consistency and repetition are the most effective patterns to learn and master new skills (Johanson et al., 2019; Schimanke et al., 2017) or to complete a therapy (Eysenbach, 2005). This behavioral pattern is arguably desired in most training, learning, and health intervention scenario. To investigate how persistent player behavior can be achieved, in publication “Casual Snacking” (Alexandrovsky et al., 2019b, F1), we conducted a systematic content analysis of 30 casual mobile games and identified single-player mechanics that promote the snacking patterns of interaction (Morris et al., 2015) – a regular but brief pattern of play (Oulasvirta et al., 2005; Vaish et al., 2014). We conceptualize the snacking pattern as an approximately daily play with a session duration of around 15 minutes. These game mechanics were assembled into a framework of five single-player game dynamics – i.e., game experiences that result from playing the specific game mechanics (Hunicke et al., 2004):

- **Rewards** are immediate gratifications or achievements for the players’ actions. Rewards address Skinner’s operant conditioning (Skinner, 1938; Skinner, 1950), which states that rewarding the desired behavior reinforces that behavior to be repeated.

- **Novelty** provides players with new game elements. It contributes to challenge (Lomas et al., 2017) and promotes novel stimuli which prevent players from boredom (Berlyne, 1970; Daffner et al., 1998; Fantz, 1964; Johnston et al., 1990).

- **Completion** describes explicitly assigned tasks, quests, or goals. Completion can fulfill the needs for competence, autonomy, and relatedness (Ryan and Deci, 2000a; Ryan et al., 2006). Moreover, unfinished tasks have an inherent pull towards the tasks (Mäntylä and Sgaramella, 1997; Syrek et al., 2017).
• **Blocking** temporally prohibits the playing. Blocking can be conceptualized as a conditioning mechanism with a fixed reward scheme, which limits the exposure and adds value to the next session.

• **Waiting** is a game mechanic that outbreaks the players through the hindrance of their progress or restricted access to resources. Waiting is closely related to Blocking; however, it operates on a different scale. It can be used as operant conditioning (Skinner, 1938) and delayed gratification (Mischel et al., 1972) mechanisms.

The framework situates these five game dynamics into theories of motivated behavior that help explain why they effectively motivate behavior. To test the mechanics, Alexandrovsky et al. (2019b, F1) designed the casual game Game of Cannons (GOC), in which the five game mechanics be could individually layered game into the game. GOC is a turn-based 2D physics puzzle with action elements inspired by games like Angry Birds™ (Jaakko Iisalo and Rovio Entertainment Ltd., 2009), a popular casual mobile game. The player is equipped with a cannon and a building with minions in it. Additionally, the player is represented by an avatar standing next to the cannon and allows for small customization with hats and skin color. On the other side of the scene, an Artificial Intelligence (AI) opponent with the same items faces the player. There is also an inventory in which in-game items are stored and can be accessed. The goal is to shoot the opponent’s minions, who live in their buildings, while protecting one’s own base. Each of the parties takes turns shooting at the opponent’s building, trying to eliminate all minions, after which the match is over. The game contains 64 levels with increasing difficulty: the buildings become more difficult to destroy, and the accuracy of the AI increases.

To provide a reasonable difficulty curve, 44 levels were adopted from Angry Birds. Initially, only the first level is unlocked, and completing a level unlocks the next one. To add randomness to the game, we introduced an additional wind mechanic that changes every turn in both strength and direction, affecting the projectile’s trajectory. The game was developed in Unity3D for the web and is entirely controlled with a mouse. A level of the game is depicted in Figure 3.1 and the gameplay videos1. To provide game elements for the different game mechanics, the game contains 5 types of projectiles for the cannon and 3 types of power-ups: Telescope provides a visual preview of the trajectory for a shot, Explosion increases the destructive power of a projectile, and Airstrike calls a UFO that throws a bomb on the opponent’s building. The different game versions build on top of the base game and implement the individual dynamics as follows:

• **Rewards**: Players are rewarded with score points and power-ups for destroying the opponent’s items. Like Angry Birds, completing a level is rewarded with up 3 stars depending on the number of turns and how many blocks were destroyed.

• **Completion**: We implemented 28 achievements that players could unlock throughout the game. Players were rewarded for in-game progression and playing daily with projectiles for the cannon and power-ups.

• **Novelty**: We implemented multiple novelty mechanics that are introduced to the player over time. First, the projectiles were tied to particular levels and introduced evenly throughout the 64 levels. Also, the levels were evenly split into four differently-themed worlds with altered visual styles. Finally, 21 levels, and in every final world level, players earned hats to customize their avatars.

• **Blocking**: This game version restricts the playtime to 30 minutes every 8 hours. Besides the playtime restriction and the waiting screen, Blocking did not differ from the base game.

• **Waiting**: In this game version, the power-ups are generated over time. Countdowns indicate the arrival of the power-ups. Players were restricted to only one of each power-up at the same time.

---

1Video figures demonstrating the gameplay of GOC: https://youtube.com/playlist?list=PLJY6D-L7S0gtXgHt_QAgvp8AIKypcYV8b, accessed: 20.02.2021.
Figure 3.1: Game of Cannons game screen (Alexandrovsky et al., 2019b, F₁).

Therefore, a power-up needed to be used before the same power-up could be collected again.

The game versions were evaluated in two comparative user studies using the crowd-sourcing platform Amazon Mechanical Turk (MTurk). First, we conducted a preliminary short-term study with \( n = 99 \) participants to validate if (a) the designed game met the intended overall experience: an enjoyable – but not immersive – game experience where players feel competent and the controls are intuitive and (b) the game versions are comparable in terms of short-term PX and motivation. We measured PX using the Intrinsic Motivation Inventory (IMI) (Ryan, 1982) and Player Experience of Need Satisfaction (PENS) (Ryan et al., 2006). As a metric of personality traits, we applied the Ten Item Personality Inventory (TIPI) (Nunes et al., 2018), which assesses personality based on the FFM model. As objective performance metrics, we logged the reached scores, unlocked levels, and the number of turns. The pre-study was conducted on MTurk with \( n = 99 \) participants. The subjects were asked to fill out the demographics surveys which after each participant played one of the five game versions for \( 15 \) min. Blocking was abandoned, as in the short-term study, it is equivalent to the baseline version. The pre-study showed equivalent levels of PX regardless of the inclusion of the mechanics and allows to suggest that the game versions were comparable in appeal over the initial \( 15 \) min of play (cf. Tab.3.1). On the performance metrics, there were no statistically significant differences between the game versions on score \( (F_{4,1080} = .922, \eta^2 = .003, p = .450) \) nor on number of turns \( (F_{4,1080} = 1.510, \eta^2 = .006, p = .197) \). However, there was a significant difference on levels completed between conditions \( (F_{4,1080} = 2.773, \eta^2 = .010, p = .026) \). These results allow to conclude that the game versions are comparable in terms of short-term PX and playability and that the game is suitable to investigate how the different dynamics affect the snacking pattern of play.

After establishing that the game is enjoyable and that the game versions are comparable in a short-term PX, we conducted the main study. The second study was a 9-day behavioral experiment which investigated whether the different game mechanics facilitate the snacking pattern. We released six versions of the game (Baseline plus five game versions for each individually layered mechanic) on MTurk to groups of players who were paid in advance. After the initial \( 15 \) minute play round, the participants could play the game as often as they liked over the experimental period. On the first day, the players completed the demographic survey, played the game for \( 15 \) minutes, and filled out the PX questionnaires. After that, the participants were free to play to their liking. At midterm, on the day 5, the participants received the IMI and Situational Motivation Scale (SIMS) (Guay et al., 2001) questionnaires. SIMS contains four subscales that measure intrinsic motivation, identified regulation,
Table 3.1: Pre-study test statistics for the IMI (top) and PENS (bottom) (Alexandrovsky et al., 2019b, F₁).

<table>
<thead>
<tr>
<th>Interest-Enjoyment</th>
<th>p</th>
<th>Intrinisic Motivation</th>
<th>Baseline</th>
<th>Novelty</th>
<th>Waiting</th>
<th>Rewards</th>
<th>Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>M(SD)</td>
<td>M(SD)</td>
<td>M(SD)</td>
<td>M(SD)</td>
<td>M(SD)</td>
<td>M(SD)</td>
<td>M(SD)</td>
<td></td>
</tr>
<tr>
<td>Interest-Enjoyment</td>
<td>0.071</td>
<td>0.991</td>
<td>0.003</td>
<td>5.11(1.70)</td>
<td>4.98(1.60)</td>
<td>5.17(1.36)</td>
<td>5.08(1.44)</td>
</tr>
<tr>
<td>Competence</td>
<td>1.376</td>
<td>0.249</td>
<td>0.058</td>
<td>4.31(1.42)</td>
<td>4.07(1.58)</td>
<td>4.86(1.39)</td>
<td>4.69(1.50)</td>
</tr>
<tr>
<td>Effort-Importance</td>
<td>0.466</td>
<td>0.441</td>
<td>0.034</td>
<td>2.84(1.64)</td>
<td>2.98(1.32)</td>
<td>2.47(1.62)</td>
<td>2.28(1.60)</td>
</tr>
<tr>
<td>Tension-Pressure</td>
<td>0.282</td>
<td>0.889</td>
<td>0.013</td>
<td>4.67(1.17)</td>
<td>4.57(1.45)</td>
<td>4.74(1.53)</td>
<td>4.88(1.40)</td>
</tr>
<tr>
<td>Need Satisfaction</td>
<td>0.104</td>
<td>0.981</td>
<td>0.005</td>
<td>3.77(1.64)</td>
<td>3.94(1.76)</td>
<td>4.06(1.57)</td>
<td>3.94(1.62)</td>
</tr>
<tr>
<td>Autonomy</td>
<td>0.247</td>
<td>0.911</td>
<td>0.011</td>
<td>3.09(1.76)</td>
<td>3.21(1.63)</td>
<td>3.48(1.25)</td>
<td>3.46(1.43)</td>
</tr>
<tr>
<td>Relatedness</td>
<td>0.52</td>
<td>0.722</td>
<td>0.023</td>
<td>3.03(1.54)</td>
<td>3.36(1.57)</td>
<td>3.61(1.37)</td>
<td>3.10(1.22)</td>
</tr>
<tr>
<td>Immersion</td>
<td>0.156</td>
<td>0.96</td>
<td>0.007</td>
<td>5.26(1.30)</td>
<td>5.05(1.23)</td>
<td>5.24(1.70)</td>
<td>5.13(1.75)</td>
</tr>
</tbody>
</table>

Table 3.2: GOC long-term study demographics (Alexandrovsky et al., 2019b, F₁).

<table>
<thead>
<tr>
<th>Extraversion (1-7)</th>
<th>Mean</th>
<th>SE</th>
<th>F&lt;sub&gt;5,114&lt;/sub&gt;</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreeableness (1-7)</td>
<td>3.3</td>
<td>0.15</td>
<td>0.37</td>
<td>0.869</td>
</tr>
<tr>
<td>Conscientiousness (1-7)</td>
<td>5.4</td>
<td>0.11</td>
<td>1.24</td>
<td>0.296</td>
</tr>
<tr>
<td>Emotional Stability (1-7)</td>
<td>5.6</td>
<td>0.10</td>
<td>1.63</td>
<td>0.157</td>
</tr>
<tr>
<td>Openness to Experiences (1-7)</td>
<td>5.1</td>
<td>0.14</td>
<td>0.55</td>
<td>0.742</td>
</tr>
<tr>
<td>Age</td>
<td>36.4</td>
<td>0.85</td>
<td>0.93</td>
<td>0.465</td>
</tr>
<tr>
<td>Identification as a Gamer (1-10)</td>
<td>6.5</td>
<td>0.25</td>
<td>3.98</td>
<td>0.002</td>
</tr>
<tr>
<td>Intrinsic Motivation (1-7)</td>
<td>5.2</td>
<td>0.12</td>
<td>1.20</td>
<td>0.312</td>
</tr>
<tr>
<td>Identified Regulation (1-7)</td>
<td>4.4</td>
<td>0.11</td>
<td>0.54</td>
<td>0.745</td>
</tr>
<tr>
<td>External Regulation (1-7)</td>
<td>1.8</td>
<td>0.10</td>
<td>0.62</td>
<td>0.688</td>
</tr>
<tr>
<td>Motivation (1-7)</td>
<td>1.9</td>
<td>0.11</td>
<td>0.41</td>
<td>0.844</td>
</tr>
</tbody>
</table>

external regulation, and amotivation. On the final day (day 9), the participants completed the exit survey, which consisted of the IMI, SIMS, and open-ended questions about the game. Throughout the study, each midnight, the participants received a notification that reminded them to reengage with the game. Like in the pre-study, we tracked the score, the number of turns, and the unlocked levels as performance metrics. We logged the exact time and duration of the play sessions for session characteristics, which we used to calculate the inter-session interval as our primary behavior metric.

We recruited 180 participants on MTurk (49.2% female, age: M=35.0, SD=9.3) across the six conditions. After filtering our sample consisted of n=128 with comparable numbers of participants in each group (N: Baseline=20; Novelty=24; Waiting=20; Rewards=21; Completion=22; Blocking=21). There were no significant differences in any demographic factors, incoming motivation, or personality, showing that the groups were comparable (cf. Tab 3.2). The only significant difference was on how much the participant identified as a gamer (F<sub>5,114</sub>=3.98, p=0.002, η<sup>2</sup> = 0.149). Games-Howell post-hoc tests revealed that participants in Rewards identified less as a gamer than those in Novelty (p=0.007) or Waiting (p=0.011). However, since our analyses compare each mechanic to Baseline, and there were no significant differences with the Baseline condition, we do not further control for the lower gamer identity in Rewards.

To determine whether the game versions yielded different PX, we conducted a Multivariate Analysis of Variance (MANOVA) on the enjoyment metrics at each day of assessment with the conditions as the between-subject factor. The MANOVA showed no significant differences in enjoyment between the conditions after day one (F<sub>5,22</sub>=0.80, p=0.556), at the midterm (F<sub>5,22</sub>=0.95, p=0.451), or after the exit survey on the final day (F<sub>5,22</sub>=0.62, p=0.687). These results corroborate the pre-study results and demonstrate that differences in enjoyment did not emerge over repeated play. We conducted a one-way Analysis of Variances (ANOVA) on the performance metrics with the conditions as the between-subjects factor to examine if the conditions yielded different performances. There were significant effects of dynamic on average score (cf. Fig. 3.2a) and levels completed (cf. Fig. 3.2b), and a marginal
3.1 Designing for Snacking Behavior

Figure 3.2: GOC long-term study performance results (Alexandrovsky et al., 2019b, F1).

Table 3.3: GOC long-term: Correlations between session performance metrics (score, turns, levels) and session characteristics (duration, inter-session interval, session count) (Alexandrovsky et al., 2019b, F1). **p<.01, *p<.05.

<table>
<thead>
<tr>
<th>metric</th>
<th>score</th>
<th>turns</th>
<th>levels</th>
<th>inter-session interval</th>
<th>session count</th>
</tr>
</thead>
<tbody>
<tr>
<td>score</td>
<td>.282**</td>
<td>.356**</td>
<td>.175*</td>
<td>-.203*</td>
<td>.259**</td>
</tr>
<tr>
<td>turns</td>
<td></td>
<td>.748**</td>
<td>.671**</td>
<td>-.091</td>
<td>-.017</td>
</tr>
<tr>
<td>levels</td>
<td></td>
<td></td>
<td>.465**</td>
<td></td>
<td>.015</td>
</tr>
<tr>
<td>duration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>inter-session interval</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>session count</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To control for the differences in performance for the behavior analysis, we correlated the performance and the session characteristics (cf. Tab. 3.3). The analysis showed significant correlations of average score with inter-session interval and session count. Therefore, we used average score as a covariate on the analysis for inter-session interval and session count. Duration correlated significantly with all performance metrics. Therefore on the analysis of duration, we controlled for performance using the number of turns, which showed the highest correlation. To determine if the game versions facilitated different play patterns, we conducted a one-way Analyses of Covariance (ANCOVA) on the session characteristics with the condition as the between-subjects factor and the respective covariate as a control for the differences in performance. There was a significant effect of condition on the inter-session interval (cf. Fig. 3.3a). The planned contrasts between the conditions and Baseline revealed a shorter inter-session interval in Novelty ($t_{122}=-2.71$, $p<0.01$) and Waiting ($t_{122}=-2.06$, $p=0.04$), but not for Rewards ($t_{122}=-0.83$, $p=0.41$), Completion ($t_{122}=-0.16$, $p=0.87$), or Blocking ($t_{122}=-1.49$, $p=0.14$). The session duration (cf. Fig. 3.3b) and the number of sessions (cf. Fig. 3.3c) did not differ significantly between the conditions.

To further investigate if the game dynamics differently influence the snacking pattern of play; i.e., sessions of around 15 minutes with an inter-session interval of 24 hours, we applied a k-Means ($k=2$) clustering for sessions characteristics (inter-session interval, duration and session count) to split the sample into ‘snacking’ and ‘non-snacking’ groups. As Table 3.4 shows, the clusters are significantly differentiated by inter-session interval and the number of sessions. However, session duration did not contribute significantly to the clustering. The average inter-session interval in cluster 1 is longer, and the number of game sessions is fewer. Cluster 2 is defined by approximately daily play sessions (22.3 hours between sessions), 8.9 game sessions over 9 days, and a session length of around 13 minutes. These characteristics correspond to the snacking pattern we are looking for. The clustering analysis identified that most players adhered to the snacking pattern. 75% of the players in the Baseline condition fell
### 3.2 Player Attrition: Snacking in Serious Games

After establishing the five snacking mechanics using a casual game, “Serious Snacking” (Alexandrovsky et al., 2021a, F₂) investigated how the snacking framework transfers into the context of a serious game with voluntary play. In contrast to previous work, we were interested in which of the game dynamics facilitates the snacking pattern and how they affect player adherence. Following the approach in the previous study (Alexandrovsky et al., 2019b, F₁), this research designed a game where each of the snacking mechanics could be layered individually. In line with the GOC study design (Alexandrovsky et al., 2019b, F₁), our goal was to design an engaging but not too immersive game in which we could include the snacking mechanics individually. However, our game should have serious attributes (Abt, 1970). Therefore, we designed “Infinitus Algebraica” – a simple math learning mobile game that addresses mental math skills – as our platform for the experiment. We chose this genre as math is free from language hurdles and is required in a wide range of fields (Roman, 2004). As a starting point for...
3.2 Player Attrition: Snacking in Serious Games

Figure 3.4: Relative membership of players in the snacking/non-snacking clusters for each game version (Alexandrovsky et al., 2019b, F1).

As modifiers of difficulty, we applied the grid size (2×2, 3×3, 4×4, 5×5, 6×6), time quota (ca. 30–380 sec.) and ranges of possible values (5–150) as parameters. In total, we generated 350 levels, which are grouped in 35 packages of 10 levels (rounds) each. With the progression through the packages, the players play on larger grids, need to perform more different math operations while having less time to complete the levels. Furthermore, we generated 4 levels of difficulty (easy, normal, hard, and master), at which the game can be replayed. The difficulty levels provide an additional modification on the range of possible values in the grid (i.e., calculating with larger numbers) and time quota; with increasing difficulty, the values becoming larger. To provide a ground for the snacking mechanics, we implemented 13 power-ups with an inventory system. Additionally, we added 8 visual skins and 5 background music tracks to be used as rewards. With two beta testers who had prior experience in game balancing and beta testing, we iteratively adjusted the difficulty levels and balanced the probabilities for the rewards to provide a sufficient curve for a play experience over the course of 3 weeks. The game is available on the Apple App store ³ and Google Play store ⁴. An exemplary gameplay round is depicted in Figure 3.5


and gameplay videos\textsuperscript{5}. The game versions were implemented as follows:

- **Baseline**: The power-ups and the inventory system are disabled. Also, all visual skins, background music, and all levels are unlocked from the beginning.

- **Rewards**: As in Alexandrovsky et al. (2019b, F\textsubscript{1}), rewards were implemented through a performance-based score system. Based on the accuracy, players could receive power-ups.

- **Completion**: In accordance with Alexandrovsky et al. (2019b, F\textsubscript{1}), we developed 48 achievements. For completing missions, the players are rewarded with power-ups, visual skins, and music soundtracks.

- **Novelty**: implements an RPG-like level-up system with 40 ranks. Depending on the performance, the players receive experience points. At milestone levels, the players are rewarded with visual skins, music, unlocks of new difficulty levels, and power-ups.

- **Blocking**: In line with Alexandrovsky et al. (2019b, F\textsubscript{1}), we restricted the playtime to 15 minutes every 8 hours.

- **Waiting**: Like in Alexandrovsky et al. (2019b, F\textsubscript{1}), power-ups are regenerated over time. Every 8 hours, the game selects and spawns new items.

![Figure 3.5: Infinitus Algebraica gameplay (Alexandrovsky et al., 2021a, F\textsubscript{2})](https://www.youtube.com/playlist?list=PLJY6D-LTSQguru1kLxS-J6bAyrj7EjsCw)\textsuperscript{6}, accessed: 20.02.2021.

\textsuperscript{5}Video figures demonstrating the gameplay of Infinitus Algebraica: \url{https://www.youtube.com/playlist?list=PLJY6D-LTSQguru1kLxS-J6bAyrj7EjsCw}, accessed: 20.02.2021.
Table 3.5: Infinitus Algebraica pre-study (Alexandrovsky et al., 2021a, F2): ANOVA results of demographics and motivation between the groups (n=49). M and SE indicate the sample means and standard errors. For TIPI and SIMS, the univariate results of the conducted MANOVAs are presented for individual subscales.

<table>
<thead>
<tr>
<th>Demographics</th>
<th>M (SD)</th>
<th>F(5,43)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>31.57 (1.22)</td>
<td>1.24 (0.31)</td>
<td>0.31</td>
</tr>
<tr>
<td>Identification as a gamer (1[not at all]-10[gamer])</td>
<td>6.99 (0.38)</td>
<td>0.14 (0.98)</td>
<td>0.43</td>
</tr>
<tr>
<td>Enjoy mental math (1[with pleasure]-7[not at all])</td>
<td>5.57 (0.19)</td>
<td>1.05 (0.43)</td>
<td>0.29</td>
</tr>
<tr>
<td>Using maths in daily life (1[every day]-10[not at all])</td>
<td>5.93 (0.04)</td>
<td>1.29 (0.29)</td>
<td>0.24</td>
</tr>
<tr>
<td>Exp. with serious games (1-10) (1[never played]-10[expert])</td>
<td>6.33 (0.04)</td>
<td>1.41 (0.24)</td>
<td>0.65</td>
</tr>
</tbody>
</table>

| Extraversion (1-7) | 4.04 (0.20) | 0.67 (0.65) | 0.99 |
| Agreeableness (1-7) | 4.88 (0.18) | 0.11 (0.99) | 0.64 |
| Conscientiousness (1-7) | 5.21 (0.16) | 1.18 (0.33) | 0.33 |
| Emotional Stability (1-7) | 4.96 (0.17) | 0.15 (0.38) | 0.86 |
| Openness to Experiences (1-7) | 4.92 (0.22) | 0.19 (0.19) | 0.96 |

<table>
<thead>
<tr>
<th>TIPI</th>
<th>M (SD)</th>
<th>F(5,43)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion (1-7)</td>
<td>4.04 (0.20)</td>
<td>0.67 (0.65)</td>
<td>0.99</td>
</tr>
<tr>
<td>Agreeableness (1-7)</td>
<td>4.88 (0.18)</td>
<td>0.11 (0.99)</td>
<td>0.64</td>
</tr>
<tr>
<td>Conscientiousness (1-7)</td>
<td>5.21 (0.16)</td>
<td>1.18 (0.33)</td>
<td>0.33</td>
</tr>
<tr>
<td>Emotional Stability (1-7)</td>
<td>4.96 (0.17)</td>
<td>0.15 (0.38)</td>
<td>0.86</td>
</tr>
<tr>
<td>Openness to Experiences (1-7)</td>
<td>4.92 (0.22)</td>
<td>0.19 (0.19)</td>
<td>0.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SIMS</th>
<th>M (SD)</th>
<th>F(5,43)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic Motivation (1-7)</td>
<td>5.29 (0.21)</td>
<td>0.55 (0.74)</td>
<td>0.47</td>
</tr>
<tr>
<td>Identified Regulation (1-7)</td>
<td>5.22 (0.19)</td>
<td>0.93 (0.47)</td>
<td>0.53</td>
</tr>
<tr>
<td>External Regulation (1-7)</td>
<td>4.04 (0.18)</td>
<td>0.84 (0.53)</td>
<td>0.56</td>
</tr>
<tr>
<td>Amotivation (1-7)</td>
<td>3.04 (0.22)</td>
<td>0.19 (0.19)</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Table 3.6: Infinitus Algebraica pre-study IMI descriptive statistics (Alexandrovsky et al., 2021a, F2).

<table>
<thead>
<tr>
<th>Interest-Enjoyment</th>
<th>Baseline M(SD)</th>
<th>Blocking M(SD)</th>
<th>Novelty M(SD)</th>
<th>Waiting M(SD)</th>
<th>Rewards M(SD)</th>
<th>Completion M(SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.46 (0.86)</td>
<td>5.69 (0.82)</td>
<td>5.77 (0.84)</td>
<td>5.38 (1.43)</td>
<td>5.60 (0.48)</td>
<td>5.36 (0.81)</td>
<td></td>
</tr>
<tr>
<td>Competence</td>
<td>5.76 (1.03)</td>
<td>5.77 (0.64)</td>
<td>6.00 (0.88)</td>
<td>5.70 (1.34)</td>
<td>5.80 (0.50)</td>
<td>5.87 (0.40)</td>
</tr>
<tr>
<td>Effort-Importance</td>
<td>4.23 (1.36)</td>
<td>4.62 (1.33)</td>
<td>4.86 (1.15)</td>
<td>5.96 (0.79)</td>
<td>5.33 (0.91)</td>
<td>4.67 (1.26)</td>
</tr>
<tr>
<td>Tension-Pressure</td>
<td>3.52 (0.84)</td>
<td>4.25 (1.52)</td>
<td>3.43 (1.68)</td>
<td>3.22 (1.45)</td>
<td>4.19 (0.66)</td>
<td>2.94 (1.52)</td>
</tr>
</tbody>
</table>

To increase the ecological validity, we designed the main study as a field study. Therefore, we published the game on mobile platforms (Google Play Store and Apple App Store) and assessed the player behavior along with subjective responses of n=100 voluntary participants over the course of three
Table 3.7: Infinitus Algebraica pre-study (Alexandrovsky et al., 2021a, F2): test statistics for the intrinsic motivation on IMI.

<table>
<thead>
<tr>
<th></th>
<th>(F_{5,43})</th>
<th>(p)</th>
<th>(\eta^2_p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest-Enjoyment</td>
<td>0.27</td>
<td>0.93</td>
<td>0.03</td>
</tr>
<tr>
<td>Competence</td>
<td>0.11</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>Effort-Importance</td>
<td>2.37</td>
<td>0.06</td>
<td>0.22</td>
</tr>
<tr>
<td>Tension-Pressure</td>
<td>1.28</td>
<td>0.23</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Table 3.8: Infinitus Algebraica pre-study (Alexandrovsky et al., 2021a, F2): ANOVA results of performance metrics between the groups. M and SE indicate the sample means and standard errors.

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>SE</th>
<th>(F_{5,43})</th>
<th>(p)-bonf</th>
<th>(\eta^2_p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>unlocked packages</td>
<td>3.531</td>
<td>0.170</td>
<td>0.338</td>
<td>1.000</td>
<td>0.038</td>
</tr>
<tr>
<td># play rounds</td>
<td>8.510</td>
<td>0.596</td>
<td>0.758</td>
<td>1.000</td>
<td>0.081</td>
</tr>
<tr>
<td># fails</td>
<td>25.878</td>
<td>4.466</td>
<td>0.205</td>
<td>1.000</td>
<td>0.023</td>
</tr>
<tr>
<td># wins</td>
<td>135.469</td>
<td>6.820</td>
<td>0.366</td>
<td>1.000</td>
<td>0.041</td>
</tr>
<tr>
<td>best combo</td>
<td>19.469</td>
<td>0.491</td>
<td>2.249</td>
<td>0.932</td>
<td>0.207</td>
</tr>
<tr>
<td>mean score</td>
<td>154.469</td>
<td>5.099</td>
<td>0.658</td>
<td>1.000</td>
<td>0.071</td>
</tr>
<tr>
<td># levels finished</td>
<td>63.367</td>
<td>3.022</td>
<td>0.425</td>
<td>1.000</td>
<td>0.047</td>
</tr>
<tr>
<td># items</td>
<td>0.204</td>
<td>0.082</td>
<td>1.337</td>
<td>1.000</td>
<td>0.135</td>
</tr>
<tr>
<td># skins</td>
<td>1.061</td>
<td>0.035</td>
<td>0.530</td>
<td>1.000</td>
<td>0.058</td>
</tr>
<tr>
<td># music</td>
<td>1.020</td>
<td>0.020</td>
<td>1.028</td>
<td>1.000</td>
<td>0.107</td>
</tr>
<tr>
<td>duration in min.</td>
<td>15.585</td>
<td>1.310</td>
<td>0.711</td>
<td>1.000</td>
<td>0.076</td>
</tr>
<tr>
<td>accuracy</td>
<td>78.755</td>
<td>3.919</td>
<td>0.449</td>
<td>1.000</td>
<td>0.050</td>
</tr>
<tr>
<td># retries</td>
<td>0.592</td>
<td>0.351</td>
<td>0.725</td>
<td>1.000</td>
<td>0.078</td>
</tr>
<tr>
<td># finished pack</td>
<td>0.714</td>
<td>0.065</td>
<td>4.884</td>
<td>0.018</td>
<td>0.362</td>
</tr>
</tbody>
</table>
3.2 Player Attrition: Snacking in Serious Games

Table 3.9: Infinitus Algebraica long-term (Alexandrovsky et al., 2021a, F) (1): demographics, motivation, and group differences (n=76). W subscript indicates Welch-correction in cases of violated homogeneity of variances. M, and SE indicate the sample means and standard errors.

<table>
<thead>
<tr>
<th>Demographics</th>
<th>M</th>
<th>SE</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>28.88</td>
<td>1.17</td>
<td>0.34</td>
<td>0.89</td>
</tr>
<tr>
<td>Identification as a gamer (1[not at all]-100[gamer])</td>
<td>59.96</td>
<td>3.91</td>
<td>0.28</td>
<td>0.85</td>
</tr>
<tr>
<td>Enjoy mental math (1[with pleasure]-7[not at all])</td>
<td>4.70</td>
<td>0.17</td>
<td>0.39</td>
<td>0.86</td>
</tr>
<tr>
<td>Using maths in daily life (1[every day]-100[not at all])</td>
<td>41.40</td>
<td>3.38</td>
<td>0.62W</td>
<td>0.69W</td>
</tr>
<tr>
<td>Exp. with serious games (1[never played]- 100[expert])</td>
<td>38.15</td>
<td>3.03</td>
<td>0.66</td>
<td>0.65</td>
</tr>
<tr>
<td>Freq. playing serious games (1[daily]-8[never])</td>
<td>5.70</td>
<td>0.22</td>
<td>0.80</td>
<td>0.50</td>
</tr>
<tr>
<td>Freq. playing video games</td>
<td>1.87</td>
<td>0.19</td>
<td>1.40W</td>
<td>0.25W</td>
</tr>
<tr>
<td>(1[daily]-8[never])</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extraversion (1-7)</td>
<td>3.91</td>
<td>0.17</td>
<td>0.26</td>
<td>0.93</td>
</tr>
<tr>
<td>Agreeableness (1-7)</td>
<td>4.82</td>
<td>0.12</td>
<td>1.25</td>
<td>0.30</td>
</tr>
<tr>
<td>Conscientiousness (1-7)</td>
<td>4.87</td>
<td>0.15</td>
<td>0.38W</td>
<td>0.86W</td>
</tr>
<tr>
<td>Emotional Stability (1-7)</td>
<td>4.86</td>
<td>0.16</td>
<td>1.07</td>
<td>0.39</td>
</tr>
<tr>
<td>Openness to Experiences (1-7)</td>
<td>4.96</td>
<td>0.11</td>
<td>0.41</td>
<td>0.84</td>
</tr>
</tbody>
</table>

weeks in which participants played the game under uncontrolled conditions. Based on the comments from the pre-study, we modified Infinitus Algebraica. We clarified the instruction, added a tutorial, fixed minor bugs, and tweaked the game balancing. As the mean playtime was around 15 minutes and also considering the lessons learned from the GOC study, to enforce the effect of Blocking, we reduced the playtime to 10 minutes per 8 hours.

The study design resembled the long-term study by Alexandrovsky et al. (2019b, F). On the first day, the participants installed the game from the store and filled out the demographics questionnaires. Afterward, they played the first 2 rounds (ca. 8 minutes) and filled out the IMI and SIMS questionnaires. Then they were allowed to play the game freely, except for the Blocking condition, that restricted the playtime. Every seven days, the participants received the IMI and SIMS questionnaires. On the final day (after 3 weeks), the participants were asked to fill out the exit survey. As we received multiple comments that the players were annoyed by the questionnaires and stopped playing due to the surveys and not due to the game’s design, we disabled the weeks 1, 2, and 3 questionnaires within the study in favor of clearer behavioral log data. Therefore, 9 out of 99 subjects participated in the study without answering the questionnaires.

149 players downloaded the Infinitus Algebraica. 99 participants played the game at least once (Android: n=83, iOS: n=16). The sample decreased drastically over the course of the study. By the end of week 2, the sample consisted of 20 participants, and only 9 participants played for the full 3 weeks. Therefore, planned repeated measure analyses on how the motivation evolved throughout the study could not be performed. Analogously to the pre-study for the questionnaire analysis, we excluded 13 participants who played less than one complete round. 11 subjects were removed because their responses showed zero variance, contradicting answers, or wrong responses to control questions. After filtering, 76 participants (40 self-identified as women, 36 self-identified as men) completed the demographics questionnaire (N: Baseline=14, Blocking=15, Completion=13, Rewards=12, Novelty=10, Waiting=12). 75 participants only filled out the SIMS. On average, the participants were M=28.88 (SD=10.18) years old. There were no significant differences between the conditions on any of the demographic factors. Table 3.9 shows details of the demographic traits.

We conducted one-way MANOVAs on the subscales of IMI and SIMS with the condition as the between-subjects factor assessed on the first day to determine if the game versions differed in short-term motivation. The analysis showed no difference on any scale (cf. Tab. 3.10). For the analysis of performance and behavior, the analysis included 99 participants (Baseline: n=18, Blocking: n=17, Completion: n=15, Novelty: n=15, Rewards: n=19 and Waiting: n=15). In total, the subjects played 512 sessions that are distributed across the conditions as following: Baseline=91; Blocking=47; Completion=169; Novelty=47; Rewards=88; Waiting=70. The play times are depicted in Figure 3.6. On the performance
metrics, we conducted one-way ANOVAs which revealed multiple significant differences, generally showing that in Blocking and Completion, the players progressed slower through the game. The detailed performance and behavior analyses are not determining for this study, but they are provided in the appendix of the corresponding publication (Alexandrovsky et al., 2021a, F_2).

To investigate how the individual mechanics affect the player attrition, we conducted a survival analysis (Collett, 2015; Miller Jr, 2011). Therefore, we grouped each player’s play sessions into bins of the mean inter-session interval (33.75 hours), leading to 32 bins in total. We conducted a Kaplan-Meier analysis (Kaplan and Meier, 1958) on the binned play sessions with the condition as the between-subjects factor. Figure 3.7 shows the predicted survival functions. Table 3.11 depicts the corresponding median survival times for the game versions. To determine how individual performance metrics affect the survival and if the game mechanics facilitate adherence differently, we conducted a Cox regression (Cox, 1972). For the goodness-of-fit, the log-likelihood ratio test was significant ($\chi^2_{11} = 70.83, p<0.001$). The concordance statistics (Harrell et al., 2005) of the model was $c = 0.719$, indicating that the model describes the data well. Table 3.12 summarizes the Cox regression results showing that only the game versions mean score and # levels finished differentiate between the survivals significantly. However, none of the other performance metrics had a significant effect on player attrition. The game versions have a significant effect on the hazard function. Planned contrast of the game versions against Baseline revealed that all game versions except Waiting yielded a negative effect on the hazard function. However, this difference was statistically significant only for Blocking.

All conditions yielded a regular play behavior with a play session duration of around 20 to 30 minutes and regular breaks of around 1 1/2 days (i.e., snacking pattern). The results of the survival analysis show that every mechanic contributed to higher engagement compared to Baseline. However, Completion yielded by far the longest participation with the game. While most other conditions stopped playing after around 10 days, most Completion participants engaged with the game for around 25 days. These results are in accordance with findings from the literature that refer to clear goals and meaning as successful strategies for playful learning experience (Garris et al., 2002; Malone, 1980; Marczewski, 2013). Previous research has identified choice as a powerful motivator (Bowey et al., 2019; Karsh and Eitam, 2015; Leotti and Delgado, 2011). Therefore, we surmise the raised sense of choice of the missions might have enhanced motivation. We assume that the increase of mission variety compared
Table 3.10: Infinitus Algebraica long-term (Alexandrovsky et al., 2021a, F\textsuperscript{2}) - One-Way ANOVA for Intrinsic Motivation and Situated Motivation on the first day.

<table>
<thead>
<tr>
<th>IMI</th>
<th>F\textsubscript{5,70}</th>
<th>p</th>
<th>η\textsuperscript{2}</th>
<th>Baseline M (SD)</th>
<th>Novelty M (SD)</th>
<th>Waiting M (SD)</th>
<th>Rewards M (SD)</th>
<th>Compl. M (SD)</th>
<th>Blocking M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest-Enjoyment</td>
<td>0.23</td>
<td>0.95</td>
<td>0.02</td>
<td>4.74 (0.83)</td>
<td>4.64 (0.71)</td>
<td>4.43 (0.75)</td>
<td>4.75 (0.68)</td>
<td>4.60 (0.94)</td>
<td>4.67 (0.80)</td>
</tr>
<tr>
<td>Competence</td>
<td>0.57</td>
<td>0.72</td>
<td>0.04</td>
<td>4.33 (0.87)</td>
<td>4.74 (0.98)</td>
<td>4.65 (1.06)</td>
<td>4.90 (1.08)</td>
<td>4.82 (1.12)</td>
<td>4.61 (0.63)</td>
</tr>
<tr>
<td>Effort-Importance</td>
<td>0.48</td>
<td>0.79</td>
<td>0.03</td>
<td>3.57 (1.10)</td>
<td>3.80 (1.01)</td>
<td>3.22 (1.13)</td>
<td>3.86 (1.16)</td>
<td>3.46 (1.16)</td>
<td>3.49 (1.31)</td>
</tr>
<tr>
<td>Tension-Pressure</td>
<td>0.25</td>
<td>0.94</td>
<td>0.02</td>
<td>3.64 (0.90)</td>
<td>3.63 (1.42)</td>
<td>3.60 (1.31)</td>
<td>3.42 (1.06)</td>
<td>3.21 (1.05)</td>
<td>3.52 (1.44)</td>
</tr>
</tbody>
</table>

Table 3.11: Infinitus Algebraica (Alexandrovsky et al., 2021a, F\textsuperscript{2}) - Median survival predicted by the Kaplan-Meier estimator.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Median Survival Time</th>
<th>SE</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.0</td>
<td>0.00</td>
<td>[0.00, 1.00]</td>
</tr>
<tr>
<td>Rewards</td>
<td>2.0</td>
<td>0.90</td>
<td>[0.24, 3.76]</td>
</tr>
<tr>
<td>Novelty</td>
<td>0.0</td>
<td>0.00</td>
<td>[0.00, 0.00]</td>
</tr>
<tr>
<td>Completion</td>
<td>5.0</td>
<td>0.93</td>
<td>[3.17, 6.83]</td>
</tr>
<tr>
<td>Waiting</td>
<td>1.0</td>
<td>0.56</td>
<td>[0.00, 2.17]</td>
</tr>
<tr>
<td>Blocking</td>
<td>4.0</td>
<td>0.88</td>
<td>[2.28, 5.71]</td>
</tr>
<tr>
<td>Overall</td>
<td>3.0</td>
<td>0.40</td>
<td>[2.23, 3.78]</td>
</tr>
</tbody>
</table>

to GOC coupled with the overall feeling of accomplishment after completing a mission helped satisfy the needs for autonomy and relatedness, which have been shown numerous times to foster player motivation (Birk et al., 2017). Interestingly, Novelty differs drastically from previous results in the GOC study (Alexandrovsky et al., 2019b, F\textsubscript{1}) and shows almost no influence on the sustained engagement. We explain this result by a lower connection of the novel elements with the game content and with the seriousness of the game. As predicted in the GOC study, in Blocking, the participants played significantly more sessions than Baseline, although Interest-Enjoyment was rated lower. These results support Nacke and Drachen’s (2011) argument that PX and adherence are not necessarily related.

### 3.3 Playful User-Generated Treatment

Simple phobias such as acrophobia (the fear of heights) or claustrophobia (the fear of closed spaces) affect around 7−9% of the population in western countries (Association and Association, 2013). Simple phobias can evoke panic and reduce overall well-being (WHO, 2019). Commonly, afflicted individuals tend to avoid spaces or situations where their phobia could be triggered. Therefore, in many cases, a therapy is desirable. The most common therapy for acrophobia (and many other phobias) is Exposure Therapy (ET) (Neudeck and Wittchen, 2005). ET can follow a paradigm of immediate extreme or gradual exposure, with the latter being more common. Over the course of such a therapy, the participants get gradually exposed to stronger stimuli of their phobia (e.g., experiencing stepping on exceedingly higher platforms). The gradual exposure therapy teaches the patients to cope with
the anxious situations and also to gradually lower the experienced intensity of the stimulus and the physiological response to it. However, due to its reliance on physical stimuli, exposure therapy may be challenging to conduct (Eaton et al., 2018). For instance, it is often difficult to travel to places with specific heights to provide the required degree of exposure. Therefore, a considerable and growing body of work provides evidence that ET can successfully take place in VR – Virtual Reality Exposure Therapy (VRET). It has been shown that virtual exposure can be as effective as real exposure (Côté and Bouchard, 2008; Emmelkamp et al., 2002; Emmelkamp et al., 2001; Parsons and Rizzo, 2008) and that VRET can successfully be applied in treatment (Emmelkamp et al., 2002; Powers and Emmelkamp, 2008). In some instances, VRET has been shown to be more effective (Coelho et al., 2009; Klinger et al., 2005; Krijn et al., 2007) and enjoyable than conventional exposure therapy (Coelho et al., 2009; Garcia-Palacios et al., 2007; Garcia-Palacios et al., 2001). In this way, VRET overcomes a range of barriers that traditional therapy faces, such as logistics and safety. Moreover, VRET allows for efficient and personalized therapy plans (Lindner et al., 2017; Powers and Emmelkamp, 2008) and can relieve therapists (Coelho et al., 2009; Gorini and Riva, 2008).

Like in other health intervention domains, mental health treatment suffers from low adherence to the therapy. Although many people suffer from simple phobias, 60 – 80% of the afflicted never seek treatment (Agras et al., 1969; Boyd et al., 1990; Magee et al., 1996) and of those who start a therapy, approximately 25% refuse exposure or drop out. Frequently cited reasons are “attitudinal barriers” (Andrade et al., 2014) as the avoidance of treatment is similar to the avoidance symptoms of the disorder itself, that the subjects do not attribute their disorder as too disabling, or that subjects had negative experiences with treatment (Andrade et al., 2014; Botella et al., 2011; Boyd et al., 1990; Garcia-Palacios et al., 2007). There is a shared consensus that motivational strategies from game design might help to motivate a therapy or support its adherence. A significant body of work in the field of GUR established serious games and gamification as effective approaches to fostering motivation for learning (Breuer and Bente, 2010; Connolly et al., 2012; Denny et al., 2018), physical activities (Barathi et al., 2018; Smeddinck et al., 2014), work (Dale, 2014; Hamari et al., 2014) and therapy (Garcia-Palacios et al., 2007; Mandryk and Birk, 2017; Smeddinck, 2016; Smeddinck et al., 2015). However, there has been little research that systematically analyzed the effects of game elements on mental health (Birk et al., 2019). Moreover, Johnson et al. (2016) reviewed literature that reports empirical evidence on the effect of gamification on health. The authors identified 10 gamification elements and pointed out that “not a single study captured game design elements on intrinsic motivation (e.g., motivation to exercise)”
### Table 3.12: Statistics and estimates of the Cox-regression (Alexandrovsky et al., 2021a, F2). Coef: Regression coefficient; predicted change in the hazard. Coef SE: Standard error of the regression coefficient. Hazard ratio Exp(Coef): Indicator of how the observed parameter changes the hazard rate. CI for Hazard Ratio: The confidence interval for the hazard ratio indicates the plausible range of the hazard ratio. \( \chi^2_{Wald} \): Test statistics determines how well the parameter fits to describe the hazard. Df: Degrees of freedom for the \( \chi^2_{Wald} \) test. p: p-value of the \( \chi^2_{Wald} \) test determines if the test statistics is significant.

<table>
<thead>
<tr>
<th></th>
<th>Coef</th>
<th>Coef SE</th>
<th>Hazard Ratio</th>
<th>95% CI for Hazard Ratio</th>
<th>( \chi^2_{Wald} )</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean score</td>
<td>0.003</td>
<td>0.001</td>
<td>1.003</td>
<td>[1.001, 1.006]</td>
<td>5.926</td>
<td>1</td>
<td>0.015</td>
</tr>
<tr>
<td>accuracy</td>
<td>0.468</td>
<td>0.598</td>
<td>1.597</td>
<td>[0.494, 5.156]</td>
<td>0.612</td>
<td>1</td>
<td>0.434</td>
</tr>
<tr>
<td># wins</td>
<td>-0.003</td>
<td>0.002</td>
<td>0.997</td>
<td>[0.993, 1.001]</td>
<td>1.814</td>
<td>1</td>
<td>0.178</td>
</tr>
<tr>
<td># fails</td>
<td>-0.006</td>
<td>0.005</td>
<td>0.994</td>
<td>[0.985, 1.004]</td>
<td>1.469</td>
<td>1</td>
<td>0.226</td>
</tr>
<tr>
<td># levels finished</td>
<td>0.014</td>
<td>0.005</td>
<td>1.014</td>
<td>[1.004, 1.024]</td>
<td>7.222</td>
<td>1</td>
<td>0.007</td>
</tr>
<tr>
<td>best combo</td>
<td>0.013</td>
<td>0.008</td>
<td>1.013</td>
<td>[0.997, 1.030]</td>
<td>2.423</td>
<td>1</td>
<td>0.120</td>
</tr>
</tbody>
</table>

- (Johnson et al., 2016) but rather game designs around rewards which build on points and achievements are dominating. Cole et al. (2015) describe such game elements that address physical or cognitive skill as **functional challenges**. In contrast, **emotional challenges** confront players with emotionally salient material through narratives, frightening scenarios, strong characters, or difficult choices players have to make (Cole et al., 2015; Denisova et al., 2017). These game elements provide enjoyment through resolutions of tensions within the narratives and overcoming negative emotions (Bopp et al., 2018; Cole et al., 2015; Endress et al., 2016; Gowler and Iacovides, 2019). To close this gap, “Playful User-Generated Treatment” (Alexandrovsky et al., 2020b, F3) investigates engaging game design for mental health therapy in VR on the explanatory use case of acrophobia treatment.

As a starting point for our research, to identify general requirements for the game design, we first conducted two semi-structured interviews with practicing therapists who had substantial expertise in the treatment of phobias using traditional ET. The goal of the interviews was to understand the general procedures of ET and to identify prescriptive target points for engaging game design. The interviews covered the following four general themes that served as a starting point for the subsequent: Techniques and Procedures, Setting and Scenarios, Tasks and Motivation, and Supplemental.

The interviews were analyzed following the qualitative content analysis method (Mayring, 2000) and investigated the data both inductively and deductively. To ensure the validity of the coding process, two authors processed the material independently. Cases of disagreement were resolved in discussions. Finally, the analysis identified five requirements of engaging game design for VRET: (i) Motivation: The VRET should emphasize the sense of autonomy. (ii) Communication: The VRET system should enable direct communication between them and their patients. (iii) Habitation: The scenarios should allow patients to reach habituation. (iv) Non-distracting tasks: The game design should build around tasks that do not distract the patients from the exposure. (v) Physiological symptoms: The VRET should not have technical flaws that might cause cybersickness, resulting in physiological symptoms that are not related to the exposure itself.

To address these requirements, for acrophobia therapy – as an archetypal example of a simple phobia – in VR, we developed Playful User-generated Treatment (PUT): a two-step approach where users first can shape and design their terrain in table-top mode (the design phase); i.e., the degree of the threatening stimulus (cf. Fig. 3.8a) and then expose themselves to the very same terrain at full-scale realistic heights (exposure phase); cf. Fig. 3.8b. Central to the approach is to empower patients designing their “fear” in a simulation (designing the exposure at a top-down view in a miniature map) before they
undergo an exposure with their fearful stimuli (view the terrain at full-scale). This method was derived from related literature on game design for mental health (Gerling et al., 2012; Mandryk and Birk, 2017), theories of motivation (Ryan and Deci, 2000a) and constructivist learning theories (Marone, 2016; Papert and Harel, 1991). For the design phase, we created a terrain editor that employed game elements from sandbox games like Minecraft (Mojang and Markus „Notch“ Persson, 2009) and sculpting methods from tools for 3D content creation such as Tilt Brush (Google and Skillman & Hackett, 2016) or Blender (OnlineCommunity, 2018). For the terrain-shaping, users were interacting using the VR controllers with a commonly applied laser pointer metaphor (Jerald, 2016; LaViola et al., 2017). We attached a body-anchored (Alexandrovsky et al., 2020a; Rzayev et al., 2019) User Interface (UI) at the controller of the non-dominant hand. From there, users get help or instructions and can select assets (e.g., spawn points, buildings, nature, or characters) to place in the terrain for decoration and personal customization. The asset library consists of 6 exemplary assets: trees, rocks, grass, bushes, stumps, and wooden cottages. Additionally, users can place spawn points as entry points for the exposure phase. The exposure phase resembles VRET examples from existing literature (cf. Emmelkamp et al., 2002; Powers and Emmelkamp, 2008). Users can enter the terrain in full scale at the spawn points and undergo exposure to the heights.

To validate the viability of the requirements and provide empirical evidence for the engaging characteristics of the PUT approach, we designed and conducted a user study with \( n=31 \) non-acrophobic subjects, which investigates the effect of content creation on player experience and height perception. The study design consisted of 2 conditions and included the two phases of PUT. First, between subjects, the participants were in the terrain editor scene. For the design phase in the PUT condition, the participants were instructed to design a terrain with three exposures at different heights levels and decorate it with virtual objects. For the three exposures, we requested the participants to design a hill of medium size (30 m), one of high size (50 m), and a 70 m tower building. In the CTRL group, the participants only were allowed to see a comparable pre-designed terrain, which contained three similar elevations for exposure along with some decoration assets. After the editor scene, both groups switched perspectives and entered the terrain at full-scale (exposure phase), where they were repeatedly exposed to the three different heights in random order. To assess how the terrain-shaping affected the intrinsic motivation, the participants filled out the first IMI (Deci and Ryan, 2003) after the design phase. Anxiety was measured using Subjective Units of Distress-Scale (SUDS) (Antony et al., 2005; Back et al., 2015) and State-Trait-Anxiety-Inventory (STAI) (Spielberger, 1983) after each task in the exposure phase. After the three exposures, the participants stated their affect using Positive and Negative Affect Scale (PANAS) (Crawford and Henry, 2004) and filled out a second IMI. The surveys were conducted on a terminal displaying the in-VR questionnaires (inVRQs).

Empirical user studies with phobics are ethically problematic since they require experienced support in case of a panic. Therefore, this early-stage research evaluated the approach with convenient subjects.
However, Robillard et al. (2003) compared emotional reactions of phobics and non-phobics to different anxiety-inducing stimuli (heights, spiders, and enclosed spaces) in VR and showed that both groups react to the stimuli similarly, differing only in the degree of reaction; with phobics being stronger affected. All subjects were pre-screened using the Acrophobia Questionnaire (AQ) (Cohen, 1977) to exclude participants showing tendencies for acrophobia. An Anxiety Score of $>45.45$ and an Avoidance Score of $>8.67$ were determined as thresholds to exclude subjects from the experiment as it is one standard deviation below the mean score for clinical acrophobia (Antony et al., 2005; Cohen, 1977). 31 participants (25% self-identifying as female) volunteered for our study. The mean age was 24.32 years ($SD=4.32$). None of the participants showed clinical tendencies for acrophobia ($M=18.87$, $SD=11.67$, avoidance: $M=3.55$, $SD=2.36$). 20 participants experienced VR once before; the others had no prior experience with VR. The groups were balanced for gender ($U=133.5$, $p=0.49$), age ($t_{29}=1.159$, $p=0.25$), avoidance ($t_{29}=-1.03$, $p=0.31$) and anxiety ($t_{29}=-0.73$, $p=0.47$).

Anxiety was assessed using STAI and SUDS after each condition. To determine if the condition affected anxiety, for both measures, we performed Mixed-factorial ANOVAs (MF-ANOVAs) with the three exposures as within-subjects factor and condition as the between-subjects factor. As Table 3.13 depicts, the MF-ANOVAs show no significant differences and no interaction effects (cf. Fig. 3.9a). Further, we investigated how the anxiety evolved over the course of the study. We conducted MF-ANOVAs with anxiety and assessment number (STAI and SUDS) as within factors and condition as between factor. Both MF-ANOVAs showed significant main effects ($SUDS: F_{3.34.67.95}=17.69$, $p<0.01$, $\eta^2_p=0.38$, Greenhouse-Geisser corrected $\epsilon=0.78$; STAI: $F_{3.58}=4.90$, $p=0.01$, $\eta^2_p=0.02$) but no significant differences between conditions nor interaction effects. Subsequent post-hoc tests revealed significant differences between platform$_1$ and platform$_3$ on both anxiety measures ($SUDS: t_{30}=2.73$, $p=0.03$ mean diff.$=3.32$, see Fig. 3.9b; STAI: $t_{30}=2.82$, $p=0.03$ mean diff.$=3.58$). For SUDS, all platforms were significantly more anxiety-inducing than the baseline (platform$_1$: $t_{30}=-5.46$, $p<0.01$, Cohen’s $d=-0.98$; platform$_2$: $t_{30}=-4.99$, $p<0.01$, Cohen’s $d=-0.90$; platform$_3$: $t_{30}=-5.54$, $p<0.01$, Cohen’s $d=-1.00$). These results indicate that there were perceivable differences between the different levels of exposures in-line with the intended effects of height exposure. Simultaneously, there are no strongly notable differences in the resulting exposure phases between conditions, as intended for the game design. To investigate if the terrain-shaping influences the affect, we conducted independent sample t-tests between the two conditions on PANAS. The positive affect ($M=37.61$, $SD=4.39$, $t_{29}=0.75$, $p=0.46$, Cohen’s $d=0.27$) as well as the negative affect ($M=14.03$, $SD=5.55$, $t_{29}=-0.99$, $p=0.33$, Cohen’s $d=0.36$) did not differ significantly. This arguably provides further corroborative evidence to comparable exposures. The results for PANAS are depicted in Figure 3.9c.

To investigate the effect of terrain-shaping on motivation, we conducted MF-ANOVAs on all IMI subscales as within-subjects factor and condition as between-subjects factor (cf. Fig. 3.9d, Tab. 3.14). The analysis revealed significant differences within subjects only on Tension-Pressure. Bonferroni-corrected post-hoc t-tests confirmed this difference ($t_{29}=-5.80$, $p<0.01$, Cohen’s $d=1.04$). Further, MF-ANOVAs showed a significant difference between the conditions and an interaction effect of Interest-Enjoyment. A Bonferroni-corrected post-hoc comparison of Interest-Enjoyment between $C_{PUT}$ and $C_{CTRL}$ was significant ($t_{29}=2.40$, $p=0.02$, Cohen’s $d=0.43$). There was a significant interaction effect of condition x Interest-Enjoyment between $C_{PUT}$ and $C_{CTRL}$ in the first assessment ($t_{29}=3.90$, $p<0.01$, Cohen’s $d=0.70$) as well as between first and second measurement in $C_{CTRL}$ ($t_{29}=-3.17$, $p=0.02$, Cohen’s $d=0.57$).

The user study results show an increase on Interest-Enjoyment for the PUT (cf. Fig. 3.9d). This

Table 3.13: PUT Mixed-factorial ANOVA of anxiety measures SUDS and STAI for all 3 exposures (Alexandrovsky et al., 2020b, F3).

<table>
<thead>
<tr>
<th></th>
<th>SUDS</th>
<th>STAI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F_{2.58}$</td>
<td>$\eta^2_p$</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.46</td>
<td>0.63</td>
</tr>
<tr>
<td>Condition</td>
<td>0.11</td>
<td>0.89</td>
</tr>
</tbody>
</table>
Table 3.14: Mixed-factorial ANOVA for both IMIs assesses (Alexandrovsky et al., 2020b, F$_3$).

<table>
<thead>
<tr>
<th></th>
<th>$C_{PUT}$</th>
<th>$C_{CTRL}$</th>
<th>IMI</th>
<th>Condition</th>
<th>IMI × Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M (SD)</td>
<td>M (SD)</td>
<td>$F_{1,29}$</td>
<td>$p$</td>
<td>$\eta^2_p$</td>
</tr>
<tr>
<td>Competence</td>
<td>5.18 (0.72)</td>
<td>5.22 (0.71)</td>
<td>4.12</td>
<td>0.05</td>
<td>0.12</td>
</tr>
<tr>
<td>Tension-Pressure</td>
<td>2.34 (1.22)</td>
<td>2.46 (1.12)</td>
<td>32.87</td>
<td>&lt; 0.01</td>
<td>0.53</td>
</tr>
<tr>
<td>Effort-Importance</td>
<td>4.69 (1.12)</td>
<td>4.55 (1.01)</td>
<td>2.25</td>
<td>0.14</td>
<td>0.07</td>
</tr>
<tr>
<td>Interest-Enjoyment</td>
<td>5.79 (0.97)</td>
<td>5.10 (0.89)</td>
<td>0.59</td>
<td>0.45</td>
<td>0.02</td>
</tr>
</tbody>
</table>

(a) State-Trait-Anxiety-Inventory (STAI). The platforms are grouped by type.

(b) Subjective Units of Distress-Scale (SUDS). The platforms are grouped by order of exposure.

(c) Positive and Negative Affect Scale (PANAS)

(d) Intrinsic Motivation Inventory (IMI). Comp: Competence; Pres: Tension-Pressure; Eff: Effort-Importance; Int: Interest-Enjoyment. The subscript number indicates the assessment of motivation 1=post design phase; 2=post exposure phase.

Figure 3.9: Subjective self-reports on motivation, anxiety and affect (Alexandrovsky et al., 2020b, F$_3$).

The whiskers indicate the standard deviation.

indicates that self-creation was perceived as playful and motivating. The comparable anxiety between the groups (cf. Fig. 3.9a and 3.9b) indicates that PUT does not interfere with the exposure, and therefore, the approach seems to be a suitable game design element for exposure therapy in VR.

Expert Perspectives on PUT

As an additional validation of the PUT concept, “Playful User-Generated Treatment: Interviews” (Volkmar et al., 2020, S$_3$) conducted an online survey with $n=13$ practicing therapists where they watched a 3 min long video explaining VRET and PUT and responded to questions regarding the applicability and further considerations of the approach. The majority of the experts were positive about the
approach. They highlighted that the active creation of the treatment might support the therapy and suggested that this approach may be translated to other disorders. Based on these findings, we surmise that this design-pattern built around empowerment and creativity is a valuable design direction that may overcome limitations of other designs for emotional challenges such as narratives as they might be difficult to balance for a non-distracting and non-frustrating experience (Gowler and Iacovides, 2019).

3.4 Discussion

Concerning RQ1, this chapter investigated how game design can be effectively applied to facilitate UE in form of player retention and emotional connection with the gameful experience. While most research on gameful/playful design tends to focus on game elements that enhance enjoyment within the game session by addressing functional challenges; i.e., designing towards flow experience (Cole et al., 2015; Johnson et al., 2016), this work investigated how game design can affect players emotionally and trigger psychological needs that make them return to the game.

The snacking framework contributes five game mechanics that address engagement in the intensity of the experience (Novelty, Rewards) and the temporal dimension beyond a single session, which triggers disengagement and re-engagement (Waiting, Blocking). Completion is a game mechanic that contributes to the intensity of the experience within the play session, working as an immediate reward mechanism and temporally by providing unfinished tasks (Mäntylä and Sgaramella, 1997; Syrek et al., 2017) which pull players back to playing. The snacking framework extends the existing frameworks on game design (e.g., MDA Hunicke et al., 2004) by adding a temporal component to the game design, which addresses the pace of playing as an additional design dimension. Contrasting MDA, the snacking framework moves designers not to think about the affective outcome (Aesthetics) the game should produce and at which pace and with what intensity. Instead, the framework provides game elements that designers can implement into their games to facilitate specific – for example, regular but brief – player behavior. However, we (Alexandrovsky et al., 2021a, F2) identified that in the context of serious games for learning, meaningful goals (i.e., Completion) are more effective to promote adherence rather than rewarding mechanisms that build around functional challenges (e.g., Instant Rewards). In line with related work on games for health (Baranowski et al., 2008), we (Alexandrovsky et al., 2020b, F3) identified that “traditional” game design might be unfavorable as they can distract from the intervention that could cause disengagement as it would negatively affect the effectiveness of the treatment. Therefore, we developed the PUT approach, which relies on creative tasks that empower players (patients) to playfully engage with their fears while not being disturbed by the game elements. The PUT design addresses O’Brien and Toms’ (2008) period of engagement and extends existing taxonomies of game design for mental health (e.g., Cheek et al., 2015; Johnson et al., 2016) with creativity and self-empowerment as a motivational strategy of engagement. While aiming for the same goal to facilitate sustained engagement, this approach stands aside to the snacking framework. It builds an alternative approach to engaging design for use cases that may benefit from self-empowering the players. Future work should combine these two design directions into a unifying theory of engaging gameful design.

In response to RQ1, we conclude that depending on the context of the game, the game elements affect engagement differently. We surmise that in casual games, players are driven by the curiosity about the novel elements (Lomas et al., 2017) and by idle-like mechanics (Fizek, 2018) (i.e., Waiting). In contrast, engagement with serious games relies on the connection between the game elements and the serious objectives such as learning outcomes or therapy progress (Malone, 1980; Tondello et al., 2018). In the case of Infinitus Algebraica, this connection was established through achievements that constantly reflect how the players develop their math skills. For VRET, the meaning was evoked through the active and creative design of the personal exposure. Therefore, to provide a sustained long-term engagement for the gameful experience with a purpose (e.g., games for change), the design should go beyond challenges that address the players skills (e.g., scores badges or simple rewards), but rather incorporate elements that provide meaning to the player beyond the in-game currency.
A significant and growing body of work has developed theories and artifacts which underline that haptic interaction supports the users’ enjoyment and may facilitate the ease of use through its directness and embodiment. However, haptic interaction – as a form of Reality-based Interaction (RBI) (Jacob et al., 2008) – may also conflict with the expressive power or the practicality of the interface. Moreover, how Haptic Experiences (HXs) (Kim and Schneider, 2020) affect the users in VEs remains an open question in HCI. Therefore, following RQ2, this chapter investigates how haptic interaction in engaging environments influences the users and how it can be applied to facilitate UE. This research thread examined static and passive shape-changing physical props in different exemplary interaction scenarios for training, learning, and creative play. Section 4.1 looks at passive, static props in a VR climbing application and investigates how HX affects the users’ anxiety. Section 4.2 discusses the effects of tangible interaction on usability and performance in a multimodal math-learning application. Section 4.3 investigates passive shape-changing props and presents a sand-shaping interface in VR for creative modeling. Finally, Section 4.4 summarizes the main findings and discusses how they contribute to RQ2.

4.1 Passive Physical Props for VR Climbing

Dealing with the fear of falling in sport climbing is challenging. However, it is crucial to overcome the fear of falling as it impedes athletes’ movements (Hörst, 2017) and negatively affects their performance (Hardy and Hutchinson, 2007). Climbers learn to overcome their fear of falling by building confidence in training and, for instance, by habituation through controlled falling (Lloyd, 2014). Fear of falling is associated with general height-induced anxiety, which manifests as acrophobia in its pathological form. The most common method to treat this condition is exposure therapy, which is the “gold standard” in psychotherapy (Powers and Emmelkamp, 2008). As discussed in Chapter 3.3, VRET received much attention in the past decades and has been shown as a viable method to treat phobias under safe conditions and with a degree of control a traditional exposure in the physical reality would not allow. Although VRET has been shown to be effective, controller-based VR might not be suitable for climbing scenarios, as it does not allow for full-body movements that athletes would normally perform during climbing. Therefore, physical props in conjunction with VR might help to provide the safety of VRET while supporting natural climbing movements. However, how interaction with physical props – i.e., tangible objects that allow for haptic sensation during interaction – affects anxiety and the sense of presence has not been closely examined.

To investigate how physical props affect the sense of presence and anxiety, in publication “Physical Props in VR” (Schulz et al., 2019, F₁) we conducted a user study with \( n=28 \) participants that applied physiological and subjective measures of UX in a climbing scenario. The study employed a within-subjects design with three conditions: (i) real climbing in 10 m height \( (C_{real}) \) (ii) physical climbing in VR with physical props at a real height of 30 cm and a virtual height of 10 m \( (C_{props}) \) (iii) climbing
in VR using controllers ($C_{ctrl}$). The setup is depicted in the video preview\(^1\) and Figure 4.1. Each of the 3 conditions consisted of three phases: The participants horizontally traversed a climbing wall, from a start point to a return point (phase #1). To ensure that the participants are aware of their exposition to height, at the return point in phase #2, the participants threw a ball down to the ground and read out the number displayed next to where the ball dropped. Finally, in phase #3, the subjects traversed back to the starting point. After each task, the participants reported their experience on multiple scales that measure anxiety, physical exertion, and the sense of presence. The order of the conditions was randomized.

In line with several studies from the literature (Slater, 2004; Slater, 2007) that suggested physiological measures as indicators of presence, we obtained Heart Rate (HR) and Heart Rate Variability (HRV) (Castaldo et al., 2015; Hynynen et al., 2009) – employed as mean R-R interval (mRRI) – using electrocardiography (ECG), Skin Conductivity Response (SCR) (Brouwer and Hogervorst, 2014; Meehan, 2001) from an Electrodermal Activity signal and Respiratory Rate (RR) from a chest strap. Biosignals are prone to physical activity (Croft et al., 2004). Therefore, in addition to physiological measures, we assessed subjective ratings of physical exertion, anxiety, and presence. As a subjective measure of exertion, we employed Rating of Perceived Exertion (RPE) (Borg, 1982) on a labeled Visual

Analog Scale (VAS) ranging from 6–20. Similarly, as a complementary subjective measurement of anxiety, we employed the Anxiety Thermometer (AT) (Houtman and Bakker, 1989) on a continuous Visual Analog Scale (VAS) with the range 0 (not anxious at all) to 10 (extremely anxious). Subjective ratings of presence were assessed using the Igroup Presence Questionnaire (IPQ) (Schubert et al., 2001) only after \( C_{\text{props}} \) and \( C_{\text{ctrl}} \) conditions. The measurements of anxiety, physical exertion, and the biosignals were analyzed using ANOVAs with repeated measures (RM-ANOVAs) with subsequent post-hoc t-tests. Since the sense of presence was assessed only in the two VR conditions, the analysis was performed using a paired-sample t-test. The descriptive plots along with the F and t-statistics are depicted in Figures 4.2 and 4.3. All subjective measures show that \( C_{\text{props}} \) provided the most intense experience. On the IPQ, \( C_{\text{props}} \) received a significantly higher score for realism over \( C_{\text{ctrl}} \) (cf. Fig. 4.2c). Both, on the AT (Fig. 4.2a) and RPE (Fig. 4.2b) \( C_{\text{props}} \) was rated higher over the other conditions. Figure 4.3b shows that HR differed between all conditions. As expected, the HR was lowest in the resting phase \( C_0 \). In the task conditions, average HR was in the following order: \( C_{\text{ctrl}} \), \( C_{\text{props}} \), \( C_{\text{real}} \). The Heart Rate Variability (cf. Fig. 4.3c) was in \( C_{\text{props}} \) lowest and was the only one that differed significantly from baseline \( (C_0) \). The RR was below baseline in all trials. However, there was no significant difference between the climbing conditions on RR (cf. Fig. 4.3d). On the Electrodermal Activity signal, we used the conductivity level (cf. Fig. 4.3e) and the rolling average for the number of peaks in time windows of 20 seconds (Fig. 4.3f) as an indicator of SCR. The skin conductivity showed only in \( C_{\text{probs}} \) an effect of condition below the baseline. Similarly, SCR peaks in \( C_{\text{real}} \) and \( C_{\text{props}} \) were significantly below baseline. In conjunction, the subjective ratings and the physiological responses show evidence that for whole-body tasks in VR, realistic physical interaction and haptic feedback have a strong effect on the sense of presence and anxiety.
4 Haptic Interaction in Engaging Environments

(a) Duration: $F_{2,44}=9.09$, $p<0.01$, $\eta^2=0.29$.

(b) Heart Rate (HR): $F_{3,63}=92.88$, $p<0.01$, $\eta^2=0.82$.

(c) mean R-R interval (mRRI): $F_{3,60}=4.33$, $p<0.01$, $\eta^2=0.18$.

(d) Respiratory Rate (RR): $F_{3,66}=28.16$, $p<0.01$, $\eta^2=0.56$.

(e) Skin Conductivity Response (SCR): $F_{3,56}=2.56$, $p=0.064$, $\eta^2=0.12$.

(f) Skin Conductivity Response peaks: $F_{3,63}=3.98$, $p=0.012$, $\eta^2=0.16$.

Figure 4.3: Physiological measurements (Schulz et al., 2019, F4).

4.2 Passive Physical Props for Math-learning

The supplementary work “Math Tangibles” (Alexandrovsky et al., 2018, S1) was situated in the context of the Federal Minestery of Education and Research (BMBF) funded research project Multimodal Algebra Larning (MAL) (Janßen et al., 2017) which investigates how TUIs (Ishii, 2008) facilitate the understanding of algebra. To investigate how tangible interaction supports algebra learning, we developed a mobile application that employs the Algebra Tiles concept (Sharp, 1995) with the objects represented as physical tiles tracked by the front-facing camera (cf. Fig. 4.4a). Algebra Tiles are a didactical tool that builds on Bruner’s (1966) concrete-representational-abstract model and the constructivistic objects-to-think-with approach (Papert and Harel, 1991) which suggest that especially for
4.2 Passive Physical Props for Math-learning

(a) Algebra Tiles application with tracked tangibles. The table-top standing tablet is equipped with a mirror at the front-facing camera to track the space in front of the tablet. The physical tiles are arranged on the playfield in front of the tablet. The play area with the physical tiles is mirrored on the tablet screen.

(b) Comparison of the UMUX scores between the tangible and the touch version of MAL, $t_{20}=2.15$, $p=0.044$, Cohen’s $d=0.916$

Figure 4.4: Setup and results of the Algebra Tile study (Alexandrovsky et al., 2018, S$_1$).

beginners, representing abstract concepts with concrete physical objects helps to conceptualize abstract phenomena.

A between-subjects study with $n=22$ seventh-graders (age 11–13) Alexandrovsky et al. (2018, S$_1$) compared the usability and the performance between the physical version and touch only variant. The task for each group was to solve 5 algebra problems at different degrees of difficulty. The problems included simple equations where the participants were required to transform the equation to solve for $x$ and word problems which required building the equations from the description before solving. At the end of each session, we conducted short interviews with the pupils targeting questions on how they proceeded when they worked with the algebra tiles and if the haptic tiles supported the problem-solving process.

As figure 4.4b shows, the usability scores on the Usability Metric for User Experience (UMUX) scale are significantly higher for the tangible condition. However, the performance, i.e., execution time ($t_{108}=-0.533581$, $p=0.595$, Cohen’s $d=0.101$) and the number of correctly solved problems ($\chi^2=0.309$, $p=0.588$, Cramer’s $V=0.053$), do not differ between the versions. In the interviews, most pupils stated that they first tried to solve the equations mentally and then validate their calculations with the tiles. 8 pupils in the tangible condition and 4 participants in the touch condition stated that the tiles helped them visualizing the steps and allowed them for exploration. The other participants mentioned that they preferred a symbolic representation of the mathematical expressions. Only one pupil in the tangible condition expressed that latency of the tracking and the indirectness of the physical props was an issue. The remaining 10 participants did not consider indirectness and latency as technical limitations. These findings are in line with literature on engaging and multimodal learning (cf. Ibrahim and Jaafar, 2009; Moreno and Mayer, 2007), allowing us to surmise that TUIs facilitate user engagement and exploration in learning environments without negatively affecting the performance.
4 Haptic Interaction in Engaging Environments

4.3 Creative Interaction Sand in VR

Most people relate to playing with sand from their childhood days. This unstructured play opens room for experimentation, exploration, or cooperation and helps to develop important skills like creativity, proprioceptive sensing, body and space awareness (Jarrett et al., 2010). Natural materials like sand have also been investigated as part of interactive systems (cf. Döring et al., 2013). These materials are used because they inherently offer rich multimodal feedback and often provoke emotional associations. Related work has explored sandboxes and their tangible affordances in AR environments; also AR sandboxes have been used for interaction with geographical phenomena (Ishii et al., 2004; Piper et al., 2002) or for playful terrain exploration and design (Beckhaus et al., 2008; Couture, 2017; Reed et al., 2014; Roo et al., 2017). While these setups already offer a great direct interaction with sand as a natural material, they are limited in their visualization and interaction capabilities.

Our work “VRBox” (Fröhlich et al., 2018, Fig. 4.5) moves a step further and presents an interactive sandbox for playful and immersive terraforming that combines augmented sandboxes with state-of-the-art VR technology and mid-air freehand interaction. This work investigated creative play with sand as a passive shape-changing input device in VR. The setup consists of a 140×80×30 cm sandbox which is tracked by 3 down-facing Kinects (Newcombe et al., 2011) anchored at the top railing and an HTC Vive VR headset with a HMD mounted Leap Motion for hand tracking. The setup is depicted in Figure 4.5a. In the VR application, the terrain is represented in a blocky Minecraft (Mojang and Markus „Notch“ Persson, 2009) like style on a 98×59 mesh with 32 height levels (cf. Fig. 4.5b). This resolution provides a sufficient compromise between the level of detail and latency. Based on the sand’s height, the visualization is mapped to different earth layers (stratum). This allows for terraforming of visually rich and plausible VEs. Along with the sand shaping, the setup supports interaction with virtual objects. Using freehand interaction (i.e., grabbing gesture), the users can select decorative objects from a menu and place them into the environment (cf. Fig. 4.5). Further, the setup allows switching perspectives from table-top mode to first-person view (Fig. 4.5c). By looking up, the users are scaled down and placed inside the terrain where they can experience the environment at a realistic full scale. Video figure2 illustrates the interaction with the VRBox. The interaction with the VRBox is closely related to the PUT design (cf. Section 3.3, Alexandrovsky et al., 2020b, F3) and translates the approach for haptic interaction in VR.

---

To evaluate the interaction concept and to guide further development, we (Fröhlich et al., 2018, F5) conducted a preliminary user study with $n=9$ (8 male, 1 female; self-identified) participants who had expertise in computer graphics ($n=4$), game design ($n=3$), HCI ($n=1$), and education ($n=1$). The age of the participants ranged from 20–39 years. 8 participants had prior experience in VR and three subjects develop VR applications on a regular basis.

The main task of the study was to play around with the interface. In preparation for each pass, we watered the sand with 1.5–2.0 liter to keep the humidity and firmness of the sand as consistent as possible. Further, we smoothed the sand to a flat surface to provide a “blank canvas” for each participant. At the beginning of each run, we explained the interface and the functionality. After that, the participants were asked to shape a terrain that contained at least one hill, a water body, and one lowland area. Further, the users should adjust the lighting mood and place some virtual objects in the terrain. The participants interacted with the VRBox for ca. 30 min. Afterwards, they rated the system on the System Usability Scale (SUS) (Brooke, 1996) using an out-VR questionnaire (outVRQ) on paper and were interviewed about their experience.

On the SUS, the setup received an average score of $M=78.6$, $SD=9.45$, which according to to Bangor et al. (2009) lies in the range within *good to excellent* usability. The interviews were targeting different areas of general experience, technical development, and use case scenarios. The participants’ responses confirmed the high SUS scores. All participants expressed the experience to be creative and enjoyable with overall good usability. However, in some cases, the resolution of the blocky terrain was criticized. In contrast, others suggested that the low resolution fits well and opens up space for imagination. The participants suggested multiple application scenarios around terraforming and discussed opportunities as well as limitations of the interface and the sand as material. These positive findings indicate this hybrid interaction design with passive shape-changing materials such as sand and VR to appear warranted for further development and study.

**VRBox v2**

Based on feedback in the pre-study, in publication “VRBox V2” (Alexandrovsky et al., 2019a, S2) we further iterated the development on the VRBox (cf. Fig. 4.6). We reduced the box size to $90 \times 60 \times 90$ cm box and exchanged the Kinects with two Intel RealSense D435i, which offer a greater resolution. The visualization of the terrain was updated from a blocky heightmap to a full 3D mesh generated using Marching Cubes (Lorensen and Cline, 1987) with compute shaders. This visualization method allows for full 3D scans of the sand and improves the overall resolution of the terrain to $64^3$ voxels at a VR suitable framerate above 90 FPS. These technical improvements extend the capabilities and the responsiveness of the interface allowing for more detailed sand shaping with higher fidelity. The updated VRBox was presented at the CHI 2019 interactivity venue, confirming that it is mobile and suitable for public spaces. The interaction with the setup is depicted in the video figure.

4.4 Discussion

In response to RQ2, this chapter investigated how haptic interaction affects UE in VEs. The aggregated results suggest that haptic interaction can create a bridge between the physical and virtual space and, therefore, intensify the experience with the interactive systems. Decomposing the findings with respect to the dimensions of UE, the results clearly show that haptic interaction primarily affects the point of

---

3 VRBox v2 demo video: [https://www.youtube.com/watch?v=GVsbnWb8cEk&h=7s](https://www.youtube.com/watch?v=GVsbnWb8cEk&t=7s), accessed: 20.02.2021.
engagement and the period of engagement stages. Haptic interaction affects most experiential qualities associated with these two stages (cf. O’Brien and Toms, 2008 in Fig. 2.6). Haptic interaction directly addresses aesthetics and sensory appeal, which may raise interest and positive affect during the interaction. The presented results are in line with Kim and Schneider (2020) who considered aesthetics beyond visual appearance with the “touchability” as a hedonic property of haptic experiences. In the case of the VRBox (Fröhlich et al., 2018, F5), the findings show that albeit high-fidelity resolution is welcomed and should be approached, low-fidelity (abstract) mapping between the physical and the virtual world can engage users through stimulating their creativity. These results confirm and refine previous work that suggested that full-body interaction in games facilitates engagement (Bianchi-Berthouze et al., 2007). With about 6 million shipped units in 20194, VR has barely reached the end-users; and mostly early adopters possess an HMD. Therefore, VR itself is mostly considered a novel experience (Rivu et al., 2021). Nevertheless, results from our work (Alexandrovsky et al., 2018, S1 and Fröhlich et al., 2018, F5) show that haptic interaction can evoke a sense of novelty and curiosity. Similarly, as the climbing task results by Schulz et al. (2019, F4) show, HX affects anxiety and the sense of presence. These results allow inferring that haptic interaction affects both attention and involvement. As Papert and Harel (1991) proposed, physical props can draw attention and afford playful exploration. This is further exemplified by the responses from the interviews in the VRBox user study Fröhlich et al. (2018, F5). The aggregated results on the usability metrics from Alexandrovsky et al. (2018, S1) and Fröhlich et al. (2018, F5) allow us to infer that passive haptic props can achieve high usability and as the raised anxiety in the climbing study (Schulz et al., 2019, F4) shows, haptic props provide additional feedback that may improve the effectiveness of (whole-body) training applications in VR. From the aggregated results presented in this chapter, we infer that haptic interaction can be utilized to intensify the experience. Responding to RQ2, we conclude that haptic interaction has a strong effect on the users’ engagement and TUIs are a valuable approach to facilitate engagement with both playful and utilitarian interactive systems.

Assessing User Engagement in Immersive Environments

Despite their deficits, the predominant assessment method of UX metrics such as presence in VR user studies are post-experience questionnaires (Schwind et al., 2019; Skarbez et al., 2017; Slater and Steed, 2000). Mostly, the questionnaires are administered in the physical reality (Alexandrovsky et al., 2020a, F6). Therefore, to fill out the surveys, the participants usually leave VR after the experience. However, our work (Putze et al., 2020, F7) argues that Breaks in Presence (BIPs) (Slater and Steed, 2000) produced by switching realities to answer questionnaires might lead to disturbances (Jerald, 2016) and negative emotions (Scherer and Ellgring, 2007) which in turn may bias the subjective ratings.

As the third line of research with the aim to improve the quality of research methods for VR user studies, work presented in this chapter follows RQ3 and investigates the embedding of questionnaires into the VE. This assessment method promises to minimize BIPs and allows users to stay closer to the experience without the need to disengage. This assessment method promises to minimize the disruption with the ongoing experience and therefore, counteract the disengagement and support the re-engagement with the VR experience. However, the understanding of the effects of question-asking in VR is fundamentally incomplete. Moreover, to date, there are no explicit guidelines on the design of in-VR questionnaires (inVRQs). This chapter presents a series of studies investigating the interaction design of inVRQs (Section 5.1). Section 5.2 presents a study that investigates how this assessment method affects the disruptions (i.e., BIPs) through subjective self-reports in VR user studies. Section 5.3 closes up the chapter with a discussion on embedded questionnaires as a design strategy to facilitate UE in VEs and to minimize disengagement (RQ3).

5.1 In-VR Questionnaire Design

In four studies, publication “InVRQ Design” (Alexandrovsky et al., 2020a, F6) investigated the design space of in-VR questionnaires (inVRQs). First, we conducted a literature review of VR user studies to identify general procedures of questionnaire administration. We searched the proceedings of ACM CHI, ACM CHI PLAY, ACM VRST, IEEE VR, and IEEE 3DUI for the time period 2016–2019. The search included papers that contained the following keywords: “virtual reality”, “head-mounted display”, “virtual environment”, “user study”, or “questionnaire”. For inclusion, we read the abstracts of the papers and include those which mentioned a VR setup with HMDs and explicitly speak of some form of evaluations with users. Our sample consisted of n=123 publications. 110 papers used questionnaires in their VR user studies. 77/110 did not report the administration method of questionnaires. 13/110 stated that the surveys were filled outside VR but did not report how they were filled out (e.g., on PC or paper). 15 publications used inVRQs in their studies. The reviewed instances of inVRQs diverse drastically in their degree of usability and presentation style. For example, Kang et al. (2018) and Cao et al. (2018) used a full-screen overlay to display the questionnaires. In contrast, Fernandes and Feiner (2016) and Oberdörfer et al. (2019) used 3D UIs anchored in the environment. Wienrich et al. (2018) also used 3D UIs, however the questionnaire was referenced to the users’ body. This diversity underlines that there is no shared agreement on the presentation methods of inVRQs and demonstrates a gap in the validity or trustworthiness of the research methods.

To narrow down the design space of VR-embedded questionnaires, in the second study, we conducted
an expert online survey that targeted opinions and considerations regarding the administration of questionnaires in VR. 67 VR experts and researchers from 13 different countries filled out the online survey. On a 6-point Likert-scale, the experts indicated to be rather experienced with designing user studies (M=3.76, SD=1.18). 56 experts were directly involved in the design of 1−5 VR user studies in the past 24 months, 8 experts with 6−10, and 3 experts indicated to be involved in more than 10 VR user studies. 29/67 experts have tried inVRQs in their study setups. These responses indicate that our participants have diverse and sound expertise on VR user research allowing us to consider the responses as an expert evaluation. Figure 5.1a shows the usefulness of inVRQs on a 6-point Likert scale (0−5). To determine if prior experience with inVRQs has an effect on the usefulness rating, we split the sample into 2 groups. On average, experts who tried inVRQs before rated the usefulness higher (M=3.72, SD=1.33) compared to the group with on experience (M=2.39, SD=1.55). A two-sided independent sample t-test confirmed this difference (t_{65}=3.69, p<.01, Cohen’s d=0.45). Further, the survey asked about concerns regarding inVRQs. We collected a broad range of reasons against using inVRQ including technical challenges, implementation effort, fear of biases, and participant overload.

To investigate objections raised by the experts and identify suitable presentation methods for inVRQs, we conducted a user study. The design study examined the usability of different anchoring and interaction modalities for inVRQ. Based on guidelines from the literature on 2D UIs (Dix, 2009; Shneiderman et al., 2016) and UIs in VR (Dingler et al., 2018; Google Developers, 2017; Oculus, 2019; Zayer, 2019), we developed 4 variants of a VR questionnaire tool. The interfaces differ in 2 dimensions: anchoring (world-referenced and body-referenced) and interaction modality (pointer and trackpad). The world-referenced UI is anchored in the VE and users can freely move around the interface. A body-referenced
5.1 In-VR Questionnaire Design

Table 5.1: Design study statistics (Alexandrovsky et al., 2020a, F6). Descriptive statistics for the four questionnaire designs on the two dimensions anchoring and interaction modality. WP: World Pointing, WT: World Touch, BP: Body Pointing, BT: Body Touch.

<table>
<thead>
<tr>
<th></th>
<th>WP (M, SD)</th>
<th>WT (M, SD)</th>
<th>BP (M, SD)</th>
<th>BT (M, SD)</th>
<th>F(3,27)</th>
<th>p</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUS</td>
<td>91.25 (8.99)</td>
<td>64.25 (24.09)</td>
<td>79.00 (14.49)</td>
<td>52.25 (19.52)</td>
<td>13.30</td>
<td>&lt;0.01</td>
<td>0.60</td>
</tr>
<tr>
<td>t (min)</td>
<td>0.61 (0.12)</td>
<td>1.00 (0.27)</td>
<td>0.69 (0.33)</td>
<td>1.19 (0.23)</td>
<td>14.99</td>
<td>&lt;0.01</td>
<td>0.62</td>
</tr>
<tr>
<td># clicks</td>
<td>10.60 (2.17)</td>
<td>10.30 (2.79)</td>
<td>10.40 (3.06)</td>
<td>11.00 (3.74)</td>
<td>0.16</td>
<td>0.83</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>rank</td>
<td>1.40 (0.52)</td>
<td>3.30 (0.82)</td>
<td>1.80 (0.79)</td>
<td>3.50 (0.53)</td>
<td>18.11</td>
<td>&lt;0.01</td>
<td>0.67</td>
</tr>
</tbody>
</table>

(a) in-VR questionnaire Likert scale question. (b) in-VR questionnaire multi-selection question. (c) The archery game.

Figure 5.2: Screenshots of the in-VR questionnaire (inVRQ) and the archery game (Alexandrovsky et al., 2020a, F6).

UI is attached to a hand-held controller. The interaction with the UI varies between a laser pointer – as Oberdörfer et al. (2019) applied for their inVRQs – and a clicking interaction where the trackpad of the controller is used to navigate through the questionnaires, similar to the interface by Schwind et al. (2019). The questionnaire tool supports 5 common types of responses: continuous values (slider), checklists, radio lists, dropdowns, and switches. The four variants were evaluated in a 2×2 within-subjects design with 2 independent variables (anchoring and interaction modality). To evaluate the usability of all response types, we developed a questionnaire that asks for common knowledge facts. These questions were chosen to ask subjects easy-to-answer but objective questions, which allow for calculating the correctness of the responses. Each condition consisted of a questionnaire that comprised all 5 question types once. In total, each participant answered 20 questions with each question type once per condition. For each condition, we logged the time required to fill out the questionnaire and the click events. After each condition in VR, the participants took off the HMD and filled out a SUS (Brooke, 1996) questionnaire on paper.

n=10 male (age M=29.9, SD=2.9) participants from the local campus volunteered for the design pre-study. All participants had prior experience in VR and game development. On all measures, we conducted RM-ANOVAs with the condition as the within-subjects factor. Table 5.1 summarizes the descriptive statistics and the outcomes of the RM-ANOVAs. The results showed overall acceptable usability of the tool (cf. Fig. 5.1b). Further, all measures confirm that world-anchoring with pointing interaction is the most usable design choice for inVRQs.

After establishing world-anchoring with pointing interaction as the most suitable design for inVRQs, we conducted a second user study, where we compared the UX between inVRQs and outVRQs using a VR archery game. The study aimed to evaluate whether the concerns regarding usability and duration regarding inVRQs raised by the experts hold and to provide guidance on question-asking methods in VR. Based on the results of the design study, we implemented the inVRQs with world-based anchoring and laser pointer interaction (cf. Fig. 5.2a). For the outVRQs, we employed LimeSurvey (LimeSurvey Project Team / Carsten Schmitz, 2012) on a 15” notebook with an external keyboard and mouse. The
### Table 5.2: User study descriptives and results of the paired sample t-tests (Alexandrovsky et al., 2020a, F6).

<table>
<thead>
<tr>
<th></th>
<th>inVRQ M (SD)</th>
<th>outVRQ M (SD)</th>
<th>t(37)</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>time (min)</td>
<td>6.87 (2.67)</td>
<td>6.32 (2.30)</td>
<td>1.22</td>
<td>0.231</td>
<td>0.22</td>
</tr>
<tr>
<td>game rating (1-10)</td>
<td>7.89 (1.71)</td>
<td>8.00 (1.73)</td>
<td>-0.52</td>
<td>0.457</td>
<td>0.06</td>
</tr>
<tr>
<td>bow control (1-10)</td>
<td>7.82 (1.66)</td>
<td>8.00 (1.77)</td>
<td>-0.78</td>
<td>0.438</td>
<td>0.11</td>
</tr>
<tr>
<td>UMUX</td>
<td>78.07 (18.03)</td>
<td>86.18 (9.55)</td>
<td>-2.95</td>
<td>&lt;0.01</td>
<td>0.56</td>
</tr>
<tr>
<td>NASA-TLX</td>
<td>19.36 (11.22)</td>
<td>14.41 (9.43)</td>
<td>4.09</td>
<td>&lt;0.01</td>
<td>0.48</td>
</tr>
<tr>
<td>Mental Demand</td>
<td>23.82 (16.29)</td>
<td>20.66 (15.08)</td>
<td>1.423</td>
<td>0.163</td>
<td>0.20</td>
</tr>
<tr>
<td>Physical Demand</td>
<td>19.61 (15.48)</td>
<td>10.66 (13.01)</td>
<td>4.143</td>
<td>&lt;0.01</td>
<td>0.63</td>
</tr>
<tr>
<td>Temporal Demand</td>
<td>11.97 (11.94)</td>
<td>10.26 (10.2)</td>
<td>1.159</td>
<td>0.254</td>
<td>0.15</td>
</tr>
<tr>
<td>Performance</td>
<td>25.26 (26.63)</td>
<td>20.13 (26.98)</td>
<td>1.483</td>
<td>0.146</td>
<td>0.20</td>
</tr>
<tr>
<td>Effort</td>
<td>20.13 (17.22)</td>
<td>12.63 (11.9)</td>
<td>3.009</td>
<td>&lt;0.01</td>
<td>0.51</td>
</tr>
<tr>
<td>Frustration</td>
<td>15.39 (12.7)</td>
<td>12.11 (9.27)</td>
<td>1.769</td>
<td>0.085</td>
<td>0.30</td>
</tr>
<tr>
<td>IPQ GP</td>
<td>4.53 (0.89)</td>
<td>4.61 (0.95)</td>
<td>-0.68</td>
<td>0.499</td>
<td>0.09</td>
</tr>
<tr>
<td>IPQ SP</td>
<td>4.87 (0.73)</td>
<td>4.88 (0.73)</td>
<td>-1.05</td>
<td>0.917</td>
<td>0.01</td>
</tr>
<tr>
<td>IPQ INV</td>
<td>3.89 (1.29)</td>
<td>3.78 (1.13)</td>
<td>0.857</td>
<td>0.397</td>
<td>0.09</td>
</tr>
<tr>
<td>IPQ REAL</td>
<td>3.12 (1.14)</td>
<td>3.00 (1.12)</td>
<td>1.01</td>
<td>0.320</td>
<td>0.11</td>
</tr>
</tbody>
</table>

The study aimed to simulate an archetypal VR experiment. To provide a realistic study setting, we designed a simple balloon archery game (cf. Fig 5.2c).

To evaluate the workload and the usability of the inVRQ tool, we employed the raw TLX (Hart, 1986) and UMUX (Finstad, 2010) after each condition. As an objective performance metric, the exact time for filling out the questionnaires was logged. The study followed a within-subjects design with the conditions inVRQs and outVRQs in randomized order. In each condition, the participants played a 90 seconds round where they should hit as many balloons as possible. The study setup is depicted in the video figure\(^1\). As a primary outcome for the quality of the VR experience, we obtained presence using IPQ (Schubert et al., 2001). Furthermore, the participants rated the game and the perceived control over the bow on a 10-ticks slider. To provide a greater variety of question types, we additionally included questions about the VE (1× numerical, 4× single choice with 2–5 items, 4× multiple choice with 5 and 16 items).

Our sample consists of \(n=38\) participants (age: \(M=27, SD=10.8, 16\) male, 22 female). The participants share a broad range of prior VR experience: 6 use VR regularly, 27 occasionally, and 5 never used VR before. 19 previously participated in other VR user studies and 3 participants used inVRQs before.

The analysis was performed using paired sample t-tests. Table 5.2 depicts the descriptive statistics and the t-test results. The time to fill out the questionnaires was around 6–8 minutes, with VR being slightly slower. However, this difference was not significant. In the post-experiment interviews, 5 participants stated that filling out the questionnaires in VR was faster, and 2 had the impression the questionnaires in VR would be shorter. In both conditions, the game and the bow control received a score of around 8/10 with no effect of condition on the rating. The UMUX scores show that both questionnaire types provide high usability, with outVRQs being slightly better rated (cf. Fig. 5.1d). The lower usability score in VR can be explained by some comments given in the interviews that sometimes the participants needed to click multiple times on a UI element for selection. According to Grier (2015), the workload was perceived low to medium with Physical Demand and Effort rated significantly higher in VR (cf. Fig. 5.1c). This could be attributed to the fact that they were standing and using the VR controller in mid-air rather than sitting on a chair and using the mouse with a resting

It appears crucial to investigate the side-effects of question-asking in VR user studies as researchers need to be aware of biases that may exist for their research methods of choice. Therefore, after identifying a usable design for inVRQs, we were interested in how question-asking in VR affects the sense of presence and if this assessment method reduces the Break In Presence (BIP). To investigate if inVRQs counteract the disruption of the study flow, in publication “InVRQ BIP” (Putze et al., 2020, F_7) we performed a systematic investigation into the effects of interrupting the VR experience through questionnaires and switching between realities using physiological data as a continuous and objective measure of presence (Skarbez et al., 2017; Slater et al., 2003). In a mixed-factorial user study with two independent variables (questionnaire modality and fidelity), we evaluated question-asking procedures after 3 short rounds of a VR shooter game. For questionnaire modality, within-subjects, the participants rated their PX on a PENS (Ryan et al., 2006) and the sense of presence on IPQ (Schubert et al., 2001) using either inVRQs or outVRQs. Since physiological responses are highly specific and can vary drastically between participants, we chose a within-subjects design with repeated measures for the questionnaire modality. This also allowed the participants to compare both assessment methods. The order of the questionnaire conditions was randomized. Fidelity is operationalized as the degree of immersion (Brogni et al., 2006; Khanna et al., 2006; Schwind et al., 2019) with two levels LoFi and HiFi with altered visual and auditory fidelity. It has been shown that visual fidelity affects immersion and thus, affects the sense of presence (Bouchard et al., 2012; Bowman and McMahan, 2007; Slater et al., 2009; Zimmons and Panter, 2003). Therefore, we hypothesize that a switch from LoFi to physical reality would cause a smaller BIP compared to switching from HiFi. To avoid transfer effects, fidelity was administered between-subjects. Our primary point of the investigation were BIP events caused by the disruption of the game to fill out the questionnaires or a switch in realities. In line with previous work by Liebold et al. (2017), we employed a 3s blackout at the initial play rounds as a reference BIP event to compare with
5 Assessing User Engagement in Immersive Environments

Figure 5.4: Flowchart of the study procedure with the randomized within-subjects variable questionnaire modality (inVRQ, outVRQ) and fidelity (HiFi, LoFi) as between-subjects conditions. States 2 and 4 in each flow diagram consist of the steps A-H. Questionnaires in states D and F are conducted in or out of VR in the respective condition. Black and grey states (B, D, F) indicate the investigated BIP events. The IPQ at the end of each condition is assessed only outside VR.

The study setup is depicted in Figure 5.4 in the video figure². The study procedure consisted of two conditions (inVRQ and outVRQ) with the following states: (1) Study preparation: Briefing. Random assignment to a fidelity condition (HiFi or LoFi) and the order of the questionnaire modality. Attach physiological sensors and synchronize biosignals with the game. Put on the HMD. (2) First questionnaire condition, steps (A)-(H) (inVRQ or outVRQ). (3) Optional 2 minute break. (4) Second questionnaire condition, steps (A)-(H) (inVRQ or outVRQ). (5) Conclusive questionnaire on a PC and debriefing. Each questionnaire condition contained the following steps: (A) Put on the HMD and play

a tutorial round (60s). (B) Initial 3s blackout BIP. (C) Game round #1 (90s). (D) PENS #1 using inVRQ or outVRQ depending on the condition. (E) Game round #2 (90s). (F) PENS #2 using same questionnaire modality as PENS #1. (G) Game round #3 (90s). (H) Take off the HMD and fill out an IPQ outside VR on PC.

53 subjects, mostly students from computer science and related areas from the local university, volunteered to participate in the user study. All participants qualified for an Amazon voucher lottery. Due to technical issues, three participants were excluded. Our sample consists of \( n=50 \) participants (8 female, 42 male, self-identified) for the analysis. The mean age was \( M=26.08, SD=4.39 \). One participant suffered from red-green color blindness, 45 participants were right-handed, 3 left-handed, and 2 bi-manual who interacted with their right hand. 20 participants had a vision correction. The four study groups were balanced for gender. The mean game experience was \( M=2.76, SD=2.17 (1–8 \text{ scale, 1 max}) \), and the mean VR experience was \( M=6.48, SD=2.22 (1–8 \text{ scale, 1 max}) \). One-way ANOVAs showed no significant differences across conditions for VR experience (\( F_{1,46}=0.01, p=0.92, \eta^2=0.0002 \)) nor for game experience (\( F_{1,46}=0.15, p=0.70, \eta^2=0.003 \)).

In all four variants, the participants shot around 50–60 drones per round (HiFi inVRQ: \( M=53.89, SD=10.93, \text{LoFi inVRQ: } M=60.05, SD=11.50, \text{HiFi outVRQ: } M=54.64, SD=11.63, \text{LoFi outVRQ: } M=59.99, SD=11.88 \)). A MF-ANOVA with fidelity as a between-subjects factor and questionnaire modality as within-subjects factor revealed a significant difference between the fidelities (\( F_{1,24}=5.71, p=0.025, \eta^2=0.19 \)) but no effect of condition and no interaction effects. The completion time of the game experience, including the time for filling out the PENS questionnaires, was on average 742.78 seconds (\( SD=78.31 \)) for the outVRQ condition and 689.90 seconds (\( SD=87.00 \)) for the inVRQ condition. A paired sample t-test showed that this difference was significant (\( t_{49}=3.71, p<0.001, \text{Cohen's } d=0.75 \)). This faster fill-out time for inVRQs underlines that these assessment methods can improve the overall study flow. The results of IPQ and PENS show that the game yields a sufficient sense of presence and a moderate PX (cf. Fig 5.5). To investigate the effect of questionnaire modality and fidelity on the sense of presence, we performed MF-ANOVA s on the subscales of the IPQ with questionnaire modality as a within-subjects factor and fidelity as between factor were performed. The analysis of IPQ showed no significant main effects or interaction effects for both independent variables. To determine if questionnaire modality or fidelity had an effect on the PX, we conducted MF-ANOVA s with assessment number and the questionnaire modality as a within-subjects factor on all PENS subscales. The analysis showed significant differences between the first and second assessment for Autonomy (\( F_{1,48}=5.89, p=0.02, \eta^2=0.11 \)), Competence (\( F_{1,48}=4.41, p=0.04, \eta^2=0.08 \)) and Intuitive Control (\( F_{1,48}=8.90, p<0.01, \eta^2=0.15 \)). However, we did not find any significant effects or interactions on questionnaire modality or fidelity. Bonferroni-Holm corrected post-hoc t-tests between outVRQ1 and outVRQ2 showed a significant decrease of Autonomy (\( t_{49}=2.33, p=0.02, \text{Cohen's } d=0.34 \)) and significant increases for Intuitive Control (\( t_{49}=-2.33, p<0.01, \text{Cohen's } d=-0.18 \)). As suggested by the
Table 5.3: Responses to all BIP event for the EDA, HR, HRV, RR signals split by the fidelity condition (Putze et al., 2020, F_7). The outVRQ event is divided into the occurrence of the notification and the HMD take off event. For each event, the mean difference between the baseline window before the event and the test window after the event and the result of the paired-sampled t-test are given. P-values are corrected with Benjamini-Hochberg correction for each signal (row). For each value: Mean difference (t-statistic), \(*p<0.01, *p<0.05.

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>Blackout_1</th>
<th>Blackout_2</th>
<th>inVRQ_1</th>
<th>inVRQ_2</th>
<th>outVRQ_1</th>
<th>outVRQ_2</th>
<th>outVRQ_1</th>
<th>outVRQ_2</th>
<th>notification HMD off</th>
<th>notification HMD off</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EDA</strong></td>
<td>23</td>
<td>0.05(2.23)*</td>
<td>0.12(2.85)*</td>
<td>0.04(2.21)*</td>
<td>0.04(2.18)*</td>
<td>0.05(2.28)*</td>
<td>0.16(2.32)</td>
<td>0.09(2.57)*</td>
<td>0.25(2.50)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HR</strong></td>
<td>24</td>
<td>-7.46(-2.38)</td>
<td>-4.09(-1.31)</td>
<td>-5.91(-2.38)</td>
<td>0.64(0.23)</td>
<td>-2.65(-0.58)</td>
<td>-10.55(-2.15)</td>
<td>1.78(-0.65)</td>
<td>1.89(0.38)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HRV</strong></td>
<td>24</td>
<td>-5.38(-1.68)</td>
<td>-1.16(-0.43)</td>
<td>-4.94(-1.72)</td>
<td>-1.63(-0.66)</td>
<td>-3.98(-1.22)</td>
<td>5.65(0.96)</td>
<td>1.85(0.89)</td>
<td>9.00(1.56)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RR</strong></td>
<td>24</td>
<td>-3.58(-3.02)</td>
<td>-2.13(-1.93)</td>
<td>-2.51(-1.58)</td>
<td>-2.86(-1.76)</td>
<td>-0.12(-0.08)</td>
<td>-0.10(-0.07)</td>
<td>1.84(-1.21)</td>
<td>-2.68(-1.75)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**LoFi**

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>Blackout_1</th>
<th>Blackout_2</th>
<th>inVRQ_1</th>
<th>inVRQ_2</th>
<th>outVRQ_1</th>
<th>outVRQ_2</th>
<th>outVRQ_1</th>
<th>outVRQ_2</th>
<th>notification HMD off</th>
<th>notification HMD off</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EDA</strong></td>
<td>24</td>
<td>0.06(2.49)*</td>
<td>0.15(3.01)*</td>
<td>0.15(3.08)*</td>
<td>0.08(2.56)*</td>
<td>0.11(2.12)*</td>
<td>0.28(2.39)</td>
<td>0.07(1.60)</td>
<td>0.20(2.81)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HR</strong></td>
<td>24</td>
<td>-2.23(-0.52)</td>
<td>0.67(0.23)</td>
<td>1.58(0.46)</td>
<td>-2.21(-0.67)</td>
<td>1.69(0.63)</td>
<td>-1.88(-0.37)</td>
<td>2.89(0.87)</td>
<td>2.42(0.56)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HRV</strong></td>
<td>24</td>
<td>-3.80(-1.46)</td>
<td>-2.27(-0.58)</td>
<td>-4.46(-1.44)</td>
<td>-1.43(-0.78)</td>
<td>1.49(0.47)</td>
<td>14.37(2.47)</td>
<td>2.21(-1.03)</td>
<td>8.79(1.39)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RR</strong></td>
<td>24</td>
<td>0.55(0.40)</td>
<td>-2.27(-1.28)</td>
<td>-4.02(-4.01)*</td>
<td>3.45(-4.01)*</td>
<td>-1.72(-1.31)</td>
<td>2.56(1.94)</td>
<td>1.07(0.99)</td>
<td>3.52(3.36)*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

F-MANOVA, the post-hoc tests indicated a trend for Competence (t_{49} = -1.95, p=0.06).

For the detection and comparison of the BIP events on the biosignals, we followed the methods by Liebold et al. (2017) and Slater et al. (2006). We expected a physiological reaction to the blackouts and the inVRQ immediately after the events. The analysis of outVRQs events is less straightforward and included two events: At the end of a game round, a text message appeared in VR that notified the participants to contact the experimenter for taking off the HMD. Taking off the HMD took on average \(M=13.78s\) (\(SD=3.85\)). Therefore, the analysis took both the occurrence of the “contact the experimenter” notification (red line, Fig. 5.6c) and the HMD-off event (black line) as post-BIP events into account. Each questionnaire condition contained three different BIP events, leading to a total of six BIP events for every participant. For the biosignals analysis, we compared the time frames of a baseline window before the event with the respective intervals after the BIP events using Benjamini-Hochberg corrected paired sample t-tests for all six BIP events for both HiFi and LoFi conditions. Table 5.3 summarises the outcomes of the comparisons for all signals.

HR and HRV was derived from the BVP signal. To detect the effect on the HR after an event, we took a window of 3 s after the event and compared it with a 3 s baseline window before the event. For the outVRQ events, the notification and HMD-off event were compared separately. In both fidelity conditions, neither HR nor HRV showed any effects. As suggested by Liebold et al. (2017), for the RR, we averaged and compared 10 s before and after the events. However, for the BVP signals, the Benjamini-Hochberg corrected paired sample t-tests showed no significant differences for the blackout events in both fidelities. For EDA, we followed the procedure by Liebold et al. (2017) and analyzed the phasic skin conductance response (Benedek and Kaernbach, 2010) as an indicator for event-related activity. One participant was excluded from the EDA analysis due to an unusually high baseline signal. For every BIP event, we compared the signal from a time interval of seconds 3 to 6 after the event with a 3 s baseline window right before the task using paired samples t-tests. Figure 5.6 shows the EDA response for all BIP events averaged over all participants. The dashed green lines illustrate the investigated intervals in Figure 5.6. All BIP events in both conditions showed a significant increase of the EDA responses between the investigated time frames. In line with Liebold et al. (2017) and Slater et al. (2006), the detection of the BIP events was robust for the EDA signals for both fidelities. Therefore, we base the further BIP analysis on the EDA signals. To examine the recovery of the EDA signals after the BIP events, we applied a 3 s sliding window for a timeframe up to 60 seconds after the BIP until the signal dropped to baseline. For each window, we conducted a paired sample t-test between the window time frame and the baseline split by questionnaire modality and fidelity. For the blackout events and inVRQs, the signal stabilized after 3 to 12 seconds. In contrast, the recovery from
5.3 Discussion

the outVRQs took by far the longest. The EDA signal was significantly above the baseline window for over 45 seconds. The corresponding t-values are visualized in Figure 5.6d. To determine whether the questionnaire modality affects the subsequent task, we compared the average EDA signals during a task directly after the conductance of the questionnaires in the previous task. We conducted a two-way ANOVA with questionnaire modality and fidelity as factors to investigate the differences in physiological responses dependent on the preceded questionnaire modality. The ANOVA revealed a main effect on questionnaire modality ($F_{1,191}=28.30, p<0.001, \eta^2=0.99$) but no main effect on fidelity and no interaction effects. The mean difference for outVRQs was $M=0.30, SD=0.53$ and for inVRQs $M=0.004, SD=0.41$. A post-hoc test confirmed the effect ($t_{193}=-5.35, p<0.001$, Cohen’s $d=0.77$).

A comparison of the biosignals between inVRQs and outVRQs during the actual filling out was not performed, as both tasks’ activity is structurally and behaviorally too different (Kristal-Boneh et al., 1995).

Neither the game experience nor prior experience in VR affected the ratings on IPQ and PENS. This asserts that the random assignment was successful. The planned interruptions were perceived as much more disturbing than the uncontrolled factors. This indicates a valid study flow. Furthermore, the plausible ratings on IPQ and PENS confirm that the design of the VE fulfilled its intended purpose and facilitated an engaging VR experience. Neither HR nor RR could robustly detect BIP events. In contrast, the physiological reactions on the EDA signals could clearly identify the BIPs events. The analysis of the EDA signals revealed that answering questionnaires in VR yields a weaker and shorter EDA response in contrast to outVRQs (cf. Figures 5.6b and 5.6c). Interestingly, this difference is not affected by the degree of visual fidelity. This indicates that the switch between realities produces BIPs independent of the visual realism. Although we did not observe a difference in performance between the questionnaire modalities, we reckon that the raised physiological activity with outVRQs might negatively affect the performance for more complex tasks. Further, the subjective ratings on Player Experience (PX) showed no significant differences between inVRQs and outVRQs indicating that inVRQs are a viable assessment method that does not produce unintended biases in the responses.

5.3 Discussion

Measuring the experience using subjective self-reports in immersive environments can be challenging since the disruptive nature of questionnaires can break the experience. This BIP is further facilitated by a change of medium, i.e., leaving VR to respond to the surveys in the physical world. Knibbe et al. (2018) discusses the “moment of exit” as an opportunity for the design of engaging VR experiences. Work presented in this chapter is in line with Knibbe et al.’s (2018) design dimensions for exiting the experience and shows that administering questionnaires in VR diminishes the BIP and raises the UE. Therefore, in response to RQ3, we argue that the measurement methods of immersive experiences should adapt to the medium and try to diminish the disengagement from the ongoing experience through embedded measurement methods. From an interaction design perspective, we (Alexandrovsky et al., 2020a, F6) identified prescriptive guidelines for the design of embedded questionnaires, which provide high usability without negatively affecting the quality of the users’ responses. In sum, these findings refine the evaluation methods of UE and move the research agenda forward towards standardized tools of subjective self-reports in VR. Moreover, this research line contributes actively to the ongoing discussion in the research community on the improvement of research methods for UX using reality-altering technology. Recent work already picked up on this discussion started adapting this measuring method (cf. Fan et al., 2020; Feick et al., 2020; Graf and Schwind, 2020; Toet et al., 2020).

Furthermore, we (Alexandrovsky et al., 2021b, S4) contribute an upcoming CHI 2021 workshop which brings researchers and practitioners from all fields together to discuss and revise measurement methods of UX for VR and AR research.
5 Assessing User Engagement in Immersive Environments

(a) Blackout

(b) inVRQ

(c) outVRQ

(d) t-values for the recovery tests. The black horizontal dashed lines indicate the critical value at for p<0.05 and 23 df.

Figure 5.6: Mean EDA responses to all BIP events (Putze et al., 2020, F7).
Figure 6.1: The triggers of User Engagement. The figure has adopted the cycle of engagement by O’Brien and Toms (2008) and highlights how the research threads presented in this thesis address the individual stages of engagement.

User Engagement (UE) has become a fundamental concern in HCI. Although many theories and guidelines on the design of engaging interaction exist, in practice, many training and intervention systems suffer from low participation and high drop-outs. This gap exemplifies that the effects of engaging design and playful/gameful design are not well understood. Therefore, designing interactive systems that are engaging is arguably a desirable goal in HCI. Yet, UE is a complex, multidimensional construct where the emphasis of different facets may diverge between contexts. This makes generalizable interaction design challenging, especially for interaction beyond the desktop, as the long-term effects of natural and multi-modal interaction are rarely explored. In line with Nacke and Drachen (2011), who suggested to pursue a multi-perspective approach to investigate engagement, the present dissertation addresses this fundamental gap of engaging design in HCI research and investigates how interactive systems should be designed to facilitate sustained user engagement. As a theoretical foundation of User Engagement (UE), this work applied O’Brien and Toms’ (2008) stages of engagement and synthesized it with perspectives on enjoyable interaction from research on PX (Landers et al., 2019; Mekler et al., 2014). This aggregated perspective utilizes UE both in terms of behavior (when do users engage?) and in terms of experience (what do users feel?) (Perski et al., 2017). To narrow down the broad research theme, this work follows Attfield et al.’s (2011) lines of research for UE (adoption of game design, immersive interaction design, and measurement methods), and subdivides it into three more specific research questions that distinctively address UE from the perspectives of game design (RQ1), haptic interaction (RQ2) and evaluation methods (RQ3). Figure 6.1 depicts an adopted version of O’Brien and Toms’ (2008) engagement model and summarizes how each line of research provides design opportunities that trigger engagement at the different stages. This multi-perspective view on interaction design aims to...
conceptualize the engaging qualities of interactive systems holistically to guide sustainable engagement with the interactive systems.

**Engaging Game Design (RQ1):** The contributions to game design show that gamefulness has the most versatile effect on UE affecting both the behavioral and experiential dimension of engagement at all stages of the engagement cycle. In response to Landers et al.’s (2019) call for the investigation on the effects of specific design elements, this thesis contributes the snacking framework (Alexandrovsky et al., 2019b, F1 and Alexandrovsky et al., 2021a, F2), which consists of five game dynamics that facilitate the snacking behavior – a regular but brief pattern of play. The framework allows to conceptualize specific gameful experiences designers might want to address. Referring to engagement as behavior, these game elements affect all stages of the engagement cycle and allow for a mapping of the behavioral attributes frequency, amount, duration, and depth (Perski et al., 2017) to prescriptive but generic game mechanics, which can be efficiently included in the design of interactive systems with a purpose.

For engagement as an experience, the snacking mechanics give designers tangible game elements that concern the affective outcomes a gameful system should produce and allow the design of experiences with varying intensity. The offered game design elements extend the existing perspectives of game elements as direct pull-in elements but utilize disengagement as a design dimension, which provides more subtle designs that address the players’ need of returning to the game (Waiting and Blocking). Regarding game design for mental health, this work is in line with Johnson et al. (2016) and Landers et al. (2019). The findings from the PUT (Alexandrovsky et al., 2020b, F3 and Volkmar et al., 2020, S3) studies extend Fleming et al.’s (2017) framework on game design for mental health through self-empowering, and creative elements, which offer a novel approach for playful mental health therapy in VR. Furthermore, PUT advocates a perspective shift from the gameful design that focuses on the state of flow towards meaningful goal-oriented and long-lasting gameful experiences with an emphasis on creativity as well as self-empowerment. These design strategies aim to facilitate engagement by evoking interest, attention, and emotion.

**Engaging Haptic Interaction (RQ2):** From the literature on presence and immersion (cf. Skarbez et al., 2017) in VR and engagement for health intervention (cf. Perski et al., 2017), this work identified immersive qualities of VEs in the form of haptic interaction as a strong factor that contributes to the intensity of ongoing experiences. As immersion affects the sense of realism (Schulz et al., 2019, F4), it brings the experience closer to the user. Also, it impacts the affective and cognitive components of engagement (Schulz et al., 2019, F4, Alexandrovsky et al., 2018, S1, and Fröhlich et al., 2018, F5), which in turn start influencing each other and modulate the experience. In terms of engagement as a subjective experience, the accumulated results support Kim and Schneider’s (2020) conceptualization of HXs and show that haptic interaction mainly addresses the intensity through the facilitated sense of realism, attention, interest and affect the point of engagement as well as the period of engagement. Moreover, the raised intensity of experience through HXs can build a foundation for physical activities to counteract sedentary behavior (e.g., exergames Oh and Yang, 2010; Smeddinck, 2016) (Mandryk et al., 2014). These findings help to improve the effectiveness of training simulations particularly for whole-body tasks in VR.

**Measurement Methods for Engagement (RQ3):** The meta-research on inVRQs investigated the design and administration of surveys in VR and advocates the embedding of self-reports into the VEs. The aggregated results show that embeddedness can minimize the disruption from the ongoing experience (cf. Putze et al., 2020, F7, Skarbez et al., 2017) and make a user study more enjoyable for the participants. The investigated interaction design lays the groundwork for future standardized assessment methods in VR user studies (Alexandrovsky et al., 2020a, F6). This line of research extends existing work on the switch between realities (cf. George et al., 2020; Knibbe et al., 2018) and adds the measurement of the experience into consideration of engaging design targeting the disengagement and re-engagement stages of the UE cycle.
Table 6.1: Classification of the contribution according to the dimensions and components of engagement. The table illustrates on which components of engagement the presented studies operate. The rows represent the stages of UE according to O’Brien and Toms (2008). The columns represent the three components of engagement according to Attfield et al. (2011) and Doherty and Doherty (2019).

<table>
<thead>
<tr>
<th>Component</th>
<th>Cognitive</th>
<th>Affective</th>
<th>Behavioral</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Point of Engagement</strong></td>
<td>“VRBox” (F₅), “Math Tangibles” (S₁)</td>
<td>“Playful User-Generated Treatment” (F₃), “Physical Props in VR” (F₄), “VRBox” (F₅)</td>
<td>“Casual Snacking” (F₁), “Serious Snacking” (F₂)</td>
</tr>
<tr>
<td><strong>Period of Engagement</strong></td>
<td>“VRBox” (F₅), “Math Tangibles” (S₁)</td>
<td>“Playful User-Generated Treatment” (F₃), “Physical Props in VR” (F₄)</td>
<td>“Casual Snacking” (F₁), “Serious Snacking” (F₂)</td>
</tr>
<tr>
<td><strong>Point of Disengagement</strong></td>
<td>“InVRQ Design” (F₆), “InVRQ BIP” (F₇)</td>
<td>“InVRQ Design” (F₆), “InVRQ BIP” (F₇)</td>
<td>“Casual Snacking” (F₁), “Serious Snacking” (F₂)</td>
</tr>
<tr>
<td><strong>Re-engagement</strong></td>
<td>“InVRQ Design” (F₆), “InVRQ BIP” (F₇)</td>
<td>“InVRQ Design” (F₆), “InVRQ BIP” (F₇)</td>
<td>“Casual Snacking” (F₁), “Serious Snacking” (F₂)</td>
</tr>
</tbody>
</table>

Towards Engaging Interactive Systems

Aggregating the presented contributions under the lens of how to design for sustained engagement, this work provides a multi-perspective view on engaging interactive systems incorporating emotional, cognitive, and behavioral elements of interaction design. The presented studies support Attfield et al.’s proposition that the three lines of research (interaction design, game design, and measurement methods) shape all aspects of engagement with interactive systems. As shown in Table 6.1, the studies in each thread contribute to the full cycle of engagement and address all major components of UE (behavioral, affective, and cognitive). The table highlights how individual design patterns affect different components of the experience and provide guidance for the design of specific experiences. This table is by far not complete and covers only aspects that were investigated within the scope of this thesis. Certainly each of the research threads, as well as other design elements and patterns not covered in this work, may affect other components and stages of engagement. However, Table 6.1 provides insights on what aspects of UE would be most likely governed by the investigated design dimensions and serves as a starting point for the design of engaging interactive systems.

The cognitive component is mainly addressed by haptic interaction and the embedded questionnaires. The VRBox (Fröhlich et al., 2018, F₅) and algebra tiles (Alexandrovsky et al., 2018, S₁) studies highlight that HXs support the users in exploration and creativity. They both serve as a trigger of engagement (point of engagement) and through the intensified experience with the application during the period of engagement. This extends the contemporary models of engagement (cf. O’Brien, 2016) with creativity and opens a design space for novel and engaging interaction design. The inVRQs (Alexandrovsky et al., 2020a, F₆ and Putze et al., 2020, F₇) as a novel assessment method for subjective self-reports in VR user studies affect the users’ involvement and provide designers control over disengagement as well as the re-engagement stages.

The affective component of engagement is triggered by all three research threads (game design, interaction design, and measurement methods). The PUT gameful design (Alexandrovsky et al., 2020b, F₃) as well as the interaction with the VRBox (Fröhlich et al., 2018, F₅) raise the interest and enjoyment
Discussion and Conclusion

through interaction that supports self-empowerment and creativity. Furthermore, in line with existing work on haptic and multimodal interaction (cf. Jacob et al., 2008; Kim and Schneider, 2020) the study on climbing in VR (Schulz et al., 2019, F4) provide supportive evidence for a link between interaction modality and the emotional response. The studies on inVRQs (Alexandrovsky et al., 2020a, F6 and Putze et al., 2020, F7) confirm previous results from the literature that BIPs negatively affect the experience (cf. Knibbe et al., 2018; Slater et al., 2003) and show that minimizing the disruption of an ongoing experience facilitates the engagement.

On the behavior dimension, both snacking studies (Alexandrovsky et al., 2019b, F1 and Alexandrovsky et al., 2021a, F2) showed that game design could affect player adherence independent of the player’s traits or incoming motivation. This allows to suggest that at least in parts, game design can modulate the player behavior independent of the player characteristics or the motivational attributes of the game mechanics themselves. However, the different results across the two studies highlight that the effect of individual game elements is dependent on the context or the genre of the game. These results suggest a perspective shift from gameful design as a motivational strategy for behavioral change towards game elements that facilitate a player behavior themselves.

It is important to note that design with the aim of altering the user’s behavior or that tries to maximize the enjoyment is valuable but also dangerous. When applied in training or therapy settings, these methods can promote engagement with content one might find difficult to adhere and improve the users’ well-being or the development of skills. However, these methods also provide a foundation for dark patterns that unethically manipulate the users forcing them into unintended behavior (Brignull, 2019; Zagal et al., 2013). For example engaging users can be exploited for monetization strategies (Altmeyer et al., 2019) or to force players into pathological behavior such as addiction or gaming disorder (WHO, 2018). These unethical designs have often dramatic consequences for the users. Therefore, in line with research on ethical UX design (Gray et al., 2018), designers are required to take social responsibility and reflect critically on the ethical implications of engaging design or designs that aim to alter the users’ behavior.

Proceeding from models on how users interact with technology (i.e., the engagement cycle), this work provides multiple target points – a toolbox – for researchers designers to counteract low participation and drop-offs. These contributions (i) provide guidance emotional and behavior-driven game design, (ii) layout a technical and conceptual grounding of interactive haptic experiences that can be implemented into both utilitarian and playful applications, and (iii) move the research agenda forward towards engaging standardized measurement methods. This work advocates the perspective shift in HCI towards a multi-angled view on potential points of adherence that will facilitate sustained engagement with interactive systems and, therefore, foster the design of truly useful interactive systems.
Limitations and Future Work

Arguably, to facilitate sustained long-term engagement, interactive systems must be designed from multiple perspectives, including the system design, the considerations of the context, and individual differences, and the design process should include different stakeholders (Cheek et al., 2015; Fleming et al., 2016; Landers et al., 2019; Thompson et al., 2010). However, this work only focused on the system design itself, excluding external factors of engagement, such as external goals or extrinsic motivation. Likewise, this work did not concern social and collaborative experiences. While acknowledging that social and multi-user scenarios are relevant for UE (cf. the need for relatedness Deci and Ryan, 2000), they are beyond the scope of this thesis. Moreover, it has been suggested that relatedness has a weak connection with enjoyment (Hassenzahl, 2008; Mekler et al., 2014). This work deliberately excluded social or multi-user setups as their design can be seen orthogonal or as an extension to single-user applications and the UE design should stand single-user scenarios first. While this research focuses only on system design and excludes external factors, the proposed design elements offer anchoring points for designers to connect the design with the particular instance of the interactive system. For example, a health game would provide meaningful achievements connected with real-life goals. The Blocking and Waiting mechanics could be adjusted to the optimal training duration, or TUIs in a math-learning application may be customized for individual preferences.

Regarding RBI, this work only investigated haptic interaction with passive props since this is the most common and straightforward interaction modality. However, to further facilitate immersion, future research should consider other senses and interaction modalities as their effects on UE may differ from haptics. Moreover, future research should look into the growing field of active TUIs such as robotic shape displays (McNeely, 1993), robotic arms (Strasnick et al., 2018) and other interfaces that “trick” the users’ perception by simulating haptic forces such as weight (Zenner and Kruger, 2017) air resistance (Heo et al., 2018; Zenner and Kruger, 2019), or stiffness (Strasnick et al., 2018). These active physical interfaces extend the expressive power of passive TUIs by using actuators to dynamically readjust the VE and allow for a bidirectional communication (He et al., 2017) between the environment and the users. These actuated physical props overcome some tradeoffs Jacob et al. (2008) proposed in the RBI framework.

This body of work looked only at the individual threads of UE. However, these dimensions are not entirely disjunctive (Attfield et al., 2011) and interfere with each other. For instance, if a VR game itself is not enjoyable, letting the players rate the game using inVRQs would most likely not fix those issues. Or if the game is well-designed, but the interaction is burdening, players might not enjoy the game since they may perceive the experience as a whole rather than a set of individually exchangeable components. However, it was necessary to identify and characterize the determinant factors of UE and how they individually modulate the experience first before investigating their interactions. Therefore, future work should look at how the individual components of UE interact and investigate their conjunction in more detail. Regarding the game design, so far, this work only identified the snacking mechanics as considerable design elements to facilitate a regular play behavior. However, to operationalize the mechanics for a determined behavior, further investigation is required to parameterize the mechanics and articulate a language of intended behaviors. Also, regarding the game design for VRET, so far, the PUT design has only been tested in a laboratory setting with convenient subjects. Although experts are positive towards the approach and related work suggests that results from convenient subjects are transferable (Robillard et al., 2003), a validation study with phobic participants is still ahead.
The present research does not offer a complete cross-product of the explored design dimensions and there are some missing links of the investigated design patterns. The snacking framework was only validated on a PC and a mobile game, but not in VR. Considering deviating results between the two investigated domains, in VR, the results might differ as well. However, both snacking studies demonstrated that most of the game mechanics could facilitate adherence. Also, since VR is deeply rooted in game design, it is reasonable to assume that the snacking framework is well transferable to immersive environments. Therefore, while expecting altering engagement patterns in VR, it is reasonable to suggest that the general effect of the snacking mechanics will persist. Similarly, so far, the VRBox was only developed as an interaction device without a connection to a specific scenario. For example, future research should evaluate the VRBox or similar devices in conjunction with PUT in a therapy context to provide a more accessible and expressive interaction. Moreover, physical shaping interfaces in VR such as the VRBox – as an interface that is easy to learn and allows for rapid prototyping – offer great potential for creative construction tasks. For instance, architects could use such a device to prototype buildings or landscapes together with their clients. Also, one might imagine such a device for the exploration of meteorological phenomena by connecting the terrain to a weather simulation. Users could terraform a landscape and explore how their changes might affect precipitation or wind energy. Concerning the findings from the inVRQs studies, this work suggests that the degree of embeddedness or disruption is a valuable design pattern for many (secondary) tasks. For instance, inVRQs could be used as quantify-self methods for VRET minimizing the chance that the patients disengage in a therapy session. However, future work should delve deeper into the connection between the inVRQs and the environment by looking at game design patterns associated with filling out the questionnaires (e.g., rewards for filling out the surveys Keusch and Zhang, 2017) and investigate a deeper contextualization of the surveys within the game environment (cf. Frommel et al., 2017).

So far, only the two studies on the snacking pattern have been looking at player behavior. For the greater proportion, this body of work investigated engagement only in single-session short-term studies under experimental conditions. Thus, it is not known how the presented findings affect users over time. Therefore, future research should put effort into the investigation of the long-term effects of the proposed design patterns. Future work also requires investigating engaging design in field studies contextualized in the users’ natural environments. Moreover, research should not only focus on regular engagement behavior but also look at how the designs could facilitate different behavioral patterns.

This body of work did not apply the User Engagement Scale (UES) (O’Brien and Toms, 2013), which produces an engagement score that could be compared across studies. Therefore, summative statements concerning O’Brien and Toms’ model of engagement cannot be made. However, this work employed measurements of related (sub-)constructs (e.g., usability, motivation, presence), which provide more nuanced details on the investigated variables. Moreover, the UES was developed in the context of an information retrieval system and is not widely applied in other sub-field of contemporary HCI research. Furthermore, to the best of our knowledge, this scale has not been applied in VR. Thus, the outcomes of this scale are not well-known. Therefore, in the context of the ongoing discussion regarding measurement methods of UX in VEs, future work should consider the UES (or its modifications) as a possible candidate for the evaluation of experiences in VR.

The proposed classification of the design patterns in Table 6.1 illustrates the designs investigated in this thesis. However, is by far not complete. For instance, the effects of interaction modality on the users’ behavior have not been concerned in this body of work. For example, Gerling et al. (2011) showed that familiarity with the input device affects the players’ performance. Likewise, Cole et al.’s (2015) definition of emotional and functional challenges fall into affective and cognitive components of engagement. This raises the open question on how does the interaction modality affect the user behavior within the application. However, the investigated design patterns and their revealed connection to the stages as well as components of UE offers a great potential to improve the engagement with interactive systems and may facilitate the effectiveness of training application and interactive systems for intervention. Moreover, such an investigation of how individual design elements affect particular components of UE is rarely discussed in the literature. Therefore, this thesis promotes a perspective shift in HCI towards design elements that address individual components of UE and open a space of interaction design that lays a foundation for a future unifying theory of user engagement.


Bibliography


Bibliography

and Gamification for Mental Health: Current Status and Promising Directions”. In: Frontiers in Psychiatry 7. ISSN: 1664-0640. DOI: 10.3389/fpsyt.2016.00215.


Gerling, Kathrin Maria, Frank Paul Schulte, Jan Smeddinck, and Maic Masuch (2012). “Game Design for Older Adults: Effects of Age-Related Changes on Structural Elements of Digital Games”. In:


Bibliography


Klimmt, Christoph, Tilo Hartmann, and Andreas Frey (Dec. 2007). “Effectance and Control as Determinants of Video Game Enjoyment”. In: CyberPsychology & Behavior 10.6, pp. 845–848. ISSN: 1094-9313, 1557-8364. DOI: 10/dr5x5.


Bibliography


Meehan, Michael (2001). “Physiological Reaction as an Objective Measure of Presence in Virtual Environments”.


Mojang and Markus „Notch“ Persson (May 17, 2009). “Minecraft”. In: Mojang Studios, Stockholm, Schweden. Last played Dez 2019,


Bibliography


Slater, Mel, Christoph Guger, Gueuter Eddinger, Robert Leeb, Gert Pfurtscheller, Angus Antley, Maia Garau, Andrea Brogni, and Doron Friedman (Oct. 2006). “Analysis of Physiological Responses to a Social Situation in an Immersive Virtual Environment”. In: *Presence: Teleoperators and Virtual Environments* 15.5, pp. 553–569. ISSN: 1054-7460, 1531-3263. DOI: 10.1162/pres.15.5.553.


Bibliography


Vorderer, Peter, Christoph Klimmt, and Ute Ritterfeld (Nov. 2004). “Enjoyment: At the Heart of Media Entertainment”. In: Communication Theory 14.4, pp. 388–408. ISSN: 1050-3292, 1468-2885. DOI: 10/d9v5fm.


Yee, Nick (Dec. 2006). “Motivations for Play in Online Games”. In: CyberPsychology & Behavior 9.6, pp. 772–775. ISSN: 1094-9313, 1557-8364. DOI: 10/cm9hfc.


### Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
<td>22</td>
</tr>
<tr>
<td>ANCOVA</td>
<td>Analysis of Covariance</td>
<td>25</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
<td>24</td>
</tr>
<tr>
<td>AR</td>
<td>Augmented Reality</td>
<td>2</td>
</tr>
<tr>
<td>AT</td>
<td>Anxiety Thermometer</td>
<td>43</td>
</tr>
<tr>
<td>AQ</td>
<td>Acrophobia Questionnaire</td>
<td>37</td>
</tr>
<tr>
<td>BIP</td>
<td>Break In Presence</td>
<td>9</td>
</tr>
<tr>
<td>BMBF</td>
<td>Federal Ministry of Education and Research</td>
<td>44</td>
</tr>
<tr>
<td>BVP</td>
<td>Blood Volume Pressure</td>
<td>54</td>
</tr>
<tr>
<td>DDA</td>
<td>Dynamic Difficulty Adjustment</td>
<td>13</td>
</tr>
<tr>
<td>ECG</td>
<td>Electrocardiography</td>
<td>42</td>
</tr>
<tr>
<td>EDA</td>
<td>Electrodural Activity</td>
<td>54</td>
</tr>
<tr>
<td>ET</td>
<td>Exposure Therapy</td>
<td>33</td>
</tr>
<tr>
<td>FFM</td>
<td>Five Factor Model</td>
<td>13</td>
</tr>
<tr>
<td>GOC</td>
<td>Game of Cannons</td>
<td>22</td>
</tr>
<tr>
<td>GUR</td>
<td>Games User Research</td>
<td>2</td>
</tr>
<tr>
<td>HCI</td>
<td>Human-Computer Interaction</td>
<td>1</td>
</tr>
<tr>
<td>HMD</td>
<td>Head-mounted Display</td>
<td>42</td>
</tr>
<tr>
<td>HR</td>
<td>Heart Rate</td>
<td>19</td>
</tr>
<tr>
<td>HRV</td>
<td>Heart Rate Variability</td>
<td>42</td>
</tr>
<tr>
<td>HX</td>
<td>Haptic Experience</td>
<td>16</td>
</tr>
<tr>
<td>IMI</td>
<td>Intrinsic Motivation Inventory</td>
<td>23</td>
</tr>
<tr>
<td>inVRQ</td>
<td>in-VR questionnaire</td>
<td>ix</td>
</tr>
<tr>
<td>IPQ</td>
<td>Igroup Presence Questionnaire</td>
<td>43</td>
</tr>
<tr>
<td>MAL</td>
<td>Multimodal Algebra Learning</td>
<td>44</td>
</tr>
<tr>
<td>MANOVA</td>
<td>Multivariate Analysis of Variance</td>
<td>24</td>
</tr>
<tr>
<td>MF-ANOVA</td>
<td>Mixed-factorial ANOVA</td>
<td>37</td>
</tr>
<tr>
<td>MR</td>
<td>Mixed Reality</td>
<td>8</td>
</tr>
<tr>
<td>mRRI</td>
<td>mean R-R interval</td>
<td>42</td>
</tr>
<tr>
<td>MTurk</td>
<td>Amazon Mechanical Turk</td>
<td>23</td>
</tr>
<tr>
<td>outVRQ</td>
<td>out-VR questionnaire</td>
<td>47</td>
</tr>
<tr>
<td>PANAS</td>
<td>Positive and Negative Affect Scale</td>
<td>8</td>
</tr>
<tr>
<td>PENS</td>
<td>Player Experience of Need Satisfaction</td>
<td>23</td>
</tr>
<tr>
<td>Psi</td>
<td>Plausibility Illusion</td>
<td></td>
</tr>
<tr>
<td>PUT</td>
<td>Playful User-generated Treatment</td>
<td>21</td>
</tr>
<tr>
<td>PX</td>
<td>Player Experience</td>
<td>7</td>
</tr>
<tr>
<td>RM-ANOVA</td>
<td>ANOVA with repeated measures</td>
<td>43</td>
</tr>
<tr>
<td>RBI</td>
<td>Reality-based Interaction</td>
<td>15</td>
</tr>
<tr>
<td>RPE</td>
<td>Rating of Perceived Exertion</td>
<td>42</td>
</tr>
<tr>
<td>RR</td>
<td>Respiratory Rate</td>
<td>42</td>
</tr>
<tr>
<td>SC</td>
<td>Skin Conductance</td>
<td>54</td>
</tr>
<tr>
<td>SCR</td>
<td>Skin Conductivity Response</td>
<td>42</td>
</tr>
<tr>
<td>SDT</td>
<td>Self-Determination Theory</td>
<td>8</td>
</tr>
<tr>
<td>SIMS</td>
<td>Situational Motivation Scale</td>
<td>23</td>
</tr>
<tr>
<td>STAI</td>
<td>State-Trait-Anxiety-Inventory</td>
<td>36</td>
</tr>
<tr>
<td>SUDS</td>
<td>Subjective Units of Distress-Scale</td>
<td>36</td>
</tr>
<tr>
<td>SUS</td>
<td>System Usability Scale</td>
<td>47</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Form</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>TIPI</td>
<td>Ten Item Personality Inventory</td>
<td>23</td>
</tr>
<tr>
<td>TLX</td>
<td>NASA Task Load Index</td>
<td>50</td>
</tr>
<tr>
<td>TUI</td>
<td>Tangible User Interface</td>
<td>15</td>
</tr>
<tr>
<td>UE</td>
<td>User Engagement</td>
<td>1</td>
</tr>
<tr>
<td>UES</td>
<td>User Engagement Scale</td>
<td>17</td>
</tr>
<tr>
<td>UI</td>
<td>User Interface</td>
<td>36</td>
</tr>
<tr>
<td>UMUX</td>
<td>Usability Metric for User Experience</td>
<td>45</td>
</tr>
<tr>
<td>UX</td>
<td>User Experience</td>
<td>3</td>
</tr>
<tr>
<td>UXQ</td>
<td>User Experience Questionnaire</td>
<td>7</td>
</tr>
<tr>
<td>VAS</td>
<td>Visual Analog Scale</td>
<td>42</td>
</tr>
<tr>
<td>VE</td>
<td>Virtual Environment</td>
<td>1</td>
</tr>
<tr>
<td>VR</td>
<td>Virtual Reality</td>
<td>i</td>
</tr>
<tr>
<td>VRET</td>
<td>Virtual Reality Exposure Therapy</td>
<td>3</td>
</tr>
</tbody>
</table>
Fundamental Publications

This dissertation is founded on seven fundamental publications which encompass several contributions of research methods, artifacts and game design theory on research around UE. In the following, the publications which solidify this dissertations are briefly summarized and the personal contribution according to the CRediT taxonomy\(^1\) are stated.

**F\(_1\) Game Dynamics That Support Snacking, Not Feasting**


**Contribution with regard to this dissertation:** This work investigates how different game elements affect the user engagement over time. In response to RQ1, the publication contributes a framework of five game mechanics that support the design of intended behavior.

**Personal contribution:** Conceptualization, data curation, formal analysis, investigation, project administration, resources, software, supervision, validation, visualization and contribution to all parts of the manuscript (80%).

**F\(_2\) Serious Snacking: A Survival Analysis of How Snacking Mechanics Affect Attrition in a Mobile Serious Game**


**Contribution with regard to this dissertation:** This publication builds on previous work from F\(_1\) and applies the snacking framework in the context of serious games. With respect to RQ1, this work extends the previous findings and demonstrates how casual game design can be transferred into a serious context in order to facilitate adherence.

**Personal contribution:** Conceptualization, data curation, formal analysis, funding acquisition, investigation, project administration, resources, software, supervision, validation, visualization and contribution to all parts of the manuscript (80%).

\(^1\)https://casrai.org/credit/
F3 Playful User-Generated Treatment: A Novel Game Design Approach for VR Exposure Therapy


This work received an Honorable Mention Award at CHI PLAY 2020.

Contribution with regard to this dissertation: This publication investigates rarely explored game design elements for emotionally challenging tasks in the context of a VRET game for acrophobia. With respect to RQ1, this work contributes a playful two-step approach, which strengthens the personal connection and self-empowerment through creativity.

Personal contribution: Conceptualization, data curation, formal analysis, funding acquisition, investigation, project administration, resources, software, supervision, validation, visualization and contribution to all parts of the manuscript (70%).

F4 The Role of Physical Props in VR Climbing Environments


Contribution with regard to this dissertation: This publication investigates the effects of haptic experience on the users’ anxiety sense of presence by comparing real-world climbing, full-body climbing in VR and controller-based (virtual) climbing in VR. The findings respond to RQ2 and highlight the importance of haptic sensation on the user experience. Further the results indicate passive haptic props as a warranted approach to improve UE.

Personal contribution: Conceptualization, data curation, formal analysis, project administration, resources, supervision, validation, visualization and contribution to all parts of the manuscript (60%).

F5 VRBox: A Virtual Reality Augmented Sandbox for Immersive Playfulness, Creativity and Exploration


This work received an Honorable Mention Award at CHI PLAY 2018.
Contribution with regard to this dissertation: This publication explores the hybrid interaction with passive shape-changing materials (i.e., sand) in VR and contributes an artifact which demonstrates the feasibility and the design space of the approach. With respect to RQ2, this work shows how reality-based haptic interaction can improve the playfulness and creative potential of an interactive system.

Personal contribution: Conceptualization, data curation, formal analysis, project administration, resources, software, supervision, validation, visualization and contribution to all parts of the manuscript (45%).

F6 Examining Design Choices of Questionnaires in VR User Studies


Contribution with regard to this dissertation: This work investigates the interaction design and effects of conducting questionnaires in VR. This work establishes a usable design for inVRQ and demonstrates the applicability of this less-disturbing measurement method. These findings provide the underlying design for the F7 study. This publication contributes to the methodology and the interaction design dimensions of the dissertation and responds to RQ3.

Personal contribution: Conceptualization, data curation, formal analysis, investigation, project administration, resources, software, supervision, validation, visualization and contribution to all parts of the manuscript (50%).

F7 Breaking The Experience: Effects of Questionnaires in VR User Studies


Contribution with regard to this dissertation: Using the in-VR questionnaire design from F6, this study investigates how the switch in realities in order to fill out questionnaires affects the BIP. The results show that filling out questionnaires in VR leads to a smaller BIP and this is less disengaging. These findings provide evidence towards using inVRQs and identify the disruptiveness of a task as a factor of disengagement. This publication contributes to the methodology dimension of the dissertation and responds to RQ3.

Personal contribution: Conceptualization, data curation, formal analysis, investigation, project administration, resources, software, supervision, validation, visualization and contribution to all parts of the manuscript (40%).
Supportive Publications

These publications are non-archival work which supplements the fundamental publications.

S₁ Exploring Interactive Systems for Algebra Learning in School


Contribution with regard to this dissertation: This work investigates how passive haptic physical props affect the engagement with an AR math learning application. This contribution follows RQ2, and provides additional evidence that passive physical props can facilitate a stronger engagement over touch interfaces.

Personal contribution: Conceptualization, data curation, formal analysis, project administration, resources, software, supervision, validation, visualization and contribution to all parts of the manuscript (50%).

S₂ Demonstrating VRBox: A Virtual Reality Augmented Sandbox


Contribution with regard to this dissertation: This technical contribution improves the apparatus presented in F₃ and demonstrates the suitability of the VRBox for public venues.

Personal contribution: Conceptualization, project administration, resources, software, validation, visualization and contribution to all parts of the manuscript (80%).

S₃ Playful User-Generated Treatment: Expert Perspectives on Opportunities and Design Challenges

Supportive Publications

**Contribution with regard to this dissertation:** This publication extends the results of the online survey in publication F₃ and provides additional insights from therapists which refine the PUT design approach.

**Personal contribution:** Conceptualization, validation and writing (20%).

**S₄ Evaluating User Experiences in Mixed Reality**


**Contribution with regard to this dissertation:** This CHI ’21 workshop disseminates the accumulated results from F₆ and F₇ to the community and facilitates the discussion on the improvement of evaluation methods in MR.

**Personal contribution:** Conceptualization, project administration, contribution to all parts of the manuscript (45%).
Game Dynamics that Support Snacking, not Feasting

Dmitry Alexandrovsky  
University of Bremen  
dimi@uni-bremen.de

Maximilian Achim Friehs  
University of Trier  
friehs@uni-trier.de

Max V. Birk  
TU/e Eindhoven  
m.v.birk@tue.nl

Rowan K. Yates  
University of Saskatchewan  
rowan.k.yates@gmail.com

Regan L. Mandryk  
University of Saskatchewan  
regan@cs.usask.ca

ABSTRACT

Updated 08.10.2019—Player experience research tends to focus on immersive games that draw us into a single play session for hours; however, for casual games played on mobile devices, a pattern of brief daily interaction—called snacking—may be most profitable for companies and most enjoyable for players. To inform the design of snacking games, we conducted a content analysis of game mechanics in successful commercial casual games known to foster this pattern. We identified five single-player game dynamics: Instant Rewards, Novelty, Mission Completion, Waiting, and Blocking. After situating them in theories of motivation, we developed a game in which game mechanics that foster each dynamic can be included individually, and conducted two studies to establish their relative efficacy in fostering the behavioural pattern of snacking, finding significant potential in Novelty and Waiting. Our work informs the design of games in which regular and brief interaction is desired.

INTRODUCTION

Various frameworks of game engagement can guide game design, while motivational theories are used to inform player experience (pX) evaluation; however, pX research is currently focused on creating immersive flow experiences that draw the player in and keep them there for hours, e.g., [119, 21, 69]. For example, flow theory [25] is defined in part by absorption and loss of a sense of time passing [39], whereas in self-determination theory [105], engagement is often operationalized as time spent in free play, e.g., [14].

Many games are designed for absorbed immersion; however, mobile casual games are often intended to be played frequently in brief play sessions [65]—a pattern of interaction called snacking [93], in which users graze regularly (e.g., daily) but for short bursts of time (e.g., 10 minutes per play session) [98, 125]. Snacking helps players resist game-related fatigue or apathy through engaging short sessions and by promoting frequent and repeated interaction [93]—a pattern that may help maximize both profits for companies and enjoyment for mobile casual game players [67, 64].

Although game mechanics that promote snacking exist, there is little guidance on how to prescriptively apply them in design. Moreover, there is no evidence on which mechanics are most effective at supporting snacking patterns. In this paper, we identify, implement, and evaluate game mechanics that promote snacking. We first examine previous work that has formalized game design and experience in frameworks and theories. We then report on a systematic content analysis of 30 commercially successful single-player casual mobile games, from which we identified the mechanics that promote snacking. We then present these mechanics in a framework of five game dynamics—i.e., game experiences that result from specific game mechanics [55] and situate them into theories of motivated behaviour that explain why they are effective. We report on the design of a casual game, in which we individually layered game mechanics to promote experiences from the five dynamics. We then report on two studies that evaluated the dynamics in our game: a preliminary study to assess equivalent pX of the five dynamics and a 9-day behavioural experiment in which we released six versions of the game (baseline game plus a version for each dynamic) and monitored objective play patterns and subjective experience. Our work suggests that including game mechanics that leverage novelty and that require waiting may be most effective at facilitating the snacking pattern of play. By providing a cohesive examination of game mechanics through a content analysis, situated within theories that explain their motivating influence, and validated by an empirical investigation of play behaviour in-the-wild, our work can help researchers and developers understand how to design more effectively for snacking-style game interactions.

BACKGROUND

Literature on game experience often attempts to grasp and formalize the features of games that engage players (c.f., [130, 107]). Games are frequently described as complex interactive systems, with strong interplay of boundaries, technology, and player emotions (c.f., [24, 107]). Boyle et. al [20] provide an overview of player experience (pX) research and engagement in games. Csíkszentmihályi’s Flow theory has been acknowledged as a key element in pX [89], and satisfaction of needs theories (e.g. Self-determination theory [27] or uses and grati-
fication [102]) describe the motives of playing, yet Boyle et al. [20] ultimately conclude that the understanding of what makes games enjoyable is “fundamentally incomplete”.

Malone [83] explored the motivational benefits of games for engaging people in learning, and identified challenge, fantasy, and curiosity as three key elements of engaging design. Mora et al. [92] summarized 18 gamification frameworks and derived 5 general design principles: Economic, Logic, Measurement, Psychology, and Interaction to guide gamification design. For instructional games, Garris et al. [44] assembled a framework with six game dimensions: Fantasy, Rules/Goals, Sensory Stimuli, Challenge, Mastery, and Control. On a lower level, Robinson and Belotti [101] established a taxonomy of six gamification elements, i.e., General Framing, General Rules and Performance Framing, Social Features, Incentives, Resources and Constraints, and Feedback and Status Information, as general rules. For casual game design, Juul [65] identified the five elements of fiction, usability, interruptibility, difficulty and feedback (juiciness). These elements give a terminology to describe casual games as such, but miss to guide casual game design for specific patterns of play. Fields characterizes casual games by their immediacy of playing and stopping [36]. These models provide value; however, they are often abstract and not easy to apply prescriptively to game development [101] to inform the choice of specific game mechanics that foster particular play experiences or behaviours.

A framework that was intended to guide game design is Hu- nicke et al.’s MDA (Mechanics-Dynamics-Aesthetics) framework [55]. According to MDA, mechanics are the specific rules, actions and behaviours a player is afforded and how the system responds (e.g., weapon upgrade option). Dynamics result from execution of the mechanics and the interactions between them and the game system; for example, a weapon upgrade (mechanic) might result in the experience of novelty (dynamic), or a cooldown timer (mechanic) might result in the experience of waiting (dynamic). Finally, the aesthetics are the emotional experiences evoked in the player. The framework can be applied in two directions: Game designers can decide on the designs (aesthetics) they wish to evoke in players, and then design play patterns (dynamics) and rules (mechanics) to achieve their intended design. From the other perspective, designers can critically assess how different mechanics yield various experiences and emotions during play.

Motivational theories (e.g., self-determination theory [105], flow theory [25]) also explain pX; however, numerous theories can explain a single behaviour, and theories are not often systematically validated in-the-wild. There are exceptions; in a series of studies, Birk et al. investigated how one type of interaction design (i.e., customizing avatars) explained by one part (i.e., autonomy) of a large theory (i.e., self-determination theory [27]) affected motivation [14], motivated behaviour [14, 16], attrition [18, 16], and ultimately efficacy [17] of an intervention. Although motivational design can help guide intended play experiences, it is difficult for designers to apply motivational theories prescriptively to inform choices on specific game mechanics that facilitate intended patterns of behaviour. Although research identifies mechanics that foster enjoyment, how mechanics contribute to play behaviour in general and to the snacking pattern in particular is not well understood.

FRAMEWORK OF MOTIVATION MECHANICS

To inform the design of games that foster snacking, we can identify game mechanics in casual mobile games that successfully promote snacking. We analyzed common mechanics in popular and successful off-the-shelf titles for mobile devices that promote the snacking pattern.

Content Analysis Methods

We began by searching online for the most popular and top-ranked mobile games, and filtered for games that fit into the snacking pattern. The criteria for screening were: 1) high ratings, 2) reviews spoke of good retention, 3) lower ratings for other reasons, but were liked by the public. After gathering this initial list of games (N=30), we kept 22 games from 5 genres (action, collection, idle, management, puzzle and strategy) as the other 8 were clones of the first 22 or did not promote snacking. We did not continue to seek more games to include, as we quickly found that analyzing new games was not adding to our knowledge. The content analysis was conducted from May 2017 to July 2017, and reflected the most popular mobile casual games in that time frame.

We deconstructed the games into their core mechanics and genres by through play testing, watching play videos online, and analyzing critic and user reviews. We use the MDA framework [55] as guidance. The game mechanics we identified were then grouped iteratively and with discussion into five categories of experiences—i.e., dynamics—that are yielded by those mechanics. For example, a cooldown timer (mechanic) resulted in the experience of waiting (dynamic).

Social interaction is known to motivate sustained engagement [27]. Ten of the games we analyzed contained multi-player components; however, the motivational pull that results from the social aspects of games is independent and orthogonal to the single-player mechanics we aimed to identify. For example, the experience of waiting for a cooldown timer to expire is not specific to a single- or multi-player game. The motivational pull from social interactions is known, complex, and does not interfere with the single-player mechanics employed, thus it is beyond the scope of this investigation, and should be explored explicitly and systematically in future work. Here, we focus on mechanics and dynamics that can be employed in both single- and multi-player contexts.

Content Analysis Results

We first describe the five identified dynamics. We then present various psychological theories of motivation that can be used to explain aspects of their success, followed by related literature on games that have used them. A summary of the content analysis is depicted in Table A in the appendix.

Instant Rewards (Reward)

Instant rewards are rewards or achievements given to the player immediately upon completion of the deserving game action; mechanics that promote the experience of instant rewards were found in 20 (91%) of the 22 games. For example, in Angry Birds, the player gets 1, 2 or 3 stars after completing
a level. Instant rewards are not always guaranteed, and can be either anticipated, i.e., the player is aware of the expected reward, or unanticipated.

**Theoretical Description of Rewards:** Instant rewards can be described by Skinner’s operant conditioning paradigms [108, 110], in which a desired behaviour that is rewarded, even by random chance [109], leads to a higher likelihood of that behaviour being repeated. Four factors influence the effectiveness of reinforcement or punishment: 1) satiation and deprivation, 2) immediacy, 3) contingency, and 4) size and amount. Initially, a continuous reward showing each of the desired behaviours leads to fast learning, but also to a fast deletion of the behaviour when the rewarding is ceased [94]. Thus, to establish long-term behaviour and shield against deletion, rewards should be given quickly and continuously in the early stages of the learning phase and intermittently in later stages, when baseline behaviour patterns have been established [54, 94]. In the context of snacking-style interaction, this means that early on, the players should be fed with new features at a high-frequency, and with increased playtime the reward frequency should decay.

**Instant Rewards in Games Research:** The immediacy of reward in digital games is a key feature that is known to draw players into computer games [73]. There have been several explorations into the efficacy of rewards [88]. Instant rewards have also been coupled with learning objectives to provide clear goals for the player [56]. With an instantly rewarding system, there is a constant feedback loop between the learning goals and player’s performance [112]. Instant rewards are often designed as a form of performance currency system (e.g., score points) [76]. While performance feedback and typical in-game rewards such as power-ups are not necessarily equivalent, they can both perceived as rewarding and enhance motivation [28] and goal completion [78].

**Mission Completion (Completion)**

When players complete missions—i.e., explicitly assigned tasks, quests, or goals—they are gratified with in-game rewards, as well as a sense of achievement, associating a positive feeling with the game; mission-based mechanics were found in 14 (63%) of the 22 games. Missions were identified as belonging to one of three categories: Daily Login, One-off, and Chains. The Daily Login missions served as the main conditioner, because all a player had to do was open the application to complete the mission. One-off missions were used to engage the player and feed a potential or existing habit (e.g., complete the level with one shot). Chains of missions were used similarly, but were not necessarily completed in one play session; with their greater time investment, they usually yielded a larger reward. Failing a mission sometimes does not affect the player, but in other cases they are punished and lose some of their accomplishments.

**Theoretical Description of Completion:** Mission completion can fulfill players’ needs for competence, autonomy, and relatedness [105, 106]. Having choice over missions promotes autonomy and once a mission is completed, a player is generally rewarded for succeeding (competence). Depending on individual differences, players might be motivated by different needs [15], or more by obtaining rewards versus avoiding losses (c.f., [53, 7]). Further, players with high motivation for success may seek completion of higher-difficulty tasks, while those avoiding failure may seek to complete tasks with lower difficulty (c.f., [66, 124, 50]). Furthermore, as the Zeigarnik Effect shows, unfinished tasks (such as uncompleted missions) have an innate pull [86, 120], and the satisfaction provided by in-game completion increases with age [15].

**Mission Completion in Games Research:** Missions can give a sense of the global picture in the game and provide clear goals [44]. Being part of the narrative, missions enhance the fantasy [83] and communicate the player’s progress and performance. In learning games, missions glue the game mechanics to the learning objectives and give players a purpose [87]. A recent review on health-related video games states that goal-setting (i.e., missions) are a prime method by which video games can change behaviour [8].

**Novelty**

Novelty is achieved by the introduction of new in-game elements, such as items, mechanics, assets, environments, or goals; mechanics that promote the experience of novelty were found in 20 (91%) of the 22 games. Novelty brings excitement to the game by providing the player with new elements that affect both gameplay and the underlying experience. Examples of novelty-promoting mechanics include: level-ups, unlocking new content, or activating new skills.

**Theoretical Description of Novelty:** As Lomas et al. identified, novelty is a contributing factor to difficulty [81]. It is important for player retention, although too much new content could lead to information overload and diminished competence [106]. Novelty itself can produce positive effects [80], and also is inversely related to negative outcomes. Novel stimuli draw attention (e.g., [34, 63, 26]) compared to repeated stimuli. Novelty prevents boredom and fosters pleasantness [13]. Stimulus factors such as repetitiveness, lack of novelty, and monotony generate boredom [111]. Furthermore, boredom has been linked to a lack of sustained attention [82] and leisure satisfaction [60]. But the effect of complex novel stimuli might vary between individuals [30, 10, 100, 33].

**Novelty in Games Research:** Unlocking new content is often designed as part of novelty. Hernandez et al. [52] deployed a 10-week trial of a networked multi-player exercise game for children with Cerebral Palsy to exercise and socialize together. To maintain interest over the 10-weeks, they included six mini-games, which were introduced progressively every two weeks. Behavioural indicators (i.e., games chosen and time played) suggest that the strategy was effective [51]. Mandryk et al. [85] deployed a 12-week trial of a neurofeedback training system that turned any off-the-shelf game into a biofeedback game to children with fetal alcohol spectrum disorder to help them learn to self-regulate. The authors reported that they originally deployed five game choices, but that participants complained of boredom after several weeks and thus new games were added to maintain interest in and enjoyment of training.
Blocking
We define blocking as a passively regenerating resource that is required to interact with the game; blocking mechanics were found in 5 (22%) of the 22 games. For example, in Last Day on Earth [3], the game world is split into different play zones on a map. When the avatar travels from one zone to another, most of the play activity is blocked. The idea behind blocking through regeneration mechanics is to limit the total time of a gameplay session and though many people complain about this limiter on a game, it can prevent overplay and burnout or loss of interest through explicit blocking [64].

Theoretical Description of Blocking: Blocking through regeneration can be viewed as a conditioning mechanic with a fixed reward scheme—it forces the player to stop playing to provide value to the next session. From this perspective, blocking through regeneration mechanics hinders satiation of the game in the context of conditioning theories. This is similar to Token Economies [62, 49], in which a secondary reinforcer called the Token (i.e., the resource needed to play the game) can be traded in for a primary reinforcer (i.e., playing the game). Blocking through regeneration mechanics limits the amount of exposure, and thus boredom is less likely to occur (see section on Novelty). However, regeneration mechanics may also interfere with satisfaction of player autonomy, which could negatively influence PX and motivation.

Blocking in Games Research: Blocking can be found in many contemporary free-to-play games [64]. The business-model restricts the playtime through a regenerative virtual currency that is required to continue. Players can either wait until their account has generated enough currency to continue or buy another few rounds. Often these games have in-game purchases for new content or faster progress [36].

Waiting
The experience of waiting results from mechanics that require time to pass, such as game resources that are instructed to produce or must be gathered over time; mechanics that promote the experience of waiting were found in 14 (63%) of the 22 games. Waiting introduces time as a resource that has to be managed and considered. In contrast to blocking, waiting allows to continue playing while the resources are regenerated. However, the players’ actions are restricted. For example, in a construction simulation game, the time and cost in order to build or update an entity might grow over time. Waiting mechanics can be applied actively or passively: If actively used, a timer progresses during active game play; whereas passive waiting mechanics progress independent of play. Passive waiting is primarily utilized in idle games, a genre that is characterized by passive and automated gameplay [37].

Theoretical Description of Waiting: Similar to blocking through regeneration, waiting mechanics can be used as an incentive to bring the player back to the game through operant conditioning [109] and delayed gratification methods [90]. In particular for idle games, Waiting addresses economy of attention, recurring gratification, compulsive gameplay and elimination of drudgery [37]. Waiting can be viewed similar to blocking, but operating on a smaller time scale. In comparison to blocking, waiting does not prevent players from engaging with the game, it just makes certain actions within the game less effective or not possible. Waiting mechanics condition a behavioural response to a trigger (e.g., a notification, showing “your resources are here”, will trigger the in-game response without much thought) and by that, establish habitual behaviour. Further, waiting mechanics can be used to keep players mentally engaged by the means of encouraging delayed gratification [90, 22]. Translated to a gaming environment, delayed gratification could result in an interest-system in which the player is rewarded for saving up their currency.

Waiting in Games Research: As with blocking, waiting is often applied within the business-model of free-to-play games [36, 72]. Waiting refuses the access to some game elements. This mechanic can be used to enforce the players to a particular style of playing or behaviour. For example, Barathi et al. implemented cooldowns to switch between sprints and paced phases in a high intensity interval training game [9].

Summary of Dynamics that Promote the Snacking Pattern
We identified various game mechanics that promote snacking and grouped them into five categories of game dynamics (Instant Rewards, Mission Completion, Novelty, Blocking, and Waiting), contextualized them in motivational theories, and grounded the game dynamics in prior use in game design research. The identified dynamics are by no means independent and could be combined (e.g., instant rewards or novel elements could be provided from completing missions). These five groups are also not intended as a comprehensive list, but meant as a starting point to identify game mechanics that yield dynamics that facilitate the snacking pattern.

DESIGNING A GAME TO COMPARE THE APPROACHES
To test how different approaches influence play patterns, we designed and developed a custom game that was playable in a basic version, but in which we could include each approach individually or in combination. The design had to fulfill requirements: 1) The gameplay had to appeal to a wide target group; 2) The core game had to be engaging enough that participants would play for several days, but not so engaging that they would want to play for hours at a time; 3) The core game mechanics had to work on their own and also with the additional snacking dynamics; 4) Each version had to emphasize only one of the additional dynamics, with the least possible interference between them; and 5) Each version had to be comparable in terms of short-term enjoyment and motivation.

Game Design
We created Game of Cannons (GOC)—a casual 2D physics puzzle with action elements inspired by games like Angry Birds [61], a popular casual mobile game. The player is equipped with a cannon and a building with minions in it. An avatar that allows for small customization (color, hats) is standing next to the cannon. On the other side of the scene, an artificial intelligence (AI) opponent with the same items faces the player; see Figure 1. There is also an inventory in which in-game items are stored and can be accessed.

The goal is to shoot the opponent’s minions, which live in their buildings, while protecting one’s own buildings. In contrast to Angry Birds, the game has targets and an AI opponent, which
adds action and is intended to promote autonomy and relatedness [106]. Moreover, it allows for a multi-player version for future research to explore social dynamics.

Each of the parties takes turns shooting at the opponent’s building, trying to eliminate all minions, after which the match is over. The game has 64 levels of increasing difficulty: the buildings become more difficult to destroy and the accuracy of the AI increases. To provide a reasonable difficulty curve, 44 levels were adopted from Angry Birds. Initially, only the first level is unlocked, and completing a level unlocks the next one. An additional wind mechanic that changes every turn in both strength and direction affects the projectile’s trajectory, adding randomness to the game. A mouse is used for game interaction: left click and drag controls the angle and the velocity of a player’s shot; angle is visually indicated by the inclination of the barrel and velocity by a loading bar above the cannon. The camera position can be adjusted by holding the right mouse button and dragging the mouse in the desired direction. Using the scroll wheel, players can zoom in and out. The zooming ranges from a full view of the entire scene to a detailed view that shows only a few meters around the cannon.

Game Versions
We implemented several game mechanics for each of the framework’s dynamics to design for snacking interactions.

Rewards: We realized instant rewards through score points for destroying minions and buildings as well as through rewarding players with power-ups. After completing a level successfully, players were rewarded with up to 3 stars depending on their number of turns and how many blocks they destroyed. We implemented three power-ups: (1) Telescope provides players with the ability to preview the trajectory of their next shot. (2) Explosion increases the destructive power of the next projectile used. (3) Airstrike calls an UFO that throws a bomb on the opponent’s building. At the beginning of each level power-ups were generated randomly with weighted probabilities to assure that weaker items appeared more frequently (Telescope: 1 out of 10, Explosion: 1 out of 25, Airstrike: 1 out of 1000). Power-ups were collected by shooting or clicking crates in the game.

Completion: We implemented mission completion through an achievement mechanism. Players were rewarded for in-game progression as well as playing daily. We implemented 28 achievements in three types: One-off (3/28, e.g., destroy the enemy with one shot), progressive (6/28, e.g., Daily Login), and chained (19/28, e.g., finish all levels in world). To make players aware of their progress, we notified them about available and earned achievements after they fulfilled the requirement for one of the 28 available achievements.

Novelty: We included several mechanics to introduce novel elements over time. First, players could choose from five projectiles (weapons) for the cannon. Each projectile could be accessed through the inventory. Weapons were tied to a particular level, and were introduced evenly throughout the levels as players progressed. Further, the 64 levels were split into four differently-themed worlds with 16 levels in each world. To visually differentiate between worlds, we altered the background of all levels for each world. Moreover, in 21 out of 64 levels and in every final world level, players earned hats that they could use to customize their avatar.

Blocking: We implemented blocking by restricting playtime to 30 minutes with a minimum of 8 hours between play sessions. Besides the playtime restriction and the waiting mechanism, blocking had no differences to the basic game.

Waiting: In the waiting condition, we used the same power-ups as in the Rewards condition—i.e., Telescope, Explosion, Airstrike—except that power-ups in the waiting condition were generated over time and not presented at the beginning of each round. To indicate arrival of each power-up, we presented three different count downs, one for each power-up. Players were restricted to only one of each power-up at the same time. Therefore, a power-up needed to be used before the same power-up could be collected again.

PRELIMINARY SHORT-TERM COMPARISON
We first evaluated the different versions in a single, short-term play session to determine how pX was affected.

Participants and Procedure
We recruited 99 participants using Amazon Mechanical Turk (MTurk)—an online platform that connects requesters with workers for Human Intelligence Tasks (HITs). Participants were paid $4 for 20-25 minutes of work. Ethical approval was obtained from the University of Saskatchewan.

Table 1. Pre-study: demographics, motivation, and group differences.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SE</th>
<th>F_{1.99}</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion (1-7)</td>
<td>3.3</td>
<td>0.18</td>
<td>0.82</td>
<td>0.316</td>
</tr>
<tr>
<td>Agreeableness (1-7)</td>
<td>5.6</td>
<td>0.11</td>
<td>1.96</td>
<td>0.433</td>
</tr>
<tr>
<td>Conscientiousness (1-7)</td>
<td>5.5</td>
<td>0.12</td>
<td>1.28</td>
<td>0.281</td>
</tr>
<tr>
<td>Emotional Stability (1-7)</td>
<td>5.0</td>
<td>0.16</td>
<td>2.44</td>
<td>0.055</td>
</tr>
<tr>
<td>Openness to Experiences (1-7)</td>
<td>5.3</td>
<td>0.14</td>
<td>0.39</td>
<td>0.816</td>
</tr>
<tr>
<td>Age</td>
<td>33.03</td>
<td>0.80</td>
<td>0.55</td>
<td>0.699</td>
</tr>
<tr>
<td>Identification as a Gamer (1-10)</td>
<td>6.9</td>
<td>0.28</td>
<td>0.99</td>
<td>0.413</td>
</tr>
<tr>
<td>Identified Regulation (1-7)</td>
<td>4.9</td>
<td>0.16</td>
<td>0.80</td>
<td>0.326</td>
</tr>
<tr>
<td>External Regulation (1-7)</td>
<td>4.3</td>
<td>0.14</td>
<td>0.64</td>
<td>0.634</td>
</tr>
<tr>
<td>Motivation (1-7)</td>
<td>2.0</td>
<td>0.12</td>
<td>0.94</td>
<td>0.44</td>
</tr>
<tr>
<td>Identification as a Gamer (1-10)</td>
<td>2.1</td>
<td>0.12</td>
<td>1.11</td>
<td>0.357</td>
</tr>
</tbody>
</table>
Table 2. 9-Day study: demographics, motivation, and group differences.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>SE</th>
<th>F (1,114)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion (1-7)</td>
<td>3.3</td>
<td>0.15</td>
<td>0.37</td>
<td>0.869</td>
</tr>
<tr>
<td>Agreeableness (1-7)</td>
<td>5.4</td>
<td>0.11</td>
<td>1.24</td>
<td>0.296</td>
</tr>
<tr>
<td>Conscientiousness (1-7)</td>
<td>5.6</td>
<td>0.10</td>
<td>1.63</td>
<td>0.157</td>
</tr>
<tr>
<td>Emotional Stability (1-7)</td>
<td>5.1</td>
<td>0.14</td>
<td>0.55</td>
<td>0.742</td>
</tr>
<tr>
<td>Openness to Experiences (1-7)</td>
<td>5.2</td>
<td>0.12</td>
<td>0.79</td>
<td>0.561</td>
</tr>
<tr>
<td>Age</td>
<td>36.4</td>
<td>0.85</td>
<td>0.93</td>
<td>0.465</td>
</tr>
<tr>
<td>Identification as a Gamer (1-10)</td>
<td>6.3</td>
<td>0.25</td>
<td>3.98</td>
<td>0.002</td>
</tr>
<tr>
<td>Intrinsic Motivation (1-7)</td>
<td>5.2</td>
<td>0.12</td>
<td>1.20</td>
<td>0.312</td>
</tr>
<tr>
<td>Identified Regulation (1-7)</td>
<td>4.4</td>
<td>0.11</td>
<td>0.54</td>
<td>0.745</td>
</tr>
<tr>
<td>External Regulation (1-7)</td>
<td>1.8</td>
<td>0.10</td>
<td>0.62</td>
<td>0.688</td>
</tr>
<tr>
<td>Motivation (1-7)</td>
<td>1.9</td>
<td>0.11</td>
<td>0.41</td>
<td>0.844</td>
</tr>
</tbody>
</table>

After providing informed consent, participants answered a demographics questionnaire, and then were directed to watch a 1:30 min tutorial video, after which they were assigned to one of five conditions, i.e., Baseline, Rewards, Completion, Novelty, or Waiting—we removed Blocking as a condition for the pre-test, because the condition is defined by preventing play for an extended time, which was not possible in a 15-minute play session. Participants played for 15 minutes before being directed to questionnaires assessing pX. They could then choose to play for another 15 minutes.

Measures

To validate consistency of game design across the conditions, we logged reached scores, unlocked levels and number of turns as performance metrics. Participants indicated their agreement to items on a 7-point Likert-scale ranging from ‘Disagree strongly’ to ‘Agree strongly’. Player Experience was measured using the Intrinsic Motivation Inventory (IMI, [103]), and the Player Experience of Need Satisfaction Scale (PENS, [106]). The IMI assesses intrinsic motivation on four sub-scales: Interest-Enjoyment, Competence, Effort-Importance, and Tension-Pressure. PENS assesses satisfaction of Competence, Autonomy, and Relatedness, and two game-focused scales: Immersion and Intuitive Control. Conscientious participants are more likely—by definition—to adhere to directions, and adherence behaviour is our main outcome measure. Therefore, to determine if one group contained more conscientious participants than another we accessed personality using the Ten-Item Personality Inventory (TIPI, [47]), which describes five dimensions: extraversion, neuroticism, agreeableness, openness, and conscientiousness.

Results

Five participants were removed because they showed zero variance in their responses across multiple scales (>2), filled in multiple scales (>4) 1SD faster than the average, or scored ±3SD from the mean across multiple sub-scales (>2). All subsequent analyses were performed with n=94 participants. We performed a one-way multivariate analysis of variance (MANOVA) and showed no statistically significant differences of condition on any measure (see Table 3). Further, a one-way ANOVA showed no significant differences between the conditions on any of the demographic factors (see Table 1).

There were no statistically significant differences of condition on score ($F_{2,1080} = 0.22, \eta^2 = 0.003, p = .450$) nor on number of turns ($F_{2,1080} = 1.510, \eta^2 = 0.006, p = .197$). However, there was a significant difference on levels completed between conditions ($F_{2,1080} = 2.773, \eta^2 = .010, p = .026$).

To determine if the measures deviate from neutral, we performed One-Sample t-Tests against a neutral score of 4. The results (all $p < 0.001$; see Table 3) show a positive difference from neutral for Enjoyment, Competence, Effort, and Intuitive Control, suggesting that the game is perceived as enjoyable, challenging, and participants are willing to invest effort. Significant negative differences show for Tension, Relatedness, and Immersion. Autonomy showed no difference from neutral, which is not surprising, considering that players had little choice and were in an experimental context.

Discussion

We showed that the game versions are comparable in terms of pX, and additionally that the game met our intended overall experience: a single-player enjoyable—but not immersive or tense—game experience, in which players felt competent, with controls that were intuitive, and that they were investing effort. The comparable scores and number of turns as well as small effect sizes provide assurance that the game versions are comparable in playability and difficulty.

EVALUATING DYNAMICS FOR THE SNACKING PATTERN

After evaluating the game’s design in a short-term study, we deployed the five versions of the game plus the baseline in a 9-day behavioural experiment to explore whether the different dynamics create the snacking play pattern.

Methods

The design of this behavioural experiment closely mirrored the design of the preliminary study, except over an extended period of time. We only report differences in methods. As the study was also deployed on MTurk, we prevented workers who participated in the preliminary study to partake in the snacking behaviour study. Participants were informed that the study was split into three HITs, which totaled to $10, and ongoing informed consent was required to enter the system.

Day 0: On the initial day, participants were paid $4 to complete demographic and trait surveys, watch the tutorial, play the game for 15 minutes, and complete pX surveys (same procedure as preliminary study, about 25 minutes). After finishing, participants could choose to continue playing.

Day 1–7: Participants received a notification at midnight each day that reminded them to reengage in the study. Participants were able to log in at any time and play the game for as long as they wished—for example in the blocking condition, which only allowed 30 minutes of play every 8 hours.

Day 5: Participants were paid $1 to complete a 2-minute midterm survey with IMI, SIMS (Situational Motivation Scale, which measures intrinsic motivation, identified regulation, external regulation, and amotivation [7]), and open questions about what they liked and disliked about GOC. Comments were related to the general game and did not reference specific dynamics, thus are not reported further.

Day 8: On the final day, participants completed the exit survey. To qualify, they had to have logged into the game at least once
Table 3. Left: ANOVA and post-hoc t-test statistics for the IMI (top) and PENS (bottom). Right: One-sample t-tests against a neutral response (4.0)

<table>
<thead>
<tr>
<th></th>
<th>F (df)</th>
<th>p</th>
<th>η²</th>
<th>M (SD)</th>
<th>M (SD)</th>
<th>M (SD)</th>
<th>M (SD)</th>
<th>M (SD)</th>
<th>M (SD)</th>
<th>M (SD)</th>
<th>M (SD)</th>
<th>M (SD)</th>
<th>M (SD)</th>
<th>Mean diff</th>
<th>p</th>
<th>Intrinsic Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interest-Engagement</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intrinsic Motivation</td>
<td>0.071</td>
<td>0.991</td>
<td>0.003</td>
<td>5.11 (1.70)</td>
<td>5.17 (1.36)</td>
<td>5.17 (1.36)</td>
<td>5.17 (1.36)</td>
<td>5.21 (1.25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.071</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Baseline</td>
<td>1.376</td>
<td>0.249</td>
<td>0.058</td>
<td>4.31 (1.42)</td>
<td>4.69 (1.50)</td>
<td>4.91 (1.60)</td>
<td>4.91 (1.60)</td>
<td>4.91 (1.60)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.376</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Novelty</td>
<td>1.09</td>
<td>0.366</td>
<td>0.047</td>
<td>5.89 (0.57)</td>
<td>6.09 (0.63)</td>
<td>5.48 (1.47)</td>
<td>5.48 (1.47)</td>
<td>5.48 (1.47)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.09</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Waiting</td>
<td>0.946</td>
<td>0.441</td>
<td>0.041</td>
<td>2.84 (1.64)</td>
<td>2.98 (1.32)</td>
<td>2.47 (1.62)</td>
<td>2.28 (1.60)</td>
<td>3.11 (1.44)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.946</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Rewards</td>
<td>0.104</td>
<td>0.754</td>
<td>0.003</td>
<td>2.80 (1.64)</td>
<td>2.98 (1.32)</td>
<td>2.47 (1.62)</td>
<td>2.28 (1.60)</td>
<td>3.11 (1.44)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.104</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Completion</td>
<td>0.247</td>
<td>0.911</td>
<td>0.011</td>
<td>3.09 (1.76)</td>
<td>3.21 (1.63)</td>
<td>3.48 (1.25)</td>
<td>3.46 (1.43)</td>
<td>3.18 (1.38)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.247</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Need Satisfaction</td>
<td>0.282</td>
<td>0.889</td>
<td>0.013</td>
<td>4.87 (1.17)</td>
<td>4.57 (1.45)</td>
<td>4.72 (1.35)</td>
<td>4.88 (1.40)</td>
<td>4.97 (1.05)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.282</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

after Day 0. Participants completed the IMI, SIMS, and open-ended questions about why they did and did not play GOC. They were paid $5 for about 10 minutes. We were careful in designing the payment structure, as previous work has shown that the timing of payment in long-term studies can affect volitional engagement and motivation [18]. We paid $5 for Day 8 to be ethical about payment for the interim days in which payment was not explicitly provided; it allowed us to interpret the play sessions from Day 1-7 as volitional engagement.

**Measures:** We logged the exact time of each play session, the start time and duration of game matches, overall progression and performance in the game (i.e., unlocked levels, collected items, number of turns in each level). We again gathered TIP1 [96] to control for conscientious participants.

**Participants:** We recruited 180 people (49.2% female, mean age=35.0, SD=9.3) across the six conditions.

**Data Analyses:** We logged data from individual game matches (i.e., a single attempt to destroy a castle) directly to the server, but were interested in *game sessions* with only short breaks. We decided to use game sessions as our metric of engagement, since common metrics of "stickiness" such as Daily Active Users (DAU) [36, 46] only measures how many users are active but does not cover the duration of active usage. In a game session, a player might play several matches, with exploration of other aspects of the game (e.g., world or level selection) in between or simply pause between matches. Because longer breaks might indicate moving to another task, we chose 3 minutes as the threshold between matches that indicates a single game session. We aggregated play sessions into a single game session, calculating performance metrics (i.e., average match score, total number of turns, starting level, ending level, number of levels completed) and game session characteristics (i.e., start time, end time, session duration, time from previous session, time to next session, number of sessions). Because we told participants that they had to play the initial session plus one other play session to be paid, we filtered out those who only completed this minimal payment threshold; remaining participants (n=128) played at least one game session not motivated by the external reward of payment.

**Characterizing the Sample and Comparing Groups**

After filtering participants, there were comparable numbers in each group (N: Baseline=20; Novelty=24; Waiting=20; Rewards=21; Completion=22; Blocking=21). As Table 2 shows, there were no significant differences in any of the demographic factors, incoming motivation, or personality, showing that random assignment to groups was effective. There was a difference in terms of how much the participant identified as a gamer on a visual analog scale (F₅,114=3.98, p=0.002, η²=0.149)—a question that was previously shown to significantly correlate with a 60-item scale on self-attributes [84]. Post-hoc tests (Games-Howell) showed that participants in the Rewards condition identified less on average as a gamer than those in the Novelty (p=.007) or Waiting (p=.011) conditions, but that no other differences were significant. Because our analyses compare each mechanic to the Baseline condition, and there were no significant differences with Baseline, we do not further control for the lower gamer identity in the Rewards group.

**Game Experience**

To determine whether the dynamics yielded different pX, we compared enjoyment, measured three times (Days 0, 5, and 8). A MANOVA showed no differences in enjoyment due to dynamic after day one (F₅,22=0.80, p=.556), at the midterm point (F₅,22=0.95, p=.451), or after the exit survey (F₅,22=0.62, p=.687). These results corroborate the results from the pre-study and also demonstrate that differences in enjoyment did not emerge over the course of repeated play.

**Game Performance and Play Behaviour**

To determine how comparable the dynamics were in terms of performance, we first conducted a one-way ANOVA on the game session performance metrics with the dynamics as the between-subjects factor. There were significant effects of dynamic on average score (F₅,122=9.138, p<0.001) and levels completed (F₅,122=2.952, p=0.015), and a marginal effect on the number of turns (F₅,122=2.713, p=0.061), suggesting that the dynamics yielded different performance.

Table 4 shows correlations between performance and game session characteristic measures. Because average score correlated significantly with inter-session interval and session count (whereas number of turns or levels completed did not), we controlled for performance using score in tests on inter-session interval and session count. As game session duration correlated significantly with score, turns, and levels, we used the performance metric with the highest correlation (turns) as the control for game session duration.

**Results**

To see if the different dynamics produced different patterns of play, we conducted one-way ANCOVAs with dynamic as
the between-subjects factor, controlling for performance, as previously described. Controlling for average score, there was a significant effect of dynamics on inter-session interval ($F_{4,121}=2.469$, $p=.036$, $\eta^2_p=.093$). The planned contrasts (comparing each dynamic to Baseline) showed that there was a shorter inter-session interval with Novelty ($p=.008$) or Waiting ($p=.041$), but not Rewards ($p=.409$), Completion ($p=.871$) or Blocking ($p=.139$); see Figure 2.

Figure 2 shows greater session durations in Waiting and Blocking compared to Baseline; however, controlling for the number of turns, the differences in session duration were not significant ($F_{5,121}=17.718$, $p=.547$). There were more game sessions in Waiting compared to Baseline; however, controlling for score, this difference was not significant ($F_{5,121}=1.163$, $p=.331$).

**Clustering Players into Pattern Types**

Looking at individual game session characteristics is helpful to inform how the dynamics may have differently influenced patterns of play; however, our goal is to observe a specific pattern of play characterized by a particular session duration (15 minutes) and inter-session interval (24 hours). As such, we used the game session duration, interval and count to cluster all players (regardless of dynamic) into two groups using k-means clustering. We chose $k=2$ with the goal of splitting the participants into ‘snacking’ and ‘non-snacking’ groups.

As shown in Table 5, session duration did not contribute significantly to clustering and there was little difference between the cluster means. However, the clusters are significantly differentiated by inter-session interval and number of sessions played. In Cluster 1, the average inter-session interval is longer and the number of game sessions is fewer. Cluster 2 is defined by approximately daily play sessions (22.3 hours between sessions, 8.9 game sessions over 9 days) that last around 13 minutes—which is the snacking pattern we are looking for.

Figure 3 shows the relative cluster membership for each dynamic play group. In the baseline condition, 75% of the players were clustered into the snacking pattern group (Cluster 2). This is not surprising as in the context of a paid experiment, participants are known to be more compliant than during app use “in-the-wild” [91] and our game was designed based on a successful mobile casual game. The remaining charts show the added value of each mechanic: a greater proportion of players were classified into the snacking cluster when any dynamic was added. To determine whether the added value of each mechanic was significant, we conducted chi-squared tests for each condition, with the 25/75 ratio of the baseline condition
as the expected case membership. Results showed that Novelty ($\chi^2_1=5.6, p=.018$) and Waiting ($\chi^2_1=4.3, p=.039$) were successful at promoting the snacking pattern of play significantly more than baseline. Rewards ($\chi^2_1=1.3, p=.257$), Completion ($\chi^2_1=0.5, p=.460$), and Blocking ($\chi^2_1=2.7, p=.101$) were not significantly improved over Baseline. Given the small number of participants in the study, and the relatively large proportion in the snacking cluster, these results should be interpreted cautiously, but do affirm the results from the ANCOVA, providing additional evidence for the potential of Novelty and Waiting.

**DISCUSSION**

The performance differences (i.e., scores and completed levels) can be explained by differences in game design. Naturally, the conditions with power-ups (Rewards and Waiting) yield higher scores. Further, in Novelty, the weapons are directly tied to specific levels and thus, are balanced differently. These differences are difficult to control for and are not apparent in a single play session. However, participants were not exposed to the scores of the other players or conditions, so scores were not expected to influence behaviours.

Novelty and Waiting had significantly lower inter-session intervals than Baseline, even when controlling for performance differences. Further, these two conditions had significantly more players classified into the snacking pattern than was predicted by classification in the Baseline condition. Blocking also showed promising results that failed to reach significance. There was no evidence that Rewards or Completion provided significant added value over the Baseline condition for promoting snacking. The sizes of the groups were relatively small and therefore, the findings cannot be conclusive. However, the observed differences in play patterns are not explained by accompanying differences in rated enjoyment, in motivation for participating in the experiment, or by personality factors, such as conscientiousness, suggesting that the mechanics themselves are responsible for the snacking pattern of play.

**Explanation of the Results**

The snacking gameplay pattern aims to establish long-lasting behaviour change. Novelty and Waiting mechanisms both contributed significantly to the snacking pattern. Based on theory, we can assume that Novelty drew the attention of players and prevented boredom, whereas Waiting added aided conditioning effects by binding certain actions to an in-game trigger. These are also characteristics of theories on habitual behaviour and addiction, in which an action is performed without volition, that help to explain our findings. Repetition of a behaviour in a consistent context progressively increases the automaticity with which the behaviour is performed when the situation is encountered [126, 129]. Automaticity is evidenced by the behaviour, displaying some or all of the following features [11]: efficiency, lack of awareness, unintentionality and uncontrollability. The salient cue itself (e.g., opening the laptop or picking up the smartphone) will trigger the behaviour, e.g., playing the game. This means that it is no longer necessary for a person to want to play since the behaviour is mostly controlled by the contextual cue [97, 127]. Consequently, according to the self-perception model, an attitude towards something (e.g., a game) can be formed by observation of our own actions, especially when internal cues are weak [12]. As a result, when there is no reason to keep on checking on the game, we are more likely to like it, because otherwise we would experience unpleasant cognitive dissonance [35]—behaviourally we revisit the game, but lack a cognitive explanation for our behaviour. Critically, achievements in the game further increases its perceived value and resulting positive valence [74, 79].

Apart from habit, addictive behaviour does not require volition and is sustained for a long time. A game that keeps players drawn in over a long period of time can be said to exhibit addictive characteristics, although it is imperative to acknowledge that repetitive and sustained behaviour is not evidence of addiction—especially if not pathological in nature (i.e., negatively affecting other aspects of one’s life) [48]. One theory, however, that might be of importance for the creation of games that leverage similar patterns of behaviour is reversal theory [4, 118, 5], which assumes two arousal states: telic (i.e., serious) and paratelic (i.e., playful). In a telic state, a person shows goal-directed behaviour, which is motivated by the achievement of future goals; whereas in a paratelic state, the focus is on the action itself and the person gains enjoyment from the process. In a telic state, arousal is undesired, whereas in a paratelic state, arousal is welcomed. Therefore, if a game offers both high and low arousal gameplay—it may be used any time for its properties by the player [2].

**Implications for Game Design**

Many games would arguably benefit from a design that fosters snacking. Keeping players returning regularly is essential for financial success in business models of free-to-play games [36]. Game designers have argued against blocking mechanisms as a heavy-handed way of enforcing session length restrictions; however, as noted by [64], the majority of mobile games in the top grossing charts include this type of control, and few can remain at the top of the rankings without embracing session length restriction. Mechanics built around blocking and waiting are necessary, as without them, casual games risk becoming boring or running out of content [67]. Although this study shows that the snacking pattern can be promoted using waiting mechanics, this play-pattern is often forced on players and exploited in casual games for monetization purposes [67, 64, 36]. Unlike players of monetized casual games, the participants in the present study had no option to circumvent the waiting mechanics (or any other mechanic). Future research should aim to investigate this aspect further.

Nevertheless, our results suggest that waiting mechanisms might help establish a consistent habitual gameplay pattern. Along with Novelty to keep players returning, we confirm the suggestion from design experts that waiting mechanics might influence behaviour regardless of demographic variables or personality traits, provided that the game itself is enjoyable.

**Serious Games**

Changing our habits and behaviours is difficult, whether it be to learn a new skill or improve our well-being. Serious games have been described as games with a purpose, that are not intended to be played solely for amusement [1], and are often designed to make beneficial but challenging activities more fun. Gamification—the use of game design elements
in non-game contexts [31]—has successfully applied motivational theories to the design of applications intended to support behaviour change, learning, or habit formation.

Although our study employed a casual game, we see potential of dynamics that support snacking for the design of serious games or games for change. There are examples of serious games that draw players in and keep them engaged in an immersive experience—for example, in Darfur is Dying [117], players experience life as a refugee in Darfur in a single immersive play session. However, there are many types of serious games that would be most effective if players were motivated to play daily for a short duration, rather than for hours in a single session. Consider intervention design for physical or mental health, in which effectiveness depends in part on repeated short bursts of engagement over time rather than lengthy periods of sustained engagement [45, 93]. Citizen science games in which player’s actions benefit a scientific or civic purpose (e.g., Foldit [70], Eye Wire [122], and Phylo [68]) would also benefit from frequent and regular engagement over time, rather than an isolated one-off play session. Similarly, spacing out verbal learning has been shown to afford people to develop strategies that scaffold success [29], which could be accounted for in a game designed around learning verbal information. Our results suggest that novel and personalized elements might reduce boredom and increase engagement in the short term, whereas waiting mechanics might prevent apathy from overplay by encouraging brief and repeated play sessions.

Limitations and Future Work

The five game dynamics were obtained from an analysis of casual games conducted in June 2017, and are not a comprehensive representation of all possible approaches in games for fostering the snacking pattern of interaction; however, they serve as a starting point to collect and evaluate game mechanics. In the future, we will expand this framework and collect quantitative metrics of game mechanics from game distribution platforms (e.g., Steam, Google Play, iTunes store).

We investigated single-player dynamics. It is known that social play mechanics can motivate behaviour and sustained participation [73, 104]; however, we were interested in first investigating the mechanics of single-player experiences and we intend to specifically investigate the mechanics and dynamics of social play for serious game design in future work.

Although blocking mechanics showed promise in our study, they failed to reach significance. Blocking forces players to wait to re-engage with the game and prevents disengagement from overplay. Our average session length was 12-13 minutes; however, our blocking mechanism was triggered at 15 minutes. The lack of success of blocking in our study may simply be a result of our game being designed to be enjoyed in a play session shorter than that triggered by blocking. Future work should investigate blocking explicitly with a compelling game that draws players in for a longer duration. Further, the success of Waiting was in the return of players, not by limiting overplaying; in future work with more engaging games, the waiting mechanic may prove even more useful.

We discuss the potential of our work to inform the design of games for change; however, we used a casual playful game in our studies. It was necessary to first establish that the mechanics identified in our content analysis of casual games translated into observable behaviour differences when deployed in the same style of game (i.e., casual playful game) from which they were identified. In future work, we plan to study whether their application in a serious games context is effective.

There are several factors that affected our results. The sample size (n=128) was small and the duration (days=9) of the experiment was too short to observe long-term behaviour. Daily notifications may have influenced motivation [16]. Although we were very considered about the design of our payment plan, it is possible that paying participants may have made them feel compelled to participate as a responsibility to us as researchers. Further, we deployed a web version of the game for desktop computers to facilitate experiment deployment and logging; however, the mechanics were drawn from games primarily played on mobile devices. We plan to conduct an in-the-wild investigation of the study that is deployed on mobile devices; it may be that the built-in notifications and persistent app button influence login behaviours. Finally, the dynamics were evaluated in isolation to determine their added value over and above a baseline condition; future studies will look at whether they can be combined for additional benefit in fostering snacking. This strictly experimental procedure might be argued to lack ecological validity and psychological concepts may not be directly transferable to the realm of games. However, it has been argued that strictly controlled immersive technological environments can facilitate ‘mundane (or experimental) realism’ [19, 61], which leads to natural participant responses while not losing experimental control [32].

CONCLUSION

Games user research often focuses on frameworks to keep players in an immersive state of flow, facilitating long-lasting game sessions. We argue that some games are most enjoyable, profitable, or effective not when played in a few long game sessions, but when played regularly and frequently for a short time (a snacking pattern). To identify which game mechanics best facilitate snacking, we conducted a content analysis of game mechanics from casual mobile games that have a strong player base with a high retention rate and classified them into five dynamics. We then examined how theories of motivation explain the efficacy of the dynamics. We designed a casual physics puzzle game, in which each of the five dynamics could be included individually or in combination. After a pre-study, in which we verified that the game met our enjoyable-but-not-immersive design criteria and that all versions were comparable in terms of pX, we conducted a web-based behavioural experiment with 180 participants over 9 days. Behavioural patterns showed that Novelty and Waiting dynamics added significant value over the Baseline game in terms of promoting the snacking pattern of play.
# APPENDIX

## TABLE OF GAMES INVESTIGATED FOR THE CONTENT ANALYSIS

<table>
<thead>
<tr>
<th>Game Title</th>
<th>Main Genre</th>
<th>Sub-Genre</th>
<th>Other Blocking</th>
<th>Novelty</th>
<th>Waiting</th>
<th>Rewards</th>
<th>Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disney Crossy Road</td>
<td>Collection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tap Quest</td>
<td>Collection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charming Keep</td>
<td>Idle Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bitcoin Billionaire</td>
<td>Idle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cookie Clicker</td>
<td>Idle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doomsday Clicker</td>
<td>Idle Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Octopus Evolution</td>
<td>Idle Puzzle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Realm Grinder</td>
<td>Idle Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fallout Shelter</td>
<td>Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simcity Buildit</td>
<td>Management</td>
<td>Player Vs. Player</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hay Day</td>
<td>Management</td>
<td>Player Vs. Player</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CivCrafter</td>
<td>Management</td>
<td>Player Vs. Player</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lords Mobile</td>
<td>Management</td>
<td>Player Vs. Player</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.A.T.S</td>
<td>Player Vs. Player</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clash Royale</td>
<td>Player Vs. Player</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triple Town</td>
<td>Player Vs. Player</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Candy Crush: Soda Saga</td>
<td>Player Vs. Player</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mr. Square</td>
<td>Puzzle Strategy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pac-Man</td>
<td>Puzzle Strategy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clash of Clans</td>
<td>Strategy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subterfuge</td>
<td>Strategy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### REFERENCES


[103] Richard M. Ryan. 1982. Control and Information in the Intrapersonal Sphere: An Extension of Cognitive...
Serious Snacking: A Survival Analysis of how Snacking Mechanics Affect Attrition in a Mobile Serious Game

Dmitry Alexandrovsky
DMLab, University of Bremen
Bremen, Germany

dim@uni-bremen.de

Maximilian Achim Friehs
University College Dublin
Dublin, Ireland

maximilian.friehs@ucd.ie

Jendrik Grittner
University of Bremen
Bremen, Germany

jendrik@uni-bremen.de

Susanne Putze
DMLab, University of Bremen
Bremen, Germany

sputze@uni-bremen.de

Max V. Birk
Eindhoven University of Technology
Eindhoven, Netherlands

m.v.birk@tue.nl

Rainer Malaka
DMLab, University of Bremen
Bremen, Germany

malaka@tzi.de

Regan L. Mandryk
University of Saskatchewan
Saskatoon, Saskatchewan, Canada

regan@cs.usask.ca

ABSTRACT

Many serious games are most effective when played regularly; however, little is known about how individual game elements support player adherence over time. This work draws on evidence from existing frameworks and game design theories as well as from the design of casual games to investigate how individual game mechanics affect player attrition in a serious game. We implemented a math-learning game in which we could individually layer various game mechanics, and over the course of 3 weeks, 99 participants played one of six versions: Baseline, Rewards, Novelty, Completion, Waiting, or Blocking. We compared the game versions by analyzing the players’ performance as well as behaviour. Using survival analysis, we identified that the addition of Completion and Blocking mechanics facilitated the strongest sustained engagement. These findings are congruent with existing theories of player experience and promote the development of guidelines on designing for sustained engagement in serious games.

CCS CONCEPTS
- Human-centered computing → HCI theory, concepts and models. Empirical studies in HCI.

KEYWORDS
Game mechanics; rewards; novelty; missions; blocking; waiting; game mechanics; design; adherence; snacking.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '21, May 8–13, 2021, Yokohama, Japan
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8096-6/21/05...
$15.00
https://doi.org/10.1145/3411764.3445689

1 INTRODUCTION

Changing our habits and behaviours is difficult, whether it be to learn a new skill or improve our wellbeing. One way to encourage change in individuals is the use of serious games. Serious games are described as games with a purpose, that are not intended to be played solely for amusement [2], and are often designed to make beneficial but challenging activities more fun. Gamification describes the method by which game design elements are used in non-game contexts [23]. Gamification has been applied successfully to design of applications intended to support behaviour change, learning, or habit formation [9, 56, 82, 109]. There are various frameworks and approaches for describing engagement [48, 78, 103] and evaluating the efficacy of engagement-enhancing gamification elements [12, 108]. However, theories of engagement frequently focus on prolonged play sessions—e.g., flow theory [16] is defined in part by absorption and loss of a sense of time passing [34], whereas in self-determination theory (SDT) [92], engagement is often operationalized as time spent in free play, e.g., [6].

There are examples of serious games that draw players in and keep them engaged in an immersive experience—for example, in Darfur is Dying [104], players experience life as a refugee in Darfur in a single immersive play session. However, there are other types of serious games (often called games for change in that they are designed to change attitudes, behaviours, or skills) that would be most effective if players were motivated to play regularly (e.g., daily) for a short duration, rather than for hours in a single session. For example, a motion-based game exercise designed to help people engage in physical rehabilitation after an injury is not only more effective if played regularly for a brief duration, but could
be dangerous for a player if they overexert themselves in a single session because they are in a state of flow [38]. Similarly, spacing out verbal learning has been shown to afford people the opportunity to consider and develop strategies that scaffold success [19, 96] which could be accounted for in a game designed around learning verbal information. Or consider, for example, intervention design in the domain of mental health, in which effectiveness depends in part on repeated short bursts of engagement over time rather than lengthy periods of deep engagement [79]. There are many examples of successful commercial games that are designed for sustained engagement over time [60]; in fact, a common measure of success for mobile app designers is "stickiness" which operationalizes repeated use as the proportion of monthly active users who engage daily [29, 75]. In particular, casual games played on mobile devices tend to encourage players to engage daily in brief play sessions as they support a pattern of interaction called "snacking" [79], in which users graze frequently, but for short bursts of time [83, 110].

In the context of serious games for learning or behaviour change, it is clear that designing for snacking could support a pattern of sustained game engagement needed for habit formation, knowledge integration, and long-term adoption; however, the problem is that game designers have little guidance on which game mechanics promote snacking and stickiness.

Alexandrovsky et al. [3] developed a framework of five game mechanics—Rewards, Novelty, Completion, Waiting, and Blocking—which encourage sustained engagement in successful commercial casual games [60] (see Section 2.2). The authors evaluated the framework using a casual game and showed that these game mechanics can facilitate player behaviour that is characterized by two aspects: not playing for too long in a single session (i.e., resisting burnout) and playing frequently and repeatedly over time (i.e., sustaining engagement). However, it is unclear whether their results translate into engagement with serious games that are played for a secondary purpose (i.e., learning, habit formation, or behaviour change), or into games engaged with completely under a player’s own volition, as their study paid participants and sent daily notification reminders to log in.

Our study advances the work from Alexandrovsky et al. [3] and differs in three ways: (1) we apply the snacking framework to a serious game for learning in contrast to a casual game for entertainment, (2) the game is evaluated in a field study in which the participants play the game under their own volition, and (3) the game is played on the participant’s own mobile device instead of a desktop PC, in line with how casual games tend to be consumed by players. Further, we employ survival analysis—a group of statistical methods that model the time span until a particular event occurs (e.g., How long do patients survive using a particular treatment?, How long will a machine last?) [14, 76]—to investigate how different game mechanics affect player adherence over time. By applying the framework in a new context, we evaluate its use for serious games played under a user’s volition on their mobile device and show how the mechanics support adherence and thwart attrition. Thus, this research is guided by the following research questions:

RQ1: Do the snacking mechanics yield different player experiences when implemented in a serious game?

RQ2: Can the snacking mechanics provoke a snacking behavioural pattern of game engagement when deployed in an uncontrolled field study?

RQ3: Which snacking mechanic facilitates the strongest adherence over time?

2 BACKGROUND

To facilitate motivation, games user research (GUR) draws on game design, and the incorporation of game elements into mundane or tedious tasks to make them more appealing, resulting in the coining of the terms serious games [2] (i.e., implementing objectives outside the game content into game design), and gamification [23] (i.e., adding game elements to utilitarian information systems) as motivational strategies. Research on player experience (PX), which is concerned with all aspects of players and their experiences, has built an understanding of how individual personalities gravitate towards specific game components. There are several frameworks of player types [5, 70, 107] or player traits [86, 87, 106], which build upon classifications of personality (e.g., Myers-Briggs type indicator (MBTI) [80], Five Factor Model (FFM) [41]) or motives (e.g., Motive Disposition Theory (MDT) [72]) and link them to theories of motivation (e.g., SDT, e.g., [45]). However, recent research argues that static or dominant player types cannot solely explain the motivation of playing certain games [1, 45, 81]. Therefore, research is moving from static player types towards dynamic player traits which allow a better conceptualization of individuals’ preferences in games [1]. Although a growing body of research has successfully applied game design to various disciplines (e.g., health [8, 64, 66], education [21, 22, 94], human computation [112] among others) to foster motivation and sustained engagement, player behaviour cannot be predicted solely from motivation, affective computing, or cognitive modeling, as games are a dynamic and ergodic media, i.e., a player interacts with them and alters their state [116]. For example, Mekler et al. investigated the effect on motivation and engagement of points, leaderboards, and levels—three commonly-applied game elements [78, 89]—in an image tagging task and found that all of these elements improved the tagging quality and facilitated player adherence over baseline, without affecting intrinsic motivation but rather by serving as indicators of progress towards a specific goal. Therefore, the authors argued that sustained long-term engagement should not exclusively build on these game mechanics and cannot be fully explained by motivation [73, 74]. Moreover, Jolley et al. investigated the role of habit and satisfaction for the retention of online gambling players. The authors differentiated between behavioural models of player retention (frequency-retention-profit) and satisfaction-retention-profit paradigms and argued that exclusively behavioural models are insufficient to inform future design iterations as they lack explanatory insights [53].

2.1 Player Behaviour and Attrition

To address a wider audience, to improve game experiences, and to maximize profits through stickiness, the game industry and games user researchers have shifted focus from informal game testing towards real-time user tracking to assess and analyze player behaviour [24, 57], adjust the game design in general [28, 99] for individualized needs, e.g., dynamic difficulty adjustment (DDA) [47], or
to balance for a personalized emotional response to avoid frustration [39]. Zoeller describes this research method as telemetry [119] in which developers collect data remotely from the users’ devices. Presently, most commercial games and multimedia platforms implement some type of user analytics [44]. For example, by default, games released on Steam\(^1\) or on mobile platforms (e.g., Google Playstore\(^2\), Apple App Store\(^3\)) track user retention and provide APIs to monitor in-depth usage.

Most commonly, performance analyses rely on user-generated events [57] such as the number of clicks, in-game actions, and player performance, e.g., number of completed levels, number of kills [25]. As estimates of behaviour, the literature frequently applies playtime, session length, and inter-session intervals [44, 98]. For example, Weber et al. [113] developed a tool that employs regression models to predict which game elements are most responsible for player retention in a football game. To inform design decisions of web-based health interventions addressing chronic conditions, lifestyle, and mental health, Kelders et al. [56] conducted a literature (n=101) review of persuasive technology elements and extracted the reported adherence of users. Using a hierarchical multiple linear regression, the authors derived 11 persuasive elements that explain 55% of the variance for user adherence [56]. Likewise, Lumsden et al. [64] reviewed the success (i.e., adherence) of gamification strategies for cognitive training from n=33 studies and report mixed effects of gamification on task performance. Pfau et al. applied deep neural networks to model player behaviour for player substitution [84] and DDA [85], whereas Frommel et al. [32] used similar computational approaches to model the affiliation that develops between players in a multiplayer game context. Other research used emotional [30, 31, 33] or physiological responses [58] to model the link between challenge and enjoyment [116].

Game telemetry is increasingly used to understand UX, but is also at the root of understanding player attrition and retention, and several metrics of adherence have been applied in the literature. Carlo et al. [13] investigated number of monthly active users (MAU) per normalized number of downloads as a metric of “stickiness” (i.e., adherence) for mobile behavioural health applications. While this and similar metrics, such as daily active users (DAU) [29, 40] or DAU/MAU (i.e., stickiness [29, 75]), are commonly applied in market research, this method lacks granularity as it measures only how many users are active but does not cover the intensity of active usage. Viljanen et al. [111] distinguish between three types of retention metrics for cohorts: (1) retention is the proportion of users who play since a particular day, (2) rolling retention is the proportion of active users for a period of time, and (3) lifetime retention is the proportion of users who play at least for a particular period of time. Based on survival analyses, the authors derived a parametric model to predict player activity in 5 commercial games. The authors tracked the player activity of 5 commercial games over the course of several months and applied a survival analysis on discretized one-day bins from which they could predict retention, rolling retention, and lifetime retention [111]. Eysenbach [26] discussed patients’ attrition when using computer-mediated interventions. Described as “the law of attrition”, the author contended that most users of eHealth applications drop off before completing their therapy and argue that although high dropout rates are natural for computer-mediated treatment, “determinants of attrition should be highlighted, measured, analyzed, and discussed” to improve the efficacy of the treatment. Eysenbach presented multiple factors that influence the non-usage of eHealth interventions and propose that survival analyses such as Kaplan-Meier analysis [54] or Cox regression [15] should be applied as additional metrics of intervention efficacy and usability [26]. Following this approach, Demediu et al. [20] performed a survival analysis using a Kaplan-Meier estimator and Cox regression to detect disengagement of League of Legends [36] players and found that the success rate of players is responsible for retention. In this work, we build on these approaches and investigate how the different snacking mechanics can be applied in order to facilitate brief but consistent player behaviour over a three-week time span.

2.2 The Snacking Framework

As this work builds on the snacking framework developed by Alexandrovsky et al. [3], in this section, we briefly describe the framework, how it was developed, and how it was evaluated in the context of a casual game. For more details, we refer to the original paper.

From a content analysis of n=22 popular casual games, Alexandrovsky et al. [3] derived five game elements that are commonly applied in contemporary mobile casual games to foster regular but brief playing behaviour—the snacking pattern—which the authors describe as daily play for around 15 minutes. The derived game elements were embedded into existing theories of behaviour and motivation and were evaluated in an online study with n=128 using an Angry Birds-like [49] casual game called Games of Cannons (GOC). The results of the user study demonstrated the efficacy of the framework by showing that Novelty and Waiting facilitated the snacking behaviour when it was included in the game. However, the other mechanics (Blocking, Completion, and Rewards) appeared to warrant further investigation. The snacking framework consists of the following five game mechanics:

**Blocking.** Blocking is defined as “passively regenerating a resource that is required to interact with the game” such as a time contingent or a credit that allows play for a limited time or players to perform a finite number of actions. Blocking prevents boredom from overexposure and can be described as a “conditioning mechanism with a fixed reward scheme”, which amplifies the perceived value of the next session [43, 50].

**Completion.** Completion is described as gratifications from explicitly assigned tasks, e.g., achievements, quests, or goals. The authors distinguish between three types of Completion: (1) Daily login is a conditioner to engage the players to return on a regular basis. (2) One-off missions are single tasks that feed “a potential or existing habit”. (3) Chains are a series of successive tasks that need to be completed within a single play session. Completion gives players a choice on how to progress through the game and satisfies the players’ needs for competence, autonomy, and relatedness [92, 93]. Moreover, Completion addresses the Zeigarnik Effect and initiates a pull to complete unfinished tasks [68, 105].
Novelty. Novelty introduces new in-game elements to the player, e.g., items, game mechanics, assets, or goals. It prevents boredom from monotony [102] by providing new stimuli to the players, which draw their attention [17, 27, 52].

Rewards. Rewards are described as gratifications that are given to the players immediately after they complete the deserving action. Frequently, rewards are designed as a currency that communicates the player’s performance, e.g., score points [62]. The triggered behaviour of rewards can be explained by Skinner’s operant conditioning paradigm in which the desired behaviour is that rewarded leads to a higher likelihood of that behaviour being repeated [101].

Waiting. The experience of waiting results from a resource that is regenerated over time and restricts the player's actions, such as cooldowns. In contrast to Blocking, with Waiting mechanics, players are still allowed to play, but their freedom of action is limited. For example, in construction strategy games, the cost to build entities grows over time. The motivational mechanism behind Waiting is explained by delayed gratification [77], which helps establish a habitual behaviour.

Alexandrovsky et al. [3] evaluated the framework by deploying a base version of GOC, along with five versions that layered on each mechanic individually. After confirming comparable experience in a pre-study, they asked players to engage with their randomly-assigned version for nine days, reminding players with daily notifications sent via email, and remunerating participants on day 1 and again on day 9 if they logged in again at least once and completed the exit survey. The authors demonstrated that Novelty and Waiting showed the most promise for supporting snacking.

3 SERIOUS GAME DESIGN

Figure 1: Game interface of Infinitus Algebraica. Left: player statistics accessible from the main menu. Center: level and difficulty selection screen. Right: screenshot of a play round.

In line with the study design in [3], our goal was to design an engaging, but not too immersive game in which we could include the snacking mechanics individually. However, our game should have serious attributes [2]. Therefore, we designed “Infinitus Algebraica”—a simple math learning mobile game that addresses mental math skills—as our platform for the experiment. We chose this genre as math is free from language hurdles and is required in a wide range of fields [90]. Furthermore, mathematics allows for a fluid adjustment of difficulty settings depending on prior knowledge and experience, helping to address a wide audience.

As a starting point for the game design, we looked at “Infini-Math” [5]. InfiniMath shows numbers and operators (+, −, × and ÷) arranged in a grid. The goal is to select all tiles in the correct order in order to solve all equations before the time runs out. An exemplary gameplay round is depicted in Figure 2.

We extended the game and modified it for our purposes: we added a logging system for all actions as well as features to display questionnaires inside the game. To create sufficient content for a long-term study, we added 2 additional operations (×, mod.), extended the procedural equations generation and difficulty levels. As modifiers of difficulty, we applied the grid size (2×2, 3×3, 4×4, 5×5, 6×6), time quota (ca. 30–380 sec.) and ranges of possible values (5–150) as parameters. In total, the game generated 350 levels, which are grouped in 35 packages of 10 levels (rounds). With the progression through the packages, the players play on larger grids, need to perform more different math operations while having less time to complete the levels. The progression of the difficulties is depicted in Table 1 in the appendix. Further, we generated 4 levels of difficulty (easy, normal, hard, and master), at which the game can be replayed. The difficulty levels add an additional modification on the range of possible values in the grid (i.e., calculating with larger numbers), and time quota; with increasing difficulty the values becoming larger. To provide a ground for the snacking mechanics, we implemented 13 power-ups with an inventory system. A detailed description of the power-ups is provided in Table 1. Additionally, we added 8 visual skins and 5 background music tracks to be used as rewards. Iteratively and with two beta testers who had prior experience in game balancing and beta testing, we adjusted the difficulty levels and balanced the probabilities for the rewards to provide a sufficient curve for a play experience over the course of 3 weeks. The final game interface is depicted in Figure 1. The game is available on the Apple App store 5 and Google Play store 6.

3.1 Game Versions

As in the study by Alexandrovsky et al. [3], we designed the game such that each of the 5 snacking mechanics can be individually layered on top of a base game version.

Base game version. For the base game, we disabled the power-ups and the inventory system. Also, all visual skins, background music, and all levels are unlocked from the beginning.

Rewards. We used a scoring mechanic with power-ups as rewards. The score was composed of the player’s accuracy in each level and the progress in the current package. Based on the score, we calculated a chance to receive a power-up.

1https://assetstore.unity.com/packages/templates/packs/infinimath-endless-math-puzzle-59911
2https://apps.apple.com/us/app/id1490357788
Table 1: Infinitus Algebraica Power-Ups

<table>
<thead>
<tr>
<th>Name</th>
<th>Duration</th>
<th>Chance</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Save Token</td>
<td>single use</td>
<td>0.06</td>
<td>Credit to save progress with a package</td>
</tr>
<tr>
<td>Time Booster</td>
<td>10 min.</td>
<td>0.1</td>
<td>30 sec. more time in each level</td>
</tr>
<tr>
<td>Combo Time Booster</td>
<td>5 min.</td>
<td>0.126</td>
<td>5 seconds more time to complete a combo</td>
</tr>
<tr>
<td>Brain Booster</td>
<td>4 min.</td>
<td>0.075</td>
<td>30% probability for highlighting corresponding tiles</td>
</tr>
<tr>
<td>XP Booster</td>
<td>5 min.</td>
<td>0.06</td>
<td>50% more experience points per round (Novelty)</td>
</tr>
<tr>
<td>Simplifier</td>
<td>2 min.</td>
<td>0.075</td>
<td>Reduces the grid size by -1</td>
</tr>
<tr>
<td>Reducer</td>
<td>3 min.</td>
<td>0.126</td>
<td>Reduces the range of possible values for the generated equations</td>
</tr>
<tr>
<td>Unlocker</td>
<td>2 min</td>
<td>0.023</td>
<td>Highlights the parts all equations at once; each equation is highlighted with a different color</td>
</tr>
<tr>
<td>Transformer</td>
<td>3 min.</td>
<td>0.06</td>
<td>Only equations with addition are generated and the range of possible values +1</td>
</tr>
<tr>
<td>Randomizer</td>
<td>single use</td>
<td>0.1</td>
<td>Activates a random other power-up.</td>
</tr>
<tr>
<td>Glitcher</td>
<td>5 min.</td>
<td>0.126</td>
<td>Last equation of a package is solved for the player</td>
</tr>
<tr>
<td>Skipper</td>
<td>single use</td>
<td>0.023</td>
<td>Skip three levels</td>
</tr>
<tr>
<td>Quantum Chip</td>
<td>4 min.</td>
<td>0.023</td>
<td>Combines and doubles the effects of Brain Booster + Reducer + Simplifier</td>
</tr>
</tbody>
</table>

Figure 2: Infinitus Algebraica Gameplay.

Completion. In line with [3], we used an achievement system for Completion. We developed 48 achievements with 16 one-off, 24 progressive, and 8 chained missions. For completing a mission, the players received power-ups, skins, and music.

Novelty. For Novelty, we implemented an (RPG-like) level-up system with 40 ranks. Based on the player’s performance and difficulty, they received experience points. At milestone levels, the player received skins, music, new levels of difficulty, and power-ups.

Blocking. In line with [3], we restricted the playtime to 15 minutes every 8 hours.

Waiting. The power-ups were generated over time. Every 8 hours the game selected and spawned power-ups. The probability of each power-up was based on the strength of the item (Table 1), with weaker items generated more frequently. For the remaining time until regeneration, we displayed a countdown and a notification inside the game.

Gameplay videos of the individual game versions are available at https://osf.io/9q6ve/files/.
4 PRELIMINARY SHORT-TERM COMPARISON

To evaluate the PX and the consistency of the six game versions of our math game, we conducted a single-session short-term study on Amazon Mechanical Turk (MTurk). The study design and the measurements closely resembled Alexandrovsky et al.’s [3] pre-study. The demographics were accessed using the Ten-Item Personality Inventory (TIPI) [18] (Cronbach’s α= 0.40 – 0.68), which determines personality using the FFM [41]. Additionally, we asked questions targeting participants’ attitudes and experiences with math. We measured motivation using the Situational Motivation Scale (SIMS) [42], which assesses motivation on the subscales Intrinsic Motivation (Cronbach’s α=0.95), Identified Regulation (Cronbach’s α=0.80), External Regulation (Cronbach’s α=0.86), and Amotivation (Cronbach’s α=0.77) and the Intrinsic Motivation Inventory (IMI) [91] with the subscales Interest-Enjoyment, Competence, Effort-Importance and Tension-Pressure. The overall reliability of IMI is Cronbach’s α=0.85 [71]. For performance metrics, we logged: the score, unlocked levels, precision, play/session duration, and play counts.

The procedure was as follows: (i) download the game from the Google Playstore (ii) registration and random assignment to one of the 6 game versions (iii) demographics questionnaires, TIPI (iv) the participants played a maximum of 6 trials to complete the first 2 packages (10 levels each, each package took approximately 4 minutes) (v) the participants filled out the SIMS and IMI questionnaires (vi) qualitative feedback (vii) the participants received their MTurk code. The study took 25 – 30 minutes with a playtime of M=15.59 (SD=9.17) minutes. The participants received $3.

4.1 Pre-Study Participants and Group Comparison

Table 2: Infinitus Algebraica Pre-study: ANOVA results of demographics and motivation between the groups (n=49), M, and SE indicate the sample means and standard errors. For TIPI and SIMS the univariate results of the conducted MANOVAs are presented for individual subscales.

<table>
<thead>
<tr>
<th>Demographics</th>
<th>M</th>
<th>SE</th>
<th>F_{4,43}</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification as a gamer</td>
<td>31.57</td>
<td>1.22</td>
<td>1.24</td>
<td>0.31</td>
</tr>
<tr>
<td>(1) Not at all</td>
<td>6.99</td>
<td>0.38</td>
<td>0.14</td>
<td>0.98</td>
</tr>
<tr>
<td>Enjoy mental math</td>
<td>5.57</td>
<td>0.19</td>
<td>1.05</td>
<td>0.43</td>
</tr>
<tr>
<td>Using maths in daily life</td>
<td>5.93</td>
<td>0.04</td>
<td>1.29</td>
<td>0.29</td>
</tr>
<tr>
<td>Exp. with serious games</td>
<td>6.33</td>
<td>0.04</td>
<td>1.41</td>
<td>0.24</td>
</tr>
<tr>
<td>Extraversion (1-7)</td>
<td>4.04</td>
<td>0.20</td>
<td>0.67</td>
<td>0.65</td>
</tr>
<tr>
<td>Agreeableness (1-7)</td>
<td>4.88</td>
<td>0.18</td>
<td>0.11</td>
<td>0.99</td>
</tr>
<tr>
<td>Conscientiousness (1-7)</td>
<td>5.21</td>
<td>0.16</td>
<td>0.67</td>
<td>0.64</td>
</tr>
<tr>
<td>Emotional Stability (1-7)</td>
<td>4.96</td>
<td>0.17</td>
<td>1.18</td>
<td>0.33</td>
</tr>
<tr>
<td>Openness to Experiences (1-7)</td>
<td>4.92</td>
<td>0.15</td>
<td>0.38</td>
<td>0.86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TIPI</th>
<th>M</th>
<th>SE</th>
<th>F_{4,43}</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic Motivation (1-7)</td>
<td>5.29</td>
<td>0.21</td>
<td>0.35</td>
<td>0.74</td>
</tr>
<tr>
<td>Identified Regulation (1-7)</td>
<td>5.22</td>
<td>0.19</td>
<td>0.93</td>
<td>0.47</td>
</tr>
<tr>
<td>External Regulation (1-7)</td>
<td>4.04</td>
<td>0.18</td>
<td>0.84</td>
<td>0.53</td>
</tr>
<tr>
<td>Amotivation (1-7)</td>
<td>3.04</td>
<td>0.22</td>
<td>0.19</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Our sample consisted n=49 (12 women, 37 men) after filtering 7 participants who did not play or had zero variance in their responses. The exit questionnaire was filled out by n=32 participants. The demographics are summarized in Table 2. To evaluate if demographics were balanced between the groups, we conducted one-way MANOVAs for TIPI and SIMS and one-way ANOVAs for the demographics. The MANOVAs were chosen because they are more robust as they account for the shared variance of multiple dependent measures. The results showed no significant differences in any of the participant characteristics between groups (see Table 2). The p-values were not corrected for multiple comparisons as it would reduce the chance of finding possible imbalances between the conditions.

4.2 Pre-Study Results

Motivation, Player Experience. A one-way MANOVA of motivation on IMI showed no significant differences between the conditions ($F_{4,43}=0.95, p=0.662, \eta^2_p=0.17$). The descriptive statistics as well as the results of the ANOVAs for each subscale on IMI are depicted in Table 3.

Performance. We conducted a one-way MANOVA on the performance metrics between the conditions ($F_{4,43}=0.95, p=0.430, \eta^2_p=0.07$). The subsequent ANOVAs show only differences on number of finished packages ($F_{4,43}=4.88, p<0.01, \eta^2_p=0.362$). Post-hoc t-tests revealed that participants in Blocking finished fewer packages than in all other conditions. The descriptive statistics and ANOVA results of the performance metrics are shown in Table 4.

**Qualitative Feedback.** In the last step of the study procedure, the participants had the option to give comments on the game. The exit questionnaire consisted of four open questions that asked what participants liked (Q1) or disliked (Q2) about the game, what prevented them from playing (Q3), and if they had additional remarks (Q4). The responses were clustered into broad topics. 14/32 participants commented that the game trained mental math skills ("It was very challenging and satisfying. It made me to love mathematics") and 14/32 stated explicitly they enjoyed playing ("I am very like to play this game this is really amazing."). 11/32 gave negative feedback regarding short time limit (7×), monotonous gameplay (3×), and missing instructions (2×). Further, we received some comments on minor software implementation bugs.

4.3 Pre-Study Discussion

In preparation for the long-term evaluation of the snacking framework in a serious game, we designed and evaluated a math-learning serious game for mobile devices. We conducted a short-term user study on MTurk to evaluate the PX of the game and to validate the comparability between the game versions.

The analysis of the demographic data shows no significant differences in the personality factors or incoming motivation between the conditions. This confirms that the random assignment was successful and that the performance and PX measurements are not biased by demographic factors. The IMI results assured that the game versions provide a comparable short-term PX. All game versions were perceived as enjoyable, but not too immersive or tense.
Table 3: Pre-study Infinitus Algebraica: test statistics for the intrinsic motivation on IMI.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Blocking</th>
<th>Novelty</th>
<th>Waiting</th>
<th>Rewards</th>
<th>Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F(5,13)</td>
<td>p</td>
<td>M(SD)</td>
<td>M(SD)</td>
<td>M(SD)</td>
<td>M(SD)</td>
</tr>
<tr>
<td>Interest-Enjoyment</td>
<td>0.27</td>
<td>0.93</td>
<td>5.46</td>
<td>5.69</td>
<td>5.77</td>
<td>5.38</td>
</tr>
<tr>
<td>Competence</td>
<td>0.11</td>
<td>0.99</td>
<td>5.76</td>
<td>5.70</td>
<td>6.00</td>
<td>5.80</td>
</tr>
<tr>
<td>Effort-Importance</td>
<td>2.37</td>
<td>0.06</td>
<td>4.23</td>
<td>4.62</td>
<td>4.86</td>
<td>5.96</td>
</tr>
<tr>
<td>Tension-Pressure</td>
<td>1.28</td>
<td>0.23</td>
<td>3.52</td>
<td>4.25</td>
<td>3.43</td>
<td>3.22</td>
</tr>
</tbody>
</table>

Table 4: Infinitus Algebraica Pre-study: ANOVA results of performance metrics between the groups. M, and SE indicate the sample means and standard errors.

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>SE</th>
<th>F(5,13)</th>
<th>p-bonf</th>
<th>p</th>
<th>n²</th>
</tr>
</thead>
<tbody>
<tr>
<td>unlocked packages</td>
<td>3.53</td>
<td>0.170</td>
<td>0.335</td>
<td>1.000</td>
<td>0.038</td>
<td></td>
</tr>
<tr>
<td># play rounds</td>
<td>8.510</td>
<td>0.596</td>
<td>0.758</td>
<td>1.000</td>
<td>0.081</td>
<td></td>
</tr>
<tr>
<td># fails</td>
<td>25.878</td>
<td>4.466</td>
<td>0.205</td>
<td>1.000</td>
<td>0.023</td>
<td></td>
</tr>
<tr>
<td># wins</td>
<td>135.469</td>
<td>6.820</td>
<td>0.366</td>
<td>1.000</td>
<td>0.041</td>
<td></td>
</tr>
<tr>
<td>best combo</td>
<td>19.469</td>
<td>0.491</td>
<td>2.249</td>
<td>0.932</td>
<td>0.207</td>
<td></td>
</tr>
<tr>
<td>mean score</td>
<td>154.469</td>
<td>5.099</td>
<td>0.658</td>
<td>1.000</td>
<td>0.071</td>
<td></td>
</tr>
<tr>
<td># levels finished</td>
<td>63.367</td>
<td>3.022</td>
<td>0.425</td>
<td>1.000</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td># items</td>
<td>0.204</td>
<td>0.082</td>
<td>1.337</td>
<td>0.000</td>
<td>0.135</td>
<td></td>
</tr>
<tr>
<td># skins</td>
<td>1.061</td>
<td>0.035</td>
<td>0.530</td>
<td>0.000</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td># music</td>
<td>1.020</td>
<td>0.020</td>
<td>1.028</td>
<td>0.000</td>
<td>0.107</td>
<td></td>
</tr>
<tr>
<td>duration in min.</td>
<td>15.585</td>
<td>1.310</td>
<td>0.711</td>
<td>0.000</td>
<td>0.076</td>
<td></td>
</tr>
<tr>
<td>accuracy</td>
<td>78.755</td>
<td>3.919</td>
<td>0.449</td>
<td>0.000</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td># retries</td>
<td>0.592</td>
<td>0.351</td>
<td>0.725</td>
<td>0.000</td>
<td>0.078</td>
<td></td>
</tr>
<tr>
<td># finished pack</td>
<td>0.714</td>
<td>0.065</td>
<td>4.884</td>
<td>0.018</td>
<td>0.362</td>
<td></td>
</tr>
</tbody>
</table>

and the players were investing effort. Also, the performance analysis showed that all game versions were similar in terms of difficulty and progression. Moreover, the comments from participants confirmed that the game was viewed as a serious game. These results confirm that the game meets our requirements for the long-term study of an enjoyable, easy-to-learn serious game with game versions based on individual snacking mechanics that are comparable in PX, playability, and difficulty.

5 LONG-TERM EVALUATION OF THE SNACKING MECHANICS IN A SERIOUS GAME

After validating the PX and establishing the comparability between the versions of Infinitus Algebraica, we designed a 3-week long-term behavioural experiment. The study was distributed on mobile devices over the course of 2 months from September to October 2019, in which the participants played in uncontrolled conditions (i.e., a field study).

5.1 Study Design and Methods

The study design closely mirrored Alexandrovsky et al.’s [3] long-term study with the GOC. Based on the comments from our pre-study, we modified Infinitus Algebraica by adding additional instruction, adding a tutorial, fixing minor bugs, and tweaking the game balancing. In the pre-study, the average playtime showed a high variance of around 15 minutes. Also, by considering the lessons-learned from [3], we reduced the playtime in the Blocking condition to 10 minutes per 8 hours to enforce its effect. We released Infinitus Algebraica over the Google Play Store (01.09.2019) and the Apple App Store (27.09.2019). We advertised the study on social networks, flyers on the local campus, and through word-of-mouth recommendations. The participation was voluntary. The course of the study was as follows:

Day 0: The participants downloaded and installed the game. Next, they read the information about the study and were asked to fill out a consent form. If confirmed, the participants were randomly assigned to one of the 6 conditions and received the TIP. After filling out the demographics, the app loaded the game interface and the participants could start exploring the game. After playing the first 2 rounds (ca. 8 minutes), the participants were required to fill out the IMI and SIMS questionnaires. After that, they were freely allowed to play (except in the Blocking condition, which limited their access).

Days 7, 14 and 21: Every seven days, the participants were asked to fill out the IMI and SIMS questionnaires. On the final day, after 3 weeks, the participants additionally filled out the exit survey.

As we received multiple comments that the players were annoyed by the questionnaires and stopped playing due to the questionnaires and not to the game’s design, we disabled the weeks 1, 2, and 3 questionnaires for all participants who started after 12.10.2019 in favor of clearer behavioural log data. Thus, our data consists only of self-reports from Android devices in the first half of September 2019. However, we are certain that the subjective responses are representative for the sample since only 9 out of 99 subjects participated in the study without answering the questionnaires.

5.2 Participants

149 players downloaded the game. 99 of them played Infinitus Algebraica at least once (Android: n=83; iOS: n=16). The number of participants decreased over the course of 3 weeks. By the end of week 2, the sample consisted of 20 participants and only 9 participants played for the full 3 weeks. Therefore, planned repeated measure analyses on how the motivation evolved over the course of the study could not be performed. Analogously to the pre-study for the questionnaire analysis, we excluded 13 participants who played less than one complete round. 11 subjects were removed, because their responses showed no variance, contradicting answers, or wrong responses to control questions. After filtering, 76 participants (40 self-identified as women) completed the demographics questionnaire (N: Baseline=14, Blocking=15, Completion=13, Rewards=12, Novelty=10, Waiting=12). The SIMS was only filled out by 75 participants. On average, the participants were M=28.88 (SD=10.18)
years old. There were no significant differences between the conditions on any of the demographic factors. Demographic details are depicted in Table 5.

6 LONG-TERM RESULTS

We present the results of three categories of data. We first compare the incoming motivation and player traits to establish comparability of the conditions. Next, we report the group differences from performance and behavioural (temporal) metrics. Finally, we report the results of a survival analysis to measure adherence, which is based on the players’ performances and behaviours.

6.1 Motivation and Player Experience

We conducted one-way MANOVAs with the game versions as a between-participants factor on the subscales of incoming (i.e., week 0) IMAs (F[3,59]=0.39, p=0.79, η²p=0.03) and SIMIs (F[3,69]=1.19, p=0.26, η²p=0.08) and found no significant differences between the conditions. Descriptive plots of the motivation on the first day are depicted in Figure 3. The overall rating for motivation was generally positive. The results of the MANOVA are shown in Tables 6 and 7.

6.2 Performance and Behaviour

For the analysis of performance and behaviour, our sample consists of 99 participants (Baseline: n=18; Blocking: n=17; Completion: n=15; Novelty: n=15; Rewards: n=19 and Waiting: n=15) who played 512 sessions in total. The play sessions are distributed across the conditions as follows: Baseline = 91; Blocking = 47; Completion = 169; Novelty = 47; Rewards = 88; Waiting = 70. The play times are depicted in Figure 4. On the performance metrics, we conducted one-way ANOVAs which revealed multiple significant differences and generally showing that in Blocking and Completion, the players progressed slower through the game. The detailed performance and behaviour analyses are provided in the Appendix, as they are in support of our main research questions, but not central to them.

6.3 Attrition

To investigate how the different conditions affect the player’s adherence, we conducted a survival analysis. Survival analysis is a group of statistical methods that model the time span until a particular event occurs (e.g., How long do patients survive using a particular treatment? How long will a machine last?). Generally, the models expect a time series of events as inputs and compute survival functions—probability functions for events to occur at a given point in time—or hazard functions that describe the chance for a state transition over time. These functions then can be used to predict future occurrences of events.

We grouped the play sessions of each player into bins with a duration of the average inter-session interval (33.75 hours) leading to a total of 32 bins for the duration of the study. We applied a Kaplan-Meier Estimator (KME) as an estimator of survival on the binned play sessions. The KME is a non-parametric estimator where the survival function is described as “the product of one minus the exit rate at each of the survival times” [51]. The KME does not require regular time intervals and is meant to deal with right-skewed data [97, 118]. Therefore, it allows us to include participants into the analysis who were not playing in the intermediate bins but came back later. Figure 5 shows the predicted survival functions for each condition and the corresponding median survival times are depicted in Table 8. Notably, the survival probability in the first bin is not 1. This is because some participants dropped out and stopped playing within the first 33 hours. To determine if the survival distributions differ, we conducted a Mantel-Cox Log-rank test [67]. The null hypothesis of the test assumes that the distributions of the survival functions are equal. The survival functions of the game versions were significantly different, χ²=4.361, p=0.0037. To investigate how the survival of the individual game versions differs, we conducted three types of pairwise Log-rank comparisons between the conditions. Log-Rank (Mantel-Cox) Test: weights each point in time equally based on median survival time. Breslow (Generalized Wilcoxon): gives more weight to the data points near time T. Tarone-Ware: gives more weight to the data points near the end of the observed time period. The results of the pairwise comparisons are depicted in Table 9.

To investigate how the players’ performance affected their retention, we applied the Cox Proportional Hazard (CPH) survival estimator. In contrast to KME which is a univariate model that can only deal with categorical variables, the CPH is a multivariate regression model that accounts for continuous and categorical factors. The model calculates the hazard ratios between groups. Hazard ratios <1 indicate a longer survival of a group. Hazard ratios >1 suggest a higher chance of an event to occur. The Cox model does not have any assumptions regarding the survival function, but it assumes that the hazard functions are proportional [46]. The predictive accuracy of CPH is commonly described by the concordance statistic which is defined as the proportion of all usable patient pairs in which the predictions and outcomes are concordant.
Figure 3: Infinitus Algebraica Long-term: Motivation assessment on the first day. The whiskers indicate SEMs. The dotted line indicates the midpoint.

Table 6: Long term Infinitus Algebraica: MANOVA for Intrinsic Motivation on Day 0.

<table>
<thead>
<tr>
<th></th>
<th>(F_{5,70})</th>
<th>(p)</th>
<th>(\eta^2_p)</th>
<th>Baseline M (SD)</th>
<th>Novelty M (SD)</th>
<th>Waiting M (SD)</th>
<th>Rewards M (SD)</th>
<th>Completion M (SD)</th>
<th>Blocking M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest-Enjoyment (1-7)</td>
<td>0.23</td>
<td>0.95</td>
<td>0.02</td>
<td>4.71 (0.83)</td>
<td>4.64 (0.71)</td>
<td>4.43 (0.75)</td>
<td>4.75 (0.68)</td>
<td>4.65 (0.94)</td>
<td>4.67 (0.80)</td>
</tr>
<tr>
<td>Competence (1-7)</td>
<td>0.57</td>
<td>0.72</td>
<td>0.04</td>
<td>4.33 (0.87)</td>
<td>4.74 (0.98)</td>
<td>4.65 (1.06)</td>
<td>4.90 (1.08)</td>
<td>4.82 (1.12)</td>
<td>4.61 (0.63)</td>
</tr>
<tr>
<td>Effort-Importance (1-7)</td>
<td>0.48</td>
<td>0.79</td>
<td>0.03</td>
<td>3.57 (1.10)</td>
<td>3.80 (1.01)</td>
<td>3.22 (1.13)</td>
<td>3.86 (1.16)</td>
<td>3.46 (1.16)</td>
<td>3.49 (1.31)</td>
</tr>
<tr>
<td>Tension-Pressure (1-7)</td>
<td>0.25</td>
<td>0.94</td>
<td>0.02</td>
<td>3.64 (0.90)</td>
<td>3.63 (1.42)</td>
<td>3.60 (1.31)</td>
<td>3.42 (0.86)</td>
<td>3.21 (1.05)</td>
<td>3.52 (1.44)</td>
</tr>
</tbody>
</table>

Table 7: Long term Infinitus Algebraica: MANOVA for Situated Motivation on Day 0.

<table>
<thead>
<tr>
<th></th>
<th>(F_{5,69})</th>
<th>(p)</th>
<th>(\eta^2_p)</th>
<th>Baseline M (SD)</th>
<th>Novelty M (SD)</th>
<th>Waiting M (SD)</th>
<th>Rewards M (SD)</th>
<th>Completion M (SD)</th>
<th>Blocking M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic Motivation (1-7)</td>
<td>1.68</td>
<td>0.15</td>
<td>0.11</td>
<td>4.30 (0.91)</td>
<td>4.30 (1.10)</td>
<td>3.71 (1.17)</td>
<td>4.77 (0.97)</td>
<td>4.23 (0.98)</td>
<td>3.83 (1.28)</td>
</tr>
<tr>
<td>Identified Regulation (1-7)</td>
<td>0.73</td>
<td>0.60</td>
<td>0.05</td>
<td>4.57 (0.86)</td>
<td>3.88 (1.43)</td>
<td>4.10 (0.95)</td>
<td>4.18 (1.14)</td>
<td>4.13 (0.91)</td>
<td>3.97 (0.87)</td>
</tr>
<tr>
<td>External Regulation (1-7)</td>
<td>0.99</td>
<td>0.43</td>
<td>0.07</td>
<td>2.34 (1.22)</td>
<td>2.77 (1.36)</td>
<td>2.75 (1.57)</td>
<td>2.05 (0.69)</td>
<td>2.00 (0.61)</td>
<td>2.60 (1.33)</td>
</tr>
<tr>
<td>Amotivation (1-7)</td>
<td>0.85</td>
<td>0.52</td>
<td>0.06</td>
<td>2.57 (1.16)</td>
<td>2.65 (1.13)</td>
<td>2.65 (0.99)</td>
<td>1.98 (0.98)</td>
<td>2.23 (0.84)</td>
<td>2.57 (1.16)</td>
</tr>
</tbody>
</table>

Table 8: Long-term Infinitus Algebraica: Median survival predicted by the Kaplan-Meier Fitter

<table>
<thead>
<tr>
<th>Condition</th>
<th>Median Survival Time</th>
<th>SE</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.0</td>
<td>0.00</td>
<td>[0.00, 1.00]</td>
</tr>
<tr>
<td>Rewards</td>
<td>2.0</td>
<td>0.90</td>
<td>[0.24, 3.76]</td>
</tr>
<tr>
<td>Novelty</td>
<td>0.0</td>
<td>0.00</td>
<td>[0.00, 0.00]</td>
</tr>
<tr>
<td>Completion</td>
<td>5.0</td>
<td>0.93</td>
<td>[3.17, 6.83]</td>
</tr>
<tr>
<td>Waiting</td>
<td>1.0</td>
<td>0.56</td>
<td>[0.00, 2.17]</td>
</tr>
<tr>
<td>Blocking</td>
<td>4.0</td>
<td>0.88</td>
<td>[2.28, 5.71]</td>
</tr>
<tr>
<td>Overall</td>
<td>3.0</td>
<td>0.40</td>
<td>[2.23, 3.78]</td>
</tr>
</tbody>
</table>

7 DISCUSSION

RQ1: Do the snacking mechanics yield different player experiences when implemented in a serious game? The responses on demographic measures show that the conditions were balanced. The
low scores on External Regulation and Amotivation indicate voluntary participation and that the field study was ecologically valid. The subjective measurements on PX and motivation confirm the results from the pre-study and reassure that the game was enjoyable but not tense. The self-reported intrinsic motivation was almost unaffected over the course of one week.

RQ2: Can the snacking mechanics provoke a snacking behavioural pattern of game engagement when deployed in an uncontrolled field study? Our results show that every mechanic contributed to higher engagement compared to Baseline. This is supported by the higher number of sessions in the snacking conditions and is further underlined by the survival analyses. However, the conditions facilitated different player behaviours. In Novelty, Waiting and Rewards, the participants played more intensively in the first days and dropped off quickly. This is supported by longer play-rounds in those conditions and the lower median survival time.

RQ3: Which snacking mechanic facilitates the strongest adherence over time? Completion yielded by far the longest participation with the game. While in most other conditions the participants stopped playing after around 10 days, the participants in the Completion condition engaged with the game for around 20 to 25 days. Supported by the comparable demographics and incoming motivation as well as the non-significant results on most performance metrics and the significant negative coefficient for Condition in the CPH estimation, these results support the inference that sustained engagement can be attributed to the different game mechanics in the different conditions. Therefore, we argue that the snacking mechanics may facilitate sustained engagement and can provoke the snacking pattern of play. These results support findings from the literature that suggests clear goals and meaning as successful strategies for educational games [37, 65, 69]. In contrast to Alexandrovsky et al., the present study implemented almost twice as many missions and presented the player with different options and goals, but not tense. The self-reported intrinsic motivation was almost unaffected over the course of one week.

Table 9: Long-term Infinitus Algebraica: Pairwise Log-Rank comparisons of the survival analysis. The bold p-values indicate significant differences. Sample sizes for the different groups: Baseline: n=18; Blocking: n=17; Completion: n=15; Novelty: n=15; Rewards: n=19; Waiting: n=15. Superscript 1: Mantel-Cox Log-Rank Test: weights each point in time equally; based on median survival time. Superscript 2: Breslow (Generalized Wilcoxon) gives more weight to the datapoints near the end of the time. Superscript 3: Tarone-Ware Test gives more weight to the datapoints near the end of the time.

Table 10: Infinitus Algebraica, Long-term: Statistics and estimates of the Cox-regression.
that the increase in mission variety coupled with the overall feeling of accomplishment after completing a mission helped satisfy the needs for autonomy and relatedness which have been shown numerous times to foster player motivation [7]. Interestingly, Novelty differs drastically from the results by Alexandrovsky et al. and shows almost no influence on the sustained engagement. This can be explained by a lower connection of the novel elements with the game content and with the seriousness of the game. As predicted in Alexandrovsky et al.’s study, in Blocking the participants played significantly more sessions compared to Baseline and at the same time Interest-Enjoyment was rated lower. All conditions yield a snacking behaviour with play sessions of around 20 to 30 minutes and regular breaks of around 1.5 days. Together, the number of rounds played and the (pure) game-play time the results show that Waiting was played more extensively.

7.1 Comparison with Prior Work
The present paper draws inference from a pre-study and a long-term study which are comparable in terms of game versions and overall study design to [3]. In conjunction, this allows us to draw a deeper understanding while acknowledging several important differences. First, while both studies used game-like features, in contrast to Infinitus Algebraica, Games of Cannons (GOC) was not a serious game. Second, the motivational pull of the two games might have been somewhat different. The reported enjoyment of the serious game was higher, which is not surprising as we iterated on the game design elements implemented in GOC. Furthermore, players might have recognized that the mathematical knowledge gained in Infinitus Algebraica can be helpful in their lives. Third, the context in which data was collected differed between studies. While GOC was a browser-based game that had to be played on a desktop PC and included a study conducted on MTurk with paid participants and daily notification reminders, Infinitus Algebraica is a mobile app that can be accessed anywhere, at any time, and participation was voluntary. Thus, the ease of access for both games differed as well as the (environmental) conditions under which the study participants played the games. However, even given these differences, the game versions were developed using the same snacking framework, and the remaining study design details were equivalent. In both studies, there was a significant improvement in player experience, performance, and playtime when any one of the five snacking game design elements was implemented. It is important to note that the impact of the experimental manipulation (i.e., the implementation of Blocking, Completion, Rewards, Novelty, and Waiting mechanics) could be replicated across context, platforms, and samples. Even though the individual contributions of design elements may vary, this supports our conclusion that the five game design elements outlined above can contribute to eliciting a snacking pattern of engagement with the game. Importantly, the design of the present study allows us to confidently draw strong conclusions about the efficacy of game design elements. Although most previous studies implicitly assume a causal relation between attrition and some variable (e.g., game performance), they cannot draw strong conclusions about the connection of the variables because of their correlative approach. The present study overcomes this gap by experimentally manipulating individual game design elements. This has the advantage, that given everything else is held constant and confounds are controlled for, any observed changes can be causally traced back to the experimental manipulation.

7.2 Limitations and Future Work
A limitation of the present study is the low sample size, which does not allow us to draw final conclusions from the results. A larger sample would have provided a more nuanced picture of player attrition as longer time intervals could have been analyzed. Further, it would have reduced potential sampling biases. Also, the demographic traits assessed the biological sex and not gender, as suggested by the "HCI Guidelines for Gender Equity and
Inclusivity” [95]. Therefore, the present results should be interpreted with caution and require further validation. Nevertheless, the significant differences of the attrition analysis show medium to large effect sizes (W=0.38–0.77) with a sufficient statistical power 1−β=0.75–0.99. Although these results do not allow to draw final conclusions, they help to build sound hypotheses that require further validation in future work. As the present study focused on the contribution and effect of a single experimental manipulation, future research should look at the potential additive effects of the five game mechanics if deployed in combination. We hypothesize that the combination of several mechanisms would lead to an even stronger motivational pull and improved player experience. Also, future work should expand beyond the snacking framework and also investigate how other game elements affect player attrition. This research focused on the transferability of the snacking mechanics into the context of serious games under volitional play in a field study; however, the mechanisms were designed rather superficially. Future research should investigate which snacking mechanics individually in-depth and operationalize its parameterization. For example, controlling the duration of Blocking and Waiting, choosing the right time to present Novelty elements or Missions, and adjusting the amount of Rewards based on players’ behaviours. Moreover, the snacking mechanics provide a fruitful foundation to implement DDA in the game. For instance, the amount of rewards or the timing of novel elements could be designed based on player performance and behaviour to further combat player attrition. So far, the snacking mechanics have been explored only in single-player contexts. Future research should address multi-player influences as it has been shown numerous times that social play can increase motivation and sustained participation [59, 92, 93]. It is plausible that an added social component might further increase motivation; however, it was important to first establish and replicate the role of the snacking mechanics in adherence in isolation and in a single-player context.

7.3 Implications for Game Design

The snacking style of gameplay is arguably important in several different scenarios, ranging from free-to-play games that rely on returning players for a profit or serious games that try to assist a player to positively change their behaviour. Our results suggest that using the five game mechanics, designers can tailor game designs to the style of game and player, and adjust it for specific serious purposes. Furthermore, the provoked behaviour is not entirely relying on motivation, and in the case of Blocking, player engagement can be sustained even when the intrinsic motivation decays. Further, the development costs of the snacking mechanics differ. While Waiting and Blocking can be easily implemented into existing game designs at almost no cost, Completion and Novelty require significantly more effort as their efficacy scales with the size of the game and amount of new content, and in most cases—except for when using procedural content creation [88]—they need to be developed manually.

It is important to note that game design with the aim of altering a player’s behaviour is valuable but also dangerous. On the one hand, these methods can foster engagement with serious content and improve the players’ well-being or the development of skills. On the other hand, these methods can be exploited unethically as “dark” design patterns [117], for example for monetization strategies [4], and force players into unintended behaviour such as addiction [115] with consequences that extend beyond the boundaries of the “magic circle” [34].

8 CONCLUSIONS

Many serious games that are designed to promote learning, habit formation, or behaviour change are most effective when played regularly and frequently in shorter bursts—known as a ‘snacking’ pattern of interaction. In the related context of health applications, Eysenbach [26] described “the law of attrition”, contending that most users drop off before completing the intended course of use, and he further argues that although high dropout rates might be expected, that “determinants of attrition should be highlighted, measured, analyzed, and discussed”.

However, in the context of serious game design, little has been known about how game elements and game design features support player adherence over time. In this work, we drew from existing frameworks and game design theories as well as from the design of casual games to investigate how five individual game mechanic groups (Rewards, Novelty, Completion, Waiting, and Blocking), affect player attrition in a serious math learning game. We compared the five game versions (plus a Baseline version) initially in a short-term study to show comparability in game experience. In a long-term study, the six versions of the game were available in common app stores for download onto participants’ own mobile phones for play in a 3-week field study. We gathered incoming self-reported player motivation, personality, and play experience, and logged the players’ performance as well as their behaviour through system use.

Our findings suggest that the groups were comparable in terms of incoming personality and motivation and that their experience of the first round of play was similar. However, using survival analysis, we identified that the addition of the Completion and Blocking mechanics facilitated the strongest sustained engagement over the weeks. While these results are not conclusive, they provide initial evidence for the snacking mechanics and allow researchers to derive sound hypotheses for future research. We demonstrate a methodological approach for evaluating the “determinants of attrition” [26] in the context of serious game design. Our findings are congruent with existing theories of player experience and our results promote the development of guidelines on designing for sustained engagement in serious games designed for learning.

OPEN ACCESS STATEMENT

The raw data along with the analyses is available at https://osf.io/9q6ve/.

REFERENCES

Evaluating Player Experience in Games at FDG '11. ACM, Bordeaux, France, 6.


[113] [4471-4769-5_7]

APPENDIX

A INFINITUS ALGEBRAICA

Table 11 illustrates how the different parameters were applied for the difficulty progression throughout the 35 packages of the game.

B LONG-TERM: PERFORMANCE ANALYSIS

For performance metrics, our sample consists of 99 participants who played 512 sessions in total. The majority of our performance variables violated the homogeneity and sphericity assumptions, therefore, for all performance variables, we conducted one-way ANOVA analyses.
Table 11: Difficulty progression of the game for the play mode normal.

<table>
<thead>
<tr>
<th>Package ID</th>
<th>Grid Size</th>
<th>Operations</th>
<th>M (SD) Level Duration in sec.</th>
<th>Values Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2 x 2, 3 x 3, 4 x 4</td>
<td>+</td>
<td>94.00 (3.16)</td>
<td>[0-15]</td>
</tr>
<tr>
<td>1</td>
<td>3 x 3</td>
<td>+</td>
<td>59.80 (1.87)</td>
<td>[0-23]</td>
</tr>
<tr>
<td>2</td>
<td>3 x 3</td>
<td>+</td>
<td>36.10 (4.91)</td>
<td>[0-25]</td>
</tr>
<tr>
<td>3</td>
<td>3 x 3, 4 x 4</td>
<td>–</td>
<td>39.00 (4.40)</td>
<td>[0-30]</td>
</tr>
<tr>
<td>4</td>
<td>3 x 3</td>
<td>–</td>
<td>37.00 (1.25)</td>
<td>[0-23]</td>
</tr>
<tr>
<td>5</td>
<td>3 x 3</td>
<td>x</td>
<td>36.70 (2.96)</td>
<td>[0-12]</td>
</tr>
<tr>
<td>6</td>
<td>3 x 3, 4 x 4</td>
<td>÷</td>
<td>35.20 (2.20)</td>
<td>[0-21]</td>
</tr>
<tr>
<td>7</td>
<td>3 x 3</td>
<td>–</td>
<td>34.50 (0.85)</td>
<td>[0-16]</td>
</tr>
<tr>
<td>8</td>
<td>3 x 3, 4 x 4</td>
<td>÷, x, –</td>
<td>37.50 (2.72)</td>
<td>[0-22]</td>
</tr>
<tr>
<td>9</td>
<td>3 x 3</td>
<td>mod</td>
<td>33.60 (5.46)</td>
<td>[0-19]</td>
</tr>
<tr>
<td>10</td>
<td>3 x 3</td>
<td>x⁴</td>
<td>39.30 (4.06)</td>
<td>[0-20]</td>
</tr>
<tr>
<td>11</td>
<td>4 x 4</td>
<td>+</td>
<td>39.30 (1.25)</td>
<td>[0-50]</td>
</tr>
<tr>
<td>12</td>
<td>4 x 4</td>
<td>–</td>
<td>48.80 (0.92)</td>
<td>[0-50]</td>
</tr>
<tr>
<td>13</td>
<td>4 x 4</td>
<td>x</td>
<td>41.80 (0.42)</td>
<td>[0-25]</td>
</tr>
<tr>
<td>14</td>
<td>4 x 4</td>
<td>÷</td>
<td>42.50 (1.27)</td>
<td>[0-32]</td>
</tr>
<tr>
<td>15</td>
<td>4 x 4</td>
<td>mod</td>
<td>52.50 (1.51)</td>
<td>[0-23]</td>
</tr>
<tr>
<td>16</td>
<td>4 x 4</td>
<td>x⁴</td>
<td>47.80 (2.04)</td>
<td>[0-60]</td>
</tr>
<tr>
<td>17</td>
<td>4 x 4</td>
<td>x⁴, mod, ÷, x, –</td>
<td>74.70 (1.34)</td>
<td>[0-23]</td>
</tr>
<tr>
<td>18</td>
<td>4 x 4</td>
<td>x⁴, mod, ÷</td>
<td>68.90 (1.52)</td>
<td>[0-23]</td>
</tr>
<tr>
<td>19</td>
<td>5 x 5</td>
<td>+</td>
<td>65.40 (3.60)</td>
<td>[0-90]</td>
</tr>
<tr>
<td>20</td>
<td>5 x 5</td>
<td>–</td>
<td>94.40 (3.60)</td>
<td>[0-53]</td>
</tr>
<tr>
<td>21</td>
<td>5 x 5</td>
<td>x</td>
<td>86.80 (1.40)</td>
<td>[0-28]</td>
</tr>
<tr>
<td>22</td>
<td>5 x 5</td>
<td>x⁴</td>
<td>92.10 (0.99)</td>
<td>[0-25]</td>
</tr>
<tr>
<td>23</td>
<td>5 x 5, 6 x 6</td>
<td>mod</td>
<td>122.20 (2.82)</td>
<td>[0-35]</td>
</tr>
<tr>
<td>24</td>
<td>5 x 5, 6 x 6</td>
<td>x, –</td>
<td>117.40 (2.62)</td>
<td>[0-45]</td>
</tr>
<tr>
<td>25</td>
<td>5 x 5, 6 x 6</td>
<td>mod, ÷, x, –</td>
<td>121.90 (2.38)</td>
<td>[0-45]</td>
</tr>
<tr>
<td>26</td>
<td>6 x 6, 5 x 5</td>
<td>x⁴, mod, ÷, x, –</td>
<td>124.10 (3.38)</td>
<td>[0-45]</td>
</tr>
<tr>
<td>27</td>
<td>6 x 6, 5 x 5</td>
<td>+</td>
<td>129.10 (1.97)</td>
<td>[0-200]</td>
</tr>
<tr>
<td>28</td>
<td>6 x 6, 5 x 5</td>
<td>–</td>
<td>129.10 (1.97)</td>
<td>[0-200]</td>
</tr>
<tr>
<td>29</td>
<td>6 x 6, 5 x 5</td>
<td>x</td>
<td>144.10 (1.97)</td>
<td>[0-105]</td>
</tr>
<tr>
<td>30</td>
<td>6 x 6, 5 x 5</td>
<td>÷</td>
<td>145.00 (3.13)</td>
<td>[0-70]</td>
</tr>
<tr>
<td>31</td>
<td>6 x 6, 5 x 5</td>
<td>mod</td>
<td>135.50 (3.84)</td>
<td>[0-50]</td>
</tr>
<tr>
<td>32</td>
<td>6 x 6, 5 x 5</td>
<td>÷, x, –</td>
<td>154.10 (1.97)</td>
<td>[0-90]</td>
</tr>
<tr>
<td>33</td>
<td>6 x 6, 5 x 5</td>
<td>x⁴, mod, ÷, x, –</td>
<td>169.40 (2.84)</td>
<td>[0-75]</td>
</tr>
<tr>
<td>34</td>
<td>4 x 4, 5 x 5</td>
<td>x⁴, mod, ÷, x, –</td>
<td>372.80 (8.63)</td>
<td>[0-400]</td>
</tr>
</tbody>
</table>

Welch-ANOVARs [114]—which minimize Type I errors [63]—with game version as a between-participant factor to determine how the snacking mechanics yielded different performance. Subsequent post-hoc tests were performed using Games-Howell [35] corrected t-tests. To control for outliers, on each performance metric, we computed the z-score and removed samples that deviated ±3 z-values from mean. Therefore, each metric consisted of different sample sizes ranging from 96 to 99 subjects and 496 to 512 sessions. The performance plots along with the ANOVA results are shown in Figure 6. Multiple variables showed significant differences, indicating that the game versions yielded different performance. Completion showed the lowest average score and the lowest accuracy (Figures 6a, 6b). Baseline and Blocking show (both similar) the lowest number of successes (# wins) per session (Fig. 6d). But Blocking has a lower failure rate per session (# fails) compared to Rewards and Completion (Fig. 6e). As indicators of total progress in the game, we operationalized the unlocked packages (Fig. 6g) passed levels, (Fig. 6h) and (actually) played levels (Fig. 6i). While unlocked packages and passed levels do not differ significantly between conditions, all three metrics indicate slower progression for Blocking and Completion.

C LONG-TERM: PERFORMANCE AND BEHAVIOUR CORRELATION

To control for performance differences in the behaviour analysis, we conducted a Spearman correlation of the performance metrics and the behaviour characteristics (Table 12), which revealed that session duration, gameplay duration and # sessions correlated significantly with most of the performance variables. Therefore, we controlled for performance variables in the behaviour analysis.

# We validated the results with MANOVAs and Kruskal-Wallis tests, yielding similar findings.
Table 12: Long-term Infinitus Algebraica: Performance and behavior metrics Spearman correlation matrix.

<table>
<thead>
<tr>
<th></th>
<th>Mean score</th>
<th>Accuracy</th>
<th># wins</th>
<th># fails</th>
<th># play rounds</th>
<th>Unlocked packages</th>
<th># levels finished</th>
<th>Best combo</th>
<th>Level number (package x level)</th>
<th>Difficulty</th>
<th>Inter-session Interval in min.</th>
<th>Session duration in sec.</th>
<th>Gameplay duration in sec.</th>
<th># sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance</strong></td>
<td>-</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>**</td>
<td>***</td>
<td>-</td>
<td>***</td>
<td></td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Mean score</td>
<td>0.535</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Accuracy</td>
<td>-0.401</td>
<td>-0.258</td>
<td>-0.283</td>
<td>-0.075</td>
<td>-0.302</td>
<td>0.159</td>
<td>-0.207</td>
<td>-0.162</td>
<td>0.176</td>
<td>-0.075</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td># wins</td>
<td>-0.401</td>
<td>-0.258</td>
<td>-0.283</td>
<td>-0.075</td>
<td>-0.302</td>
<td>0.159</td>
<td>-0.207</td>
<td>-0.162</td>
<td>0.176</td>
<td>-0.075</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td># play rounds</td>
<td>-0.401</td>
<td>-0.258</td>
<td>-0.283</td>
<td>-0.075</td>
<td>-0.302</td>
<td>0.159</td>
<td>-0.207</td>
<td>-0.162</td>
<td>0.176</td>
<td>-0.075</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Unlocked packages</td>
<td>0.258</td>
<td>-0.081</td>
<td>0.889</td>
<td>0.741</td>
<td>0.146</td>
<td>0.174</td>
<td>-0.528</td>
<td>-0.33</td>
<td>-0.441</td>
<td>0.259</td>
<td>0.23</td>
<td>-0.107</td>
<td>0.005</td>
<td>-</td>
</tr>
<tr>
<td># levels finished</td>
<td>0.634</td>
<td>0.378</td>
<td>0.088</td>
<td>-0.273</td>
<td>0.003</td>
<td>0.313</td>
<td>0.245</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Best combo</td>
<td>-0.247</td>
<td>-0.317</td>
<td>0.236</td>
<td>0.355</td>
<td>0.204</td>
<td>0.04</td>
<td>0.169</td>
<td>-0.19</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Level number (package x level)</td>
<td>-0.482</td>
<td>-0.264</td>
<td>0.127</td>
<td>0.146</td>
<td>0.174</td>
<td>0.528</td>
<td>0.33</td>
<td>-0.441</td>
<td>0.536</td>
<td>0.259</td>
<td>0.23</td>
<td>-0.107</td>
<td>0.005</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 6: Longterm Infinitus Algebraica: performance metrics by condition. The F-statistics represent the main effect of condition. The whiskers indicate SEMs. Abbreviations next to the bars indicate significant differences from the post-hoc Games-Howell corrected t-tests. BL = Baseline, B = Blocking, C = Completion, R = Rewards, N = Novelty, W = Waiting.

D LONG-TERM: BEHAVIOUR ANALYSIS

To determine if player behaviour differed between the game versions, we conducted one-way ANCOVAs with game version as the between-participant factor, the levels finished as covariates and the session characteristic metrics as dependent variables. We chose levels finished because this performance metric showed the strongest
correlation with the behaviour variables. Similar to the perfor-
mance analysis, we excluded participants with ±3 z-scores of the
respective variable. The post-hoc t-tests were also Games-Howell
corrected. Figure 7 shows the results of the behaviour analysis.
Completion has by far the most number of sessions played (Fig.7a).
This difference is also confirmed when controlling for levels finished
\(F_{1,505}=6.64, p<0.05, \eta^2_p=0.01\). The # play rounds per session differed
between the conditions (Fig.7b); also when controlling for levels
finished \(F_{1,494}=1758.70, p<0.01, \eta^2_p=0.78\). Blocking shows the low-
est # play rounds per session (close to baseline) played. Whereas
in Waiting, the participants played more longer sessions and more
rounds. There were no differences for inter-session intervals. All
game versions fell on average into a daily pattern with breaks
around 1 1/2 day (Fig. 7c). The total session duration only differed
between the conditions when controlling for # wins \(F_{1,501}=5.79,
p=0.02, \eta^2_p=0.01\) otherwise did not differ significantly between the
conditions (Fig.7d). For playtime per session, both the main effect
(Fig. 7e) and when controlling for levels finished \(F_{1,495}=351.49,\n\eta^2_p<0.42\) the ANCOVA was significant.
Playful User-Generated Treatment: A Novel Game Design Approach for VR Exposure Therapy

Dmitry Alexandrovsky∗
Georg Volkmar∗
dimi@uni-bremen.de
gvolkmar@uni-bremen.de
DMLab, University of Bremen
Germany

Maximilian Spleithöver
Stefan Finke
University of Bremen
Germany

Marc Herrlich
herrlich@eit.uni-kl.de
University of Kaiserslautern
Germany

Tanja Döring
tdoerine@uni-bremen.de
DMLab, University of Bremen
Germany

Jan David Smeddinck
jan.smeddinck@newcastle.ac.uk
Newcastle University
UK

Rainer Malaka
malaka@tzi.de
DMLab, University of Bremen
Germany

ABSTRACT
Overcoming a range of challenges that traditional therapy faces, virtual reality exposure therapy (VRET) yields great potential for the treatment of phobias such as acrophobia, the fear of heights. We investigate this potential and present playful user-generated treatment (PUT), a novel game-based approach for VRET. Based on a requirement analysis consisting of a literature review and semi-structured interviews with professional therapists, we designed and implemented the PUT concept as a two-step VR game design. To validate our approach, we conducted two studies. (1) In a study with 31 non-acrophobic subjects, we investigated the effect of content creation on player experience, motivation and height perception, and (2) in an online survey, we collected feedback from professional therapists. Both studies reveal that the PUT approach is well applicable. In particular, the analysis of the user study shows that the design phase leads to increased interest and enjoyment without notably influencing affective measures during the exposure session. Our work can help guiding researchers and practitioners at the intersection of game design and exposure therapy.

CCS CONCEPTS
• Human-centered computing → Human computer interaction (HCI): Virtual reality.

KEYWORDS
virtual reality, exposure therapy, user-generated content, game design

∗Both authors contributed equally to this research.

1 INTRODUCTION
Simple phobias such as acrophobia (the fear of heights) or claustrophobia (the fear of closed spaces) cause problems that affect many people. In western countries, 7−9% of the population suffer from simple phobias [7], which can evoke panic [114] and can reduce the quality of life. Commonly, individuals tend to avoid spaces or situations where their phobia could be triggered. Therefore, a therapy is desirable in many cases. The most common therapy for acrophobia (and many other phobias) is exposure therapy [84]. Exposure therapy can follow a paradigm of immediate extreme or of gradual exposure, with the latter being more common. The gradual exposure therapy aims to teach the patients coping strategies when facing situations that may trigger anxiety or panic and also to gradually lower the experienced intensity of the stimulus and the physiological response to it.

Due to relying on physical stimuli, exposure therapy can be difficult to implement and manage. For example, it is often not possible to travel to places with certain heights. However, a considerable and growing body of work is evidencing that exposure therapy can successfully take place in virtual reality (VR). It has been shown that virtual exposure can be as effective as real exposure [32, 42, 43, 85] and that VRET can successively be applied in treatment [43, 86]. In some cases, virtual therapy can even be more effective [28, 66, 68] and enjoyable [28, 50, 51]. In this way, VRET overcomes a range of challenges that traditional therapy faces, e.g. logistics and safety. Moreover, VRET allows for the efficient and scalable design of individually adapted therapy plans [73, 86] and can relieve therapists [28, 54]. Although 7−9% of the population suffers from simple phobias [7], only very few people undergo a therapy [74] and even if they do, many abandon it, often due to a lack of motivation [5, 16, 20, 50].
Games user research (GUR) established serious games [1] and gamification [39] as effective approaches to foster motivation for learning [22, 31, 38], physical activities [10, 105], work [36, 58] and therapy [50, 77, 103, 104]. While existing literature provides a good understanding of motivational game design (e.g. MDA [62]), it can be argued that common game design strategies for fostering motivation require careful consideration for the context of exposure therapies, as their implementation may interfere with requirements for a successful therapy. Therefore, in this work we investigate the potential of motivational game design elements, including patient-generated content for motivational games, in VRET. Our work was guided by the following research questions: RQ1: What are the specific requirements for a game-based virtual reality exposure therapy application? RQ2: How can motivational game design be applied to virtual reality exposure therapy? With the more specific follow-up questions for conceptual validation: RQ3: For the selected motivational game design strategy can measurable differences in motivation be achieved? RQ4: For the selected motivational game design strategy can the resulting exposure experience be expected to be comparable to a non-modified VRET approach?

To address these research questions, our research is composed of the following parts: (1) To inform the requirements (RQ1) we conducted a literature review that is summarized in Section 2 and two interviews with professional therapists who had a background in traditional exposure therapy that are discussed in Section 3. (2) Based on these results, we developed a two-step concept for motivational games for exposure therapy called PUT which lets users create the anxiety-inducing experience themselves followed by an exposure phase. We built a VR game for exposure therapy (RQ2) that implements the concept in a prototypical fashion (Section 4). (3) To begin answering RQ3 and RQ4, we subsequently conducted a lab-based user study with 31 non-acrophobic subjects to investigate the effect of content creation on player experience, motivation and height perception compared to a baseline condition without the PUT element (Section 5). (4) To provide early-stage ecological validation of our outcomes, we conducted an online survey with 6 professional therapists (Section 6).

Our studies reveal that the PUT approach is well applicable. In particular, the analysis of the user study shows that the design phase leads to increased interest and enjoyment without notably influencing affective measures during the exposure session. Our work provides guidance for game design of computer-mediated exposure therapy.

2 BACKGROUND
Our work is informed by conventional and VR-based exposure therapy, motivational game design and games for mental health.

2.1 Exposure Therapy
Cognitive behavioral therapy (CBT) is based on a cognitive model of mental illness, which links thoughts, behavior and emotion [45]. The model assumes that one’s ‘emotions, behavior and psychology are influenced by their perception of events. It is not a situation in and of itself that determines what people feel, but rather how they construe a situation’ [12]. CBT is problem-oriented focusing on improving the patient’s current state with mutually agreed SMART-goals: specific, measurable, achievable, realistic and time-limited [45]. CBT-based mental treatment methods such as exposure therapy (ET) are recognized as the most effective therapy methods [43, 84]. “Exposure therapy is a psychological treatment that was developed to help people confront their fears. [...] In this form of therapy, psychologists create a safe environment in which to ‘expose’ individuals to the things they fear and avoid. The exposure to the feared objects, activities or situations in a safe environment helps reduce fear and decrease avoidance” [4]. ET is based on two main mechanisms: (1) Natural habituation describes a natural decay in physiological response after frequent exposure to the anxiety stimulus. (2) Cognitive revaluation is a mechanism that comprises the patients’ reflection on the exposure and their fear reaction [84]. The therapy procedure consists of three main phases: preparation, exposure, and reflection. Over time, ET typically varies between gradual and concentrated increase in the stimulus strength (e.g. increases in height). Scharfenberger discriminates three types of exposure therapy: in-sensu where the participants only imagine the exposition to the stimulus, in-vivo – an exposure to real stimuli and in-virtuo [110] an exposure to virtual stimuli (i.e. VR) [95]. In-virtuo exposure, as realized by VRET, offers several advantages over the exposure in-vivo, since the treatment can be conducted in the therapist’s office or in remote settings and on patients who are too anxious to undergo an in-vivo exposure [43]. Furthermore, VRET allows for flexible adjustments to individual needs [86]. VRET uses immersive displays – most commonly head-mounted displays (HMDs) –, spatial audio and frequently reality-based user interfaces [63] for interaction to create strong immersion [19, 100, 102] and a sense of presence [41, 98, 99, 115]. A substantial body of research has applied psycho-physiological measures in the process of VRET and reported on studies that showed effectiveness of exposure in VR and VRET as viable options for treatment of claustrophobia [18, 75], fear of heights [40, 48, 61, 69], fear of flying [71], anxiety disorder [54, 70], and public speaking [67] among others. Emmelkamp et al. showed that exposure to heights in VR can achieve the same effect as in-vivo therapy [42]: a result that has been reproduced multiple times [28, 66, 83, 86]. Further meta-analyses on VRET studies provide strong arguments for applying VRET in clinical contexts [17, 28, 50, 56, 86].

We add to this body of research by developing a new approach for VRET that applies a two-step game design. Insights from previous work on motivational game design and game design for mental health further informed the design of these 2 phases.

2.2 Motivational Game Design
Literature on game experience often aims to understand features of games that shape player engagement (c.f., [96, 116]) and to transfer this knowledge into recommendations, guidelines, or principles for the design of more appealing games and engaging interactive systems with a purpose [21, 39]. An array of theoretical frameworks has been developed that helps to structure engaging characteristics
of games [21, 111]. These theories are adjacent to flow theory [35] as well as self-determination theory (SDT) [92] which remain the most commonly applied theories to inform and validate game design and metrics of player experience [21, 111].

In his early work in the field, Malone identified challenge, fantasy, and curiosity as 3 key elements of engaging design [76]. For individual game mechanics, a broad array of frameworks exists [2, 52, 82, 89] that identified common structures in game design and linked them with experimentally-focused empirical research. While a large portion of the literature builds on functional challenges, which address physical or cognitive skills of the players [30], games with emotional challenges received much attention recently. These games confront players with emotionally salient material through narratives, frightening scenarios, strong characters or difficult choices players have to make [30, 37]. In such settings, the players' gratifications can result from resolutions of tensions within the narratives and overcoming negative emotions [15, 30, 44, 55]. However, emotional challenges require a careful design as they can also lead to frustration and disengagement with the game [55].

In relation to ETs, facing and overcoming negative emotions have been recognized as capable elements for shaping engagement and motivation in game design. Accordingly, Illinx [11, 25] or vertigo games [23, 24] are built around core experiences of vertigo that can become enjoyable in the context of play. As fear tends to be avoided by individuals, games that provide enjoyment from an inherently negative experience could positively contribute to engagement in exposure therapy. In our work, we found that emotional challenges rather than functional challenges in the game should play an important role during an exposure phase in VRET. We also present suggestions how these emotional challenges could be designed.

2.3 Game Design For Mental Health

Games for health offer considerable potential for a broad range of application areas and enable not only fostering engagement and motivation, but also guidance (e.g. on treatment protocols in the absence of health professionals) and analysis (through tracing behaviour and/or engagement) [103]. According to Fleming et al. [46], game design offers 3 potentials for health interventions: (1) appealing potential by reaching out to target groups without access to treatment otherwise [13, 46, 77], (2) engaging potential [17, 93] due to games’ enjoyable nature, and (3) effectiveness potential since they allow for sensory rich interactive learning experiences. Coyle et al. [33] identified mental health as a key challenge that faces society and argue towards technology-facilitated intervention methods. The authors provide development guidelines for mental-health interventions (MHiAs) and suggest that HCI experts and therapists should work in conjunction to achieve the desired therapeutic objectives [13, 77]. Additionally, therapists should be involved in the design process as these applications are to be used in a collaborative therapy setting [47].

However, as of now there are few tested game design patterns and no generalizable guidelines for the design of therapeutic games in the domain of VRET that were derived based on the expertise of therapy practitioners. To fill this gap, we conducted and analyzed semi-structured interviews with professional therapists who have a background in treating patients using traditional exposure therapy.

3 REQUIREMENT ANALYSIS

Designing VRET games is challenging since concepts of traditional exposure therapy and game design elements need to be combined to achieve the desired therapeutic objectives [13, 77]. Additionally, therapists should be involved in the design process as these applications are to be used in a collaborative therapy setting [47]. However, as of now there are few tested game design patterns and no generalizable guidelines for the design of therapeutic games in the domain of VRET that were derived based on the expertise of therapy practitioners. To fill this gap, we conducted and analyzed semi-structured interviews with professional therapists who have a background in treating patients using traditional exposure therapy.

3.1 Interview Design & Structure

In preparation of the interviews, a semi-structured document was composed, consisting of bullet points from various themes that were of interest to us to address RQ1. As we aimed for unexpected input by the therapists to arise, the structure of each interview was kept rather flexible, allowing the examiner to adapt to the situation by adding or rephrasing certain questions. The preparation process of the interview followed Helferich’s method of qualitative analysis from the social sciences domain [59] and included the following 4 steps: (1) Collection, (2) Inspection, (3) Sorting and (4) Subsuming.

Following this approach, the interview document was divided into 4 categories: Techniques and Procedures (C1), Setting and Scenarios (C2), Tasks and Motivation (C3) and Supplemental (C4). C4 carried all items that could not be categorized into one of the identified clusters but still remained relevant to address RQ1.

3.2 Interview Participants

In total, 2 experts, both self-identifying as female, agreed to participate in the inquiry. Both could draw on substantial expertise in traditional ET. One expert held a master’s degree in clinical psychology, had finished clinical training in CBT and was currently working as psychotherapist specialized in posttraumatic stress disorder (PTSD), depression and the influence of childhood maltreatment. The other interviewee held a diploma in psychology and was also working as a psychological therapist offering a variety of therapeutic methods in individual or group sessions. Both had experience in using ET to treat specific phobias on a regular basis.

3.3 Conduct of Interview

The interviews were conducted as 30 to 40 minutes long face-to-face conversations in a location of the respective therapist’s choosing. Following an introductory conversation, the experts signed a consent form. This detailed the usage of audio recordings and
anonymized further processing of data gathered throughout the interview. As a next step, the examiner began to work through the semi-structured interview.

### 3.4 Interview Analysis

We fully transcribed the audio recordings and conducted a deductive qualitative content analysis [78] using the categorization approach described above. The content analysis was conducted deductively since a basic categorisation had been carried out already in preparation of the interviews. We refrained from deploying an inductive approach as our overall objective was to derive requirements that a technical VRET implementation should account for. As C1 - C4 were created with the aim for a technical solution to ET, they served as a meaningful foundation for the analysis process. In the first step of the analysis, each statement given by the therapists was coded into these four basic categories (C1-C4). A single coding item could be one or multiple sentences belonging to one response. After the material was processed for a first screening, the basic categorization was revised. To ensure validity of the re-categorization and overall coding process, we conducted an inter-coder agreement check [78]. For that purpose, two examiners processed the material independently, created their own categories and coded the data accordingly. As a result of discussion between both coders, 9 final sub-categories emerged. Techniques and Procedures (C1) was divided into: Therapy Procedure (C1.1), Role of Therapist (C1.2), Motivation of Patients (C1.3) and Possible Symptoms (C1.4). Setting and Scenarios (C2) was split into: Impact of Environment (C2.1) and Environment Characteristics (C2.2). Tasks and Motivation (C3) was divided into: Rewards (C3.1) and Possible Tasks (C3.2). Lastly, the additional category Supplemental (C4) was replaced by: Practical Applicability (C4). The categories along with the coding scheme are provided as supplementary materials: https://osf.io/4cq3k.

### 3.5 Interview Results

We provide an overview of summarized insights derived from the interviews and link them to game design considerations.

Regarding therapy procedure, the first step in ET is referred to as probatory, which serves to discuss and collaboratively decide the steps of ET between patients and therapists. This is necessary in preventing therapy from being experienced to be "other-directed" or imposed upon oneself from the patient’s perspective. In the following course of therapy, patients are confronted with their phobia multiple times until a state of habituation is achieved. In game design, this can be linked to gradual, possibly customized or adaptive increases in challenge relative to one’s own skills. Such patterns are closely linked to flow and the competence dimension of SDT.

During therapy, therapists educate patients regarding the effectiveness of the therapeutic approach as well as potential challenges and difficulties. On top of that, therapists have the role of motivators and companions, especially in the first sessions while they gradually recede from directly intervening with the process. In game design, this can be linked to the relatedness dimension in SDT, but should be considered in interplay with autonomy. It also relates to a range of social and multiplayer game design patterns.

A vital factor determining the outcome of therapy is the patient’s motivation. In line with CBT’s SMART goals [45], motivation is achieved by introducing moments of partial success that are linked to the patient’s small objectives. Motivation is greatly increased by making the patients feel autonomous, e.g. when they finally take the step to seek out phobia-triggering situations on their own. Rewards are a common component of ET and should be used to facilitate motivation. They should be defined by the patients themselves and come in a variety of forms. Both interviewees emphasized the importance of social support and its role as a reward system.

In game design terms, this outcome provides clear motivation to explore the potential of games to foster motivation and engagement. More specifically, it invites the consideration of traditional game design elements, such as points, badges or leaderboards, and more complex patterns around personal development (e.g. from RPGs).

During exposure, certain physiological symptoms (e.g. extensive sweating, accelerated heart rate and shaking legs) are expected to appear. Therapists can make use of heart rate measurements and anxiety meters to observe these symptoms. In relation to games, this outcome is important, as traditional game design patterns would overlook this element. It can however, as detailed below, be considered through patterns that emphasize the role of an involved therapist. This can also be linked to promising potential around using modern interaction devices that can record physiological signals, such as wearables, camera-based analysis, etc.

Based on the probatory phase, therapists determine a variety of scenarios for their patients to be exposed to. These scenarios address visual and tactile senses and have some relation to the patient’s daily life. The scenarios usually come in rich variety while not switching between settings too rapidly before reaching habituation. The patients autonomously define which environments they prefer and what level of intensity they choose to confront in a session. In general game design terms, this links to generative / customizable / personalized content and – as further detailed below – this also offers an opportunity to consider user-generated content.

The interviews indicate that one of the simplest and yet most useful tasks in ET is doing nothing at all while focusing on the environment (e.g. the edge of a cliff) and the symptoms it evokes. In relation to games, this constitutes another “unusual” element. When relating to game design patterns, “atmospheric” games and the associated patterns are relevant. This also marks the crucial consideration that the exposure phase may not be easily compatible with the majority of established game design patterns, which typically focus on functional challenges [30] (skill-based tasks) and narrative [9], which are both designed to be captivating.

In summary, the experts agreed that a VRET system could generally be used effectively in ET. They pointed out that the application’s content should not be random but scientifically sound and thus, incorporate traditional therapy concepts. The interviewees stated that VRET can become a vital part of therapy but should not replace real exposure. This calls for the consideration of possible game design patterns that can serve a dual-flow purpose in the sense of seeking alignment between (serious) therapy aims with game-oriented motivation mechanics [103].
3.6 Requirements & Design Implications

Based on the insights from the expert interviews, we discuss how game design patterns can be combined to build a VRET system that is tailored to support therapists who conduct traditional ET in treatment of acrophobia. We derive a variety of requirements (R1-R5) that should be taken into account when designing a VRET system:

R1: Motivation. From the interviews, we gathered that one key aspect of motivation in the context of ET is autonomy. Patients are motivated by pro-actively determining the course of therapy. More precisely, by defining sub-goals and deciding which situations to expose themselves to, they are more engaged in the process which helps them to eventually reach habituation. In summary, a VRET application should emphasize the sense of autonomy. This can, for example, be achieved by giving users the opportunity to choose or shape a scenario and the respective tasks.

R2: Communication. In line with requirements frequently stated in the literature [26, 47], the communication between therapists and patients was identified to be another crucial element of ET. Since therapists function as motivators, educators and companions during therapy, a VRET system should enable direct communication between them and their patients. This can be achieved by either placing both in the virtual scene (e.g. as avatars [48, 67]) or at least allowing audio feedback to guide patients through the experience.

R3: Scenario habituation. Scenarios should give patients a chance to reach habituation. Therefore, users should have enough time to become familiar with their surroundings. Scenes should come in a variety of different aesthetics in dependence of the respective phobia but are not allowed to be switched too quickly.

R4: Non-distracting tasks. Regarding tasks that provide a meaningful occupation in the virtual scene, the experts proposed some additional requirements. The typical activity during traditional ET involves exposing oneself to the situation in absence of any other specific activities. As a result, tasks in a virtual environment (VE) should not be distracting, to avoid shifting the focus from the situation to completing some arbitrary task. Notably, this excludes the majority of common game design patterns, which are often built around particular functional challenges and narratives. To enhance motivation, the activities in the VE have to be designed with care and should be linked to real-life rewards.

R5: Physiological symptoms. Physiological symptoms are direct results of phobia exposure and help the therapists to monitor the situation and react accordingly. A VRET system should be designed in a way that prevents additional symptoms due to technical flaws. Visual stuttering, an unstable frame rate or other visual glitches have to be eliminated. Otherwise, physiological symptoms might be wrongly attributed to the virtual exposure although they emerged on account of technical defects.

4 GAME DESIGN FOR PLAYFUL USER-GENERATED TREATMENT

Our approach splits the VRET experience into 2 distinct phases (Fig. 1): (1) The design phase, where participants use a terrain editor in VR to create their exposure (Fig. 1a and 1b); (2) The actual exposure phase, in which the participants enter their (self-designed) terrain at full scale (Fig. 1c).

The key of the concept is to allow users to design their exposure in a simulation (top-down view of a miniature map) before they experience it in the exposure at full-scale from a first-person perspective. This approach is in line with recommendations by Mine who found that user-generated content motivates creativity and self-expression as well as that world-in-miniature models can help conceptualizing the VE [80]. This “sandbox” approach is designed to foster intrinsic motivation by creating an engagement with – and a degree of personal relevance of – the exposure through playful creative action (R1 Motivation). Further, the approach empowers patients to adjust the degree of exposure to their specific needs, assess their limits and reflect on the progress, all in collaboration with the therapists. The terrain editor can also be included in the preparatory talks between the therapists and the patients in VR (R2 Communication) and help visualize the anxiety-producing stimuli. Since a self-paced scenario habituation (R3) was regarded as an important aspect, both phases needed to be designed in a way that gives users enough time and the right interaction options to either shape and customize (design phase) or explore and experience (exposure phase) the virtual scenario. Moreover, for the exposure phase, the interaction needs to be kept simple and focus on allowing the attentive perception of the exposure, explicitly avoiding any potentially distracting tasks (R4). Finally, we opted for a proven consumer-grade VR setup (HTC Vive), as the technical setup itself should not cause any additional physiological symptoms (R5).

4.1 Design Phase

For the design phase, we created a terrain editor that employed game elements from sandbox games [49, 90, 108] (e.g. Minecraft [81])
and interaction techniques from applications for 3D content creation (e.g. Tilt Brush [53], Blender [14]), which allow designing own worlds and offer great platforms for customization [112]. In the terrain editor, users interact with the VE in tabletop mode using the VR controllers and a commonly used laser pointer metaphor [64, 72]. The terrain is displayed in miniature form situated in a lobby room. To shape the terrain, users press the up and down buttons on the touch pad to raise or lower the terrain respectively. We attached a body-anchored [3, 94] UI at the controller of the non-dominant hand. From there, users get help or instructions and can select assets (e.g. buildings, nature or characters) to place in the terrain for decoration and personal customization. The asset library consists of 6 exemplary decoration objects: trees, rocks, grass, bushes, stubs, and wooden cottages. The spawn points can be placed as viewing platforms at different points of height (e.g. on buildings or mountains), which then become entry points in the exposure phase. Therapists can pre-select specific assets for the patients that are convenient for the individual cases. They also have control over the minimum and maximum heights and slopes for the VR as these parameters are most significant for shaping the intensity of the stimulus.

4.2 Exposure Phase
The exposure phase resembles examples from existing literature (cf. [43, 86]). To further support a clear focus on the experience, context menus, teleportation and terrain editing tools are disabled. As a general safety precaution, we implemented a panic button: when pressing all four grip-buttons simultaneously, the screen fades out and users immediately teleport back to the lobby.

We implemented our concept using Unity3D and a HTC Vive with the bundled hand-held controllers as the VR platform. This described approach presents an exemplary instance of an implementation that adheres to the requirements and illustrates a specific response to RQ2 in addition to the general requirements discussed above.

5 LAB-BASED USER STUDY
To validate the viability of the requirements and the specific approach described in response to RQ2 above and to provide empirical evidence with respect to RQ3 and RQ4, we designed and conducted a user study with non-acrophobic subjects, which investigates the effect of content creation on player experience and height perception. The study employed the acrophobia VRET setup with a playful terrain editor and an exposure to heights in VR as described above. The study took place in a lab, in which users wore an HTC Vive head-mounted display and could move around in a tracking space of approx. 2×3 m. The overall size of the virtual landscape at full scale simulates a world of approximately 40×60 m with heights up to 70 m. However, the surrounding skybox indicates a much larger scene, we restricted the placement of the spawn points to specific plausible regions (e.g. viewing platform or rooftops of a building). In our study, we used the prototype as described in Section 4 with the following adjustments. To only allow valid viewpoints in the scene, we restricted the placement of the spawn points to specific plausible regions (e.g. viewing platform or rooftops of a building). Additionally, after completion of the design phase, the following adjustments were applied to the terrain: a) a straight abyss down to the ground level was cut at the view point and b) the surface of each mountaintop was flattened (without notably changing the total height). For the final spawn points, we calculated the position on the spawn platform that was furthest away from the abyss. The rotation was set to look away from the edge. To avoid users watching down the "end of the world", we restricted the rotation of the buildings so that the viewing platform would always face towards the center of the terrain. For consistency between the trials in the design phase, we provided only one single circular shaped terraforming brush with a medium strength.

As we aimed to assess repeated self-reports in VR, to avoid breaks in presence [101] we added questionnaire terminals (see Fig. 1d) as world-anchored in-VR questionnaires (inVRQs) [3, 87]. In the terrain scene, we positioned the questionnaires on the opposite side of the ledge to minimize interference with the exposure when responding.

we were viewing or shaping in the second scene. We employed subjective self-reports as measures of intrinsic motivation, affect and anxiety. The assignment to the groups was randomized after balancing for gender. The study received an ethical approval.

To examine how the playful sandbox-style shaping of the exposure environment affects motivation (RQ3) and the perception of height (RQ4), we derived the following hypotheses:

Hypothesis H1: The activity of shaping terrains provides a measurably higher motivation than viewing a predefined terrain.

Hypothesis H2: There is a measurable difference on subjective ratings of anxiety induced by exposure to height between a self-created terrain and a predefined terrain.

5.1 Participants
We advertised the study on campus, via university mailing lists and through word-of-mouth. During acquisition, all subjects were pre-screened using the acrophobia questionnaire (AQ) [29] to exclude participants showing tendencies for acrophobia. An Anxiety Score of >45.45 and an Avoidance Score of >8.67 were determined as thresholds to exclude subjects from the experiment as it is one standard deviation below the score averages of clinical acrophobics [6, 29]. 31 participants (25% self-identifying as female) volunteered for our study. The mean age was 24.32 years (SD=4.32). None of the participants showed clinical tendencies for acrophobia (anxiety: M=18.87, SD=11.67, avoidance: M=3.55, SD=2.36). 20 participants experienced VR once; the others had no prior experience with VR. The groups were balanced for gender (U=133.5, p=0.49), age (t29=1.159, p=0.25), avoidance (t29=-1.03, p=0.31) and anxiety (t29=0.73, p=0.47).

5.2 Apparatus
In our study, we used the prototype as described in Section 4 with the following adjustments. To only allow valid viewpoints in the scene, we restricted the placement of the spawn points to specific plausible regions (e.g. viewing platform or rooftops of a building). Additionally, after completion of the design phase, the following adjustments were applied to the terrain: a) a straight abyss down to the ground level was cut at the view point and b) the surface of each mountaintop was flattened (without notably changing the total height). For the final spawn points, we calculated the position on the spawn platform that was furthest away from the abyss. The rotation was set to look away from the edge. To avoid users watching down the "end of the world", we restricted the rotation of the buildings so that the viewing platform would always face towards the center of the terrain. For consistency between the trials in the design phase, we provided only one single circular shaped terraforming brush with a medium strength.

As we aimed to assess repeated self-reports in VR, to avoid breaks in presence [101] we added questionnaire terminals (see Fig. 1d) as world-anchored in-VR questionnaires (inVRQs) [3, 87]. In the terrain scene, we positioned the questionnaires on the opposite side of the ledge to minimize interference with the exposure when responding.
5.3 Measurements

We assessed intrinsic motivation using the Intrinsic Motivation Inventory (IMI) [91] on the 4 sub-scales Interest-Enjoyment, Competence, Effort-Importance, and Tension-Pressure with 7-point Likert scales. To get an impression of the participants’ emotional state, we applied the Positive and Negative Affect Scale (PANAS) [34]. PANAS consists of 2 sub-scales (10-items each) that assess positive and negative affect respectively on 5-item Likert-scales.

Cleworth et al. [27] showed that non-phobic subjects rate the exposure to different heights with different ratings. Therefore, we included subjective measures of anxiety as to validate the effectiveness of the VE. To measure levels of anxiety induced by the exposure to heights, we used the 20-item State-Trait-Anxiety-Inventory (STAI) [106]. STAI contains 2 sub-scales (10 items each) that access the propensity to be anxious (trait anxiety) and a temporary anxiety with fluctuating intensity (state anxiety). As an additional measure of affliction, we used the Subjective Units of Distress-Scale (SUDS) [6, 8] – a single-item visual analog scale ranging from 0 (no anxiety) to 100 (highest anxiety).

5.4 Procedure and Tasks

We first informed the participants about the study procedure and gained their consent for participation. Next, the subjects stated basic demographics and were randomly assigned to one of the conditions (CPUT or CCTRL). Subsequently, the participants entered the lobby scene. Depending on the conditions, we instructed the subjects differently. In both conditions, they initially entered an empty lobby where we explained the panic switch, navigation and interaction with the inVRQs. After the tutorial, the participants rated their anxiety on the SUDS and we activated the terrain. For CPUT, we explained the controls of the terrain editor and asked the participants to shape and decorate the landscape to their liking, but with the constraint that the terrain should contain 3 viewing platforms with different heights each (mid high hill 30 m, high hill 50 m and a tower building 70 m). We chose these heights because all exposures should evoke a sense of notable height at different intensities for convenient subjects (explicitly not suffering from acrophobia). For CCTRL, we pre-designed a terrain that contained the same types of elements available for placement and modification in the other condition. To create a meaningful duration for the pre-exposure phase, the subjects were instructed to inspect and memorize the scene. In both conditions, the participants thereby engaged with the terrain for 3–5 min. After 2 min in, the participants gave a second SUDS rating. After finishing the editing or memorizing task respectively, the participants completed the IMI as well as a third SUDS and were further instructed to proceed to the next scene (teleport to the next location).

In random order, the participants teleported to all 3 spawn points and underwent an exposure to heights from each platform. To ensure that the participants were exposed to the heights and did not have their eyes shut, we implemented a secondary task. The participants should throw down a ball and read a series of numbers displayed on the ground when they looked down the pit. Although the secondary tasks can potentially facilitate an unintended playful experience or a distraction, this or similar tasks have been applied in the experimental setups to encourage participants engaging with the exposure task [40, 79, 97]. Each trial consisted of the following steps: (1) Participants pick up a ball and approach the ledge; (2) They extend their arm over the ledge so the ball is above the abyss; (3) They let the ball fall and follow it with their sight; (4) Participants read out numbers shown on the ground floor when the ball hits the ground; (5) They approach the terminal and rate their anxiety on STAI and SUDS. We assessed trait anxiety after the first trial. State anxiety was rated after every exposure. After all 3 trials, the participants filled out a second IMI and a PANAS and left VR. We then conducted a semi-structured interview with the participants to gain additional insights about their player experience. On average, participants spent 23.07 min (SD=3.86) in VR (CPUT: M=24.2, SD=4.02; CCTRL: M=21.63, SD=3.24; t23.38=2.14, p=0.02, Cohen’s d=0.76), with the difference resulting from a varying duration in the first phase. A detailed analysis showed the difference was only significant in the lobby scene (t11.65=11.65, p<0.01, Cohen’s d=0.76) with 2.07 min (SD=2.00) in CPUT and 3.99 min (SD=0.36) in CCTRL. The total study duration was about 30 min.

5.5 Results

Intrinsic Motivation. For all IMI sub-scales, we conducted mixed-factorial ANOVAs with the respective subscale as a within (repeat-measures) factor and condition as between factor (Fig. 2). The analysis showed significant differences within subjects only on Tension-Pressure. A Bonferroni-corrected post-hoc t-test confirmed this difference (t23=5.80, p<0.01, Cohen’s d=1.04). The MF-ANOVA revealed a significant difference between the conditions and an interaction effect on the Interest-Enjoyment subscale. A post-hoc comparison of Interest-Enjoyment with Bonferroni-correction between the CPUT and CCTRL was significant (t23=2.40, p=0.02, Cohen’s d=0.43). There was a significant interaction of CONDITION × INTEREST-ENJOYMENT between CPUT and CCTRL of the first assessment (t23=3.90, p<0.01, Cohen’s d=0.70) as well as between first and second assessment in CCTRL (t23=−3.17, p=0.02, Cohen’s d=−0.57). The results of the MF-ANOVA are summarized in Table 1. This indicates that significantly higher motivation potential can be achieved on the Interest-Enjoyment dimension (H1, RQ3), while the more
CHI PLAY ’20, November 2–4, 2020, Virtual Event, Canada

Table 1: Mixed-factorial ANOVA for both IMI assesses.

<table>
<thead>
<tr>
<th>IMI</th>
<th>Condition</th>
<th>IMI × Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tension-Pres</td>
<td>F1,29</td>
<td>p</td>
</tr>
<tr>
<td></td>
<td>4.12</td>
<td>0.05</td>
</tr>
<tr>
<td>Effort-Import</td>
<td>F1,29</td>
<td>p</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.85</td>
</tr>
<tr>
<td>Interest-Enj</td>
<td>F1,29</td>
<td>p</td>
</tr>
<tr>
<td></td>
<td>3.29</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table 2: One-sample t-tests against a neutral response (4.0) for both IMIs. Top: IMI1 assessment directly after terrain editing. Bottom: IMI1 assessment after all 3 exposures.

<table>
<thead>
<tr>
<th>IMI1</th>
<th>t30</th>
<th>p</th>
<th>mean diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tension-Pres</td>
<td>9.14</td>
<td>&lt; .01</td>
<td>1.96</td>
</tr>
<tr>
<td>Effort-Import</td>
<td>-15.40</td>
<td>&lt; .001</td>
<td>-2.15</td>
</tr>
<tr>
<td>Interest-Enj</td>
<td>7.45</td>
<td>&lt; .001</td>
<td>1.41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IMI2</th>
<th>t30</th>
<th>p</th>
<th>mean diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence</td>
<td>9.88</td>
<td>&lt; .001</td>
<td>1.34</td>
</tr>
<tr>
<td>Tension-Pres</td>
<td>-4.73</td>
<td>&lt; .001</td>
<td>-1.05</td>
</tr>
<tr>
<td>Effort-Import</td>
<td>3.46</td>
<td>&lt; .001</td>
<td>0.74</td>
</tr>
<tr>
<td>Interest-Enj</td>
<td>9.07</td>
<td>&lt; .001</td>
<td>1.51</td>
</tr>
</tbody>
</table>

Table 3: Mixed-factorial ANOVA of anxiety measures SUDS and STAI for all 3 exposures.

<table>
<thead>
<tr>
<th>Condition</th>
<th>SUDS</th>
<th>STAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxiety</td>
<td>p</td>
<td>ηp²</td>
</tr>
<tr>
<td></td>
<td>0.63</td>
<td>0.99</td>
</tr>
<tr>
<td>Anxiety × Condition</td>
<td>p</td>
<td>ηp²</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.10</td>
</tr>
</tbody>
</table>

functional (competence / effort) and potentially negatively associated (with respect to the application scenario) Tension-Pres dimension is not likely to differ dramatically (RQ4). The lack of difference in the exposure phase and increased Tension-Pres after the exposure indicate some functioning of the phase as intended (RQ2, RQ4).

To determine if the IMI measures deviate from neutral, we performed one-sample t-tests against a neutral score of 4 (see Table 2). The results (all p<0.001) show a positive difference from neutral for Competence, Effort-Importance and Interest-Enjoyment, suggesting that the experience was perceived as challenging, enjoyable and that the participants are willing to invest effort. Tension-Pres showed a significant negative difference from midpoint, which can be linked to the exposures not resulting in notable anxiety (as expected with non-acrophobic convenient subjects).

**Affect and Anxiety.** We measured affect using PANAS after each exposure task. We conducted independent t-tests to compare the participants’ affect between the conditions. Both positive affect (M=37.61, SD=4.39, t29=0.75, p=0.46) and negative affect (M=14.03, SD=4.55, t29=-0.99, p=0.33) did not differ significantly. This arguably provides further corroborative evidence to comparable exposures (RQ4). For STAI (see Fig. 3) and SUDS as measures of anxiety, we conducted MF-ANOVAs with the 3 exposures (tower, high, and mid) as within factors and condition as between factor. The results show no significant differences and no interaction effects for both measurements (see Table 3). To investigate how the anxiety evolved over the course of the study, we conducted MF-ANOVAs with anxiety and assessment number (STAI and SUDS) as within factors and condition as between factor. For SUDS, we used the assessment at the end of the first phase as baseline for anxiety. Both analyses showed significant main effects (SUDS: F3,34,67.95=17.69, p<0.01, ηp²=0.38, Greenhouse-Geisser corrected ε=0.78; STAI: F2,58=4.90, p=0.01, ηp²=0.02) but no significant differences between conditions nor interaction effects. Subsequent post-hoc tests revealed significant differences between platform1 and platform2 on both anxiety measures (SUDS: t30=2.73, p=0.03 mean diff.=3.32, see Fig. 4; STAI: t30=2.82, p=0.03 mean diff.=3.58). For SUDS, all platforms were significantly more anxiety-inducing than the baseline (platform1: t30=-5.46, p<0.01, Cohen’s d=−0.98; platform2: t30=-4.99, p<0.01, Cohen’s d=−0.99; platform3: t30=-5.54, p<0.01, Cohen’s d=−1.00). These results indicate that there were perceivable differences between the different levels of exposures in-line with the intended effects (RQ2), while there are no strongly notable differences in the resulting exposure phases between conditions, as intended (RQ4, H2a). There was a strong correlation of STAI and Tension-Pres (Pearson’s τ29=0.72, p<0.01) and a medium correlation with Effort-Importance (Pearson’s τ29=0.43, p=0.02). This underlines a positive applicability for ET (RQ2, RQ4).

**Qualitative Feedback.** At the end of each session, we asked the participants to comment on their experience and their relationship with the VE. We list paraphrased statements ordered by their frequency: “I felt related to the VE” (CPUT 1×, Cctrl 1×), “The terrain shaping was creative” (CPUT 10×), “The controls were intuitive” (CPUT 10×, Cctrl 2×), “The experience was interesting or enjoyable” (CPUT 9×, Cctrl 6×), “I had a vertigo experience when I was looking down” (CPUT 6×, Cctrl 7×), “The environment felt realistic” (CPUT 6×, Cctrl 0×), “I had difficulties with the pointer or teleporter” (CPUT 5×, Cctrl 1×), “Reading the UI was difficult” (CPUT 4×), “I would like to have more assets/controls” (CPUT 4×), “Looking down did not make me anxious” (CPUT 1×, Cctrl 7×),
Playful User-Generated Treatment: A Novel Game Design Approach for VR Exposure Therapy

CHI PLAY ’20, November 2–4, 2020, Virtual Event, Canada

Figure 3: Bar plots of STAI. The whiskers indicate the SD.

There were too many questionnaires” (CPUT 1x, Cctrl 0x), “I learned something about myself” (CPUT 1x, Cctrl 0x), “The experience was not interesting or enjoyable” (CPUT 0x, Cctrl 6x), “I did not feel related to the VE” (CPUT 0x, Cctrl 4x), “The environment felt unrealistic” (CPUT 0x, Cctrl 2x).

5.6 Outcomes

The results show that the VR experience created some sense of anxiety (RQ2), although not to the point of resulting in strong perceived tension or negative affect. This is supported by the raised SUDSs of all platforms compared to baseline and relatively raised Tension-Pressure of IMI2 after the exposure. Also, the significant decay in anxiety both on SUDS and STAI from first to third exposure indicates a habituation which is normal for people without acrophobia. The strong correlation of Tension-Pressure and Effort and the anxiety measurements further corroborates these outcomes. Low negative and high positive PANAS scores as well as positive ratings of IMI show that the experience was generally engaging and positively accepted by the participants (RQ2).

Regarding H1, the higher interest and enjoyment in CPUT shows that user-generated content can bring up interest. This is further underlined by the qualitative comments which indicate that the process is perceived as creative and supports the forming of a personal VE. This is a positive indication for the concept of playful user-generated therapy (RQ3). Moreover, only in CPUT the subjects stated that the environment felt realistic. Contrary, only in Cctrl the participants stated they felt not related, had a low sense of vertigo or lack of realism. These results show evidence in favor of H1, that user-generated content facilitates aspects of intrinsic motivation. We observed a significant increase of Interest-Enjoyment in Cctrl between IMI1 and IMI2. This results most likely from the non-interactive experience in the lobby. Conversely, we observed a drop in Interest-Enjoyment for CPUT between IMI1 and IMI2 to a level comparable with IMI2 in Cctrl. This can be explained by the simplicity of the experience in the exposure phase and by the participants’ low anxiety towards heights.

We could not find conclusive evidence to support H2. For anxiety, we only observed differences between the height levels but not between the conditions. These results, together with the qualitative comments, indicate that the exposure felt realistic and anxiety-inducing without a notable effect of the game elements on the perception of heights. We therefore reject H2 and count this as positive evidence towards the requirement that the game elements should not notably interfere with the exposure (RQ4). However, the results show fluctuations with small effect sizes and a marginal significance. Therefore, it is undue to conclude that there are no effects and these indications require further validation.

With regard to RQ2, we found that the probatory is a suitable phase of the therapy session to include game elements that do not interfere with the therapy. The self-creation of the phobic stimuli is a welcomed approach for the therapists and allows them to adjust the exposure to the patients’ individual needs; a feature that is crucial for effective computer-mediated therapy [60]. This is in line with the SMART goals [45] and with Thompson et al.’s guidelines and offers opportunities to address the Goal Setting, Problem Solving, Motivational Statements and Feedback in game design for therapy [109]. Empowering self-creation mechanics can facilitate creativity [113, 117] and interest for the therapy and address Malone’s curiosity dimension of educative game design [76] in contrast to frequently applied elements such as rewards or challenges.

6 ONLINE SURVEY EXPERT EVALUATION

To validate our approach regarding its potential applicability in a real-life therapy scenario, we conducted an online study with practicing therapists. For further validation, we included therapists with a wider range of expertise than in the initial interviews. This online survey was not meant to be a final evaluation of a fully functional system but a way to obtain an additional expert perspective on the PUT concept. The survey consisted of a consent form, 12 questions concerning the proposed game design and 13 items targeting the respondents’ professional background, technical expertise and other demographic data. To illustrate the design approach, the survey contained an introductory text with corresponding images and an embedded video explaining VRET in general and demonstrating the
PUT concept. The 3 min long video was compiled from several documentaries from public media on VR and demo clips of VR games and VR applications which corroborate the feasibility of VRET as well as a demo of the terrain editor and the exposure phase accompanied by an explanation of PUT. For copyright reasons, the video cannot be published. After perceiving the material, the therapists filled out the forms which took between 15 and 20 minutes.

6.1 Participants

6 therapists (5 female) filled out the online survey. The group of interviewees included 3 therapists specialized in depth psychology, 2 behavioral therapists and 1 expert in ET. Their ages ranged from 28 to 64 years ($M=49.33$, $SD=15.51$). 5 experts held an approbation and 1 a master’s degree in psychology. Work experience as a professional therapist was reported to be between 1 and 31 years ($M=11.17$, $SD=9.91$). All therapists were asked to rate their experience regarding VR on a 5-point Likert-scale ranging from “No Experience (1)” to “Expert (5)”. On that scale, responses lay between 1 and 2 ($M=1.17$, $SD=0.37$). None of the experts used VR in a professional capacity or for entertainment. Responses on frequency of utilizing CBT methods in therapy included “never” ($n=1$), “once a year” ($n=1$), “multiple times a year” ($n=1$), “once a week” ($n=1$) and “multiple times a week” ($n=2$). The frequency of treating acrophobia ranged from “never” ($n=1$) to “once a year” ($n=1$) and “multiple times a year” ($n=4$).

6.2 Results

Quantitative Results. Regarding applicability in a real-life therapy scenario, the therapists rated the PUT design approach on a 5-point Likert-scale ranging from “Not Useful” (1) to “Very Useful” (5). On average, the design received a score of $M=3.67$ ($SD=0.75$) (RQ2). All experts agreed that giving the patients the opportunity to create the environment for later exposure is a valuable approach. The majority of participants ($n=5$) stated that separating therapy into 2 steps (design and exposure phase) may have a positive impact on the course of therapy. The remaining expert expressed it would not affect the therapy. Regarding the influence of the patient’s contact with the scaled-down miniature terrain (RQ4), the experts reported that it may lead to positive effects ($n=4$) or no effects ($n=2$). Since non-distracting tasks were identified to be one core requirement of VRET design, we asked the experts if the design phase may distract from the actual exposure (scale ranging from “No Distraction” (1) to “Full Distraction” (5). This item received an assessment of $M=2.00$ ($SD=1.15$) (RQ4). Regarding communication between therapists and patients, 4 therapists would like to accompany their patients during the design phase, whereas the remaining 2 experts stated this would not be necessary.

Qualitative Feedback. 2 experts stated that a playful approach would be beneficial in preparation to real-life exposure as the situation itself would not be perceived to be as terrifying as in the real world. However, since the virtual representation of phobia-inducing stimuli lacks in realism, 3 experts explicitly stated that it should not replace real exposure. Additionally, 2 experts pointed out that communication should be enabled by the system whereas 1 therapist wished to enter the virtual world along with the patient. 1 expert reported that PUT may lead to a higher sense of control (RQ3) and perceived self-efficacy. Another therapist reinforced this assessment by stating that PUT design could give patients a sense of security. As a suggestion for future development, 1 therapist proposed the idea of exposing patients to virtual scenes that were not created by them in later stages of therapy.

7 DISCUSSION

With regard to RQ1, we identified 5 considerations for VRET game design. For an auspicious computer-mediated ET, patients should have control over the course of therapy by defining tasks, goals and situations themselves (R1). VRET should allow for direct communication between therapists and their patients during exposure (R2). Scenarios should come in rich variety but leave patients enough time to get accustomed to and reach habituation (R3). As for potential in-game tasks, they should be linked to real-life rewards but not be too distracting from the exposure or be perceived as tests of courage (R4). The system itself should not cause any additional physiological symptoms that might be wrongly attributed to exposure (R5). Most of these requirements are in line with the SMART goals [45] and existing frameworks on game design for interventions (cf. [33, 46, 109]). However, we found the combination of motivating (R1) and non-distracting game design (R4) a specific requirement for VRET that is rarely discussed in the literature on games for change. The outcomes of the user study show evidence that self-generated anxiety stimuli can raise the intrinsic motivation for a simple task (RQ3, H1) while not interfering with the sense of anxiety (RQ4, H2a) and thus, our design approach fulfills R1 and R4 explicitly. The experts involved in the online survey gave an overall positive assessment of PUT design in terms of applicability in a real therapy scenario (RQ2). A minority of the therapists was concerned that the patients would avoid challenging themselves. Most therapists acknowledged the PUT concept to be applicable, to have a positive effect on the motivation (RQ3) and not being too distracting (RQ4). Moreover, the survey identified potential for improvements for future iterations such as a communication interface between the therapists and the patients. The experts emphasized that communication between therapists and patients is crucial and should be anchored deeply in the design. They highlighted that VRET should be considered as an addition but not as a replacement for conventional ET. These triangulated results corroborate that PUT appears to be a viable concept that warrants further development and study.

7.1 Limitations and Future Work

Empirical user studies with phobics are ethically problematic, since they require experienced support in case of a panic. Therefore, this early-stage research evaluated the approach with convenient subjects. However, Robillard et al. [88] compared emotional reactions of phobics and non-phobics to different phobias in VR and showed that both groups react to phobic stimuli similarly but to a different degree; with phobics being affected by the VE stronger. Interestingly, the authors found no differences in the perception of game elements. This shows evidence that evaluation of game design for exposure-based VR with non-phobics should generalize to phobics as well and may be transferred to other phobias. According to Coyle et al. [33], this work is situated in the first phase...
of the development cycle. In future work, we aim to investigate PUT with afflicted subjects under the supervision of therapists. In the study, we did not assess autonomy since the study was conducted with convenient subjects and therefore, we did not expect a raise on autonomy, as no deregulation compared to traditional ET would be notable to the subjects. However, literature on motivation shows that user-generated content can raise the sense of autonomy and competence [92, 113, 117] and that creativity is linked to self-empowerment [107]. As the study was mainly concerned with the interference of game design, height perception and intrinsic motivation, assessment of autonomy was beyond scope. This should, however, be addressed explicitly in future work with phobic patients and SDT. Further, the subjective ratings on anxiety (SUDS and STAI) show high variances that are likely attributed to low sensitivity towards heights. Although this work is based on validated measurements, future work should consider physiological measures of anxiety such as heart rate or galvanic skin response for clearer and more reliable data.

As fear generally embodies avoidance, it is difficult to create an intrinsic motivation to coping with one’s phobias. Thus, presenting anxiety inducing stimuli in an enjoyable way is a challenging demand for game design. Our research shows that game design for non-intrusive playful experiences is rarely explored in the literature and requires further attention. Likewise, further traditional game elements could be employed, especially in the design phase, as well as game mechanics that may contribute positively to the therapeutic potential of the exposure phase, e.g. around vertigo experiences or emotional challenges associated with phobias. Both are underexplored in the literature and should be considered in future research on VR ET.

8 CONCLUSION

We presented a novel game design approach for VR exposure therapy: playful user-generated treatment (PUT). We investigated this concept with a multi-angled approach: a requirement analysis, the implementation of a VR-based game, a user study with convenient subjects, and an online survey with a distinct group of therapists. The requirement analysis revealed that PUT is considered to be a useful addition to conventional therapy and that VR ET game design demands specific considerations, which are rarely addressed in the literature. The requirements led to the PUT design approach that separates ET in VR into design and exposure phases. Our two validation studies with the VR game show that PUT is well applicable. In particular, the user study reveals that the users’ content creation leads to increased interest and enjoyment without notably influencing affective measures during the exposure session. The positive indications with convenient subjects also suggest that further study and validation in an applied therapeutic context appear warranted.

ACKNOWLEDGMENTS

This project was partially funded by the project Scalable Pervasive Health Environments of the German Research Foundation (DFG) as part of the SPP 2199.

REFERENCES


The Role of Physical Props in VR Climbing Environments

Peter Schulz
University of Bremen
Bremen, Germany
schulzp@uni-bremen.de

Dmitry Alexandrovsky
University of Bremen
Bremen, Germany
dimi@uni-bremen.de

Felix Putze
University of Bremen
Bremen, Germany
felix.putze@uni-bremen.de

Rainer Malaka
University of Bremen
Bremen, Germany
malaka@informatik.uni-bremen.de

Johannes Schöning
University of Bremen
Bremen, Germany
schoening@uni-bremen.de

ABSTRACT
Dealing with fear of falling is a challenge in sport climbing. Virtual reality (VR) research suggests that using physical and reality-based interaction increases the presence in VR. In this paper, we present a study that investigates the influence of physical props on presence, stress and anxiety in a VR climbing environment involving whole body movement. To help climbers overcoming fear of falling, we compared three different conditions: Climbing in reality at 10 m height, physical climbing in VR (with props attached to the climbing wall) and virtual climbing in VR using game controllers. From subjective reports and biosignals, our results show that climbing with props in VR increases the anxiety and sense of realism in VR for sport climbing. This suggests that VR in combination with physical props are an effective simulation setup to induce the sense of height.

CCS CONCEPTS
• Human-centered computing → Virtual reality; Haptic devices; Empirical studies in HCI.

KEYWORDS
Virtual reality; Climbing; Biosensors; Fear of Falling; Physical props; Presence.

Figure 1: (a) A subject is climbing horizontally on the wall while wearing a head-mounted display and virtually climbing at a height of 10 m. (b)-(d) Example views of the VR scene from the user’s perspective.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-5970-2/19/05...$15.00
https://doi.org/10.1145/3290605.3300413

ACM Reference Format:
1 INTRODUCTION & MOTIVATION

In sport climbing, it is important to overcome the fear of falling as it holds back climbing athletes by impeding movements [28] and effects on the athlete’s performance [26]. Climbers learn to overcome their fear of falling by building confidence in training and, for instance, by habituation through controlled falling [48]. Beyond the fear of falling, sport climbing is also associated with general height-induced anxiety, which in its pathological form is acrophobia. One way of treating this condition is exposure therapy, the ‘golden standard’ in the field of psychotherapy [59].

Starting in the mid-1990s, scientists began experimenting with virtual reality exposure therapy (VRET), which gave them more control over the grading of exposure and reduced the inhibition threshold for a therapy [59]. VRET uses head-mounted displays (HMDs), spatial audio and e.g., data gloves to create a high sense of presence for the subjects. Presence is considered the defining factor for VRET to work [15, 27]. In their study, Emmelkamp et al. showed that exposure to heights in virtual reality (VR) achieves the same effects as in-vivo therapy [17], a result that has been reproduced multiple times since then [36, 55, 59]. In a large scale study with 100 acrophobia patients over the course of two weeks, Freeman et al. [21] showed significant improvement in overcoming fear of height after few weeks of treatment with an automated VRET in comparison to regular in-vivo therapy. Besides presence, immersion is an important measure in VR. From an extensive content analysis of 83 studies on presence and immersion, Cummings and Bailenson conclude that the immersion provided by increased levels of user tracking and the quality of the mediated virtual environment has a medium effect on improving the sense of presence in the virtual environment (VE) [13]. For sport climbers and industrial climbers, a VR training environment allows climbing training to counter fear of falling even in times of recovery or when a real climbing environment is not available.

In this paper, we report on the design of a VR climbing environment that supports overcoming the fear of falling. Based on traditional training of handling fear of falling in sport climbing and VRET from psychology, we investigate to what extent realistic physical sensation is necessary to create an immersive climbing experience which comes with a natural anxiety induced by a feeling of height. We conducted an experiment in VR to compare the manifestation of anxiety (measured directly and indirectly through induced stress) and presence at subjective and physiological levels. We compared climbing, a complex whole body movement, using props (hand- and footholds) with controller-based climbing simulation and observed higher subjective anxiety and presence for props.

2 RELATED WORK

In this section, we discuss how anxiety is related to sport climbing and present related work on physiological measurement of anxiety and presence as a foundation for our study design. Furthermore, we discuss VR, augmented reality (AR) and mixed reality (MR) in climbing scenarios.

Sport climbing is a superset of various different activities. The most prominent activities are lead climbing, top-rope climbing and bouldering. When doing a lead climb, a climber is attaching the rope with quickdraws to the wall to secure him while ascending the route. When climbing a route top-rope, the climber is secured by the rope through an anchor at the top of the route. Bouldering characterizes a climbing activity in low heights (typically below 6m) without any belaying equipment.

Arousal, Fear, Anxiety and Stress in Sport Climbing

Breckner describes fear as an essential factor in all levels of sport, from mass sport to elite sport [7]. He further describes, that reasons for fear depend on the personality and quotes Allmer and Schulz, who supposed that there are different trigger situations: (i) risk (ii) limitation, handicap or failure (while feeling the pressure to succeed) (iii) insecurity (when confronted with unknown situations significant for the subject) (iv) competition [1]. For climbers, it is unpleasant to face an imminent risk of falling, yet they are motivated to venture falling and mastering the hardest parts of the route (“conquer the crux”). At the same time, it is linked to physiological arousal [43, 44] with typical somatic symptoms like uneasiness, pale skin, cold hands and sweating. Also, increased muscle tensions, raised respiratory rate (RR) and heart rate (HR), elevated urinary frequency and adrenaline/no-adrenalin level have been observed [6]. Tietjens also considers specific behaviors such as fleeing and avoidance as typical fear reactions [71].

Anxiety is triggered by threats that are perceived as uncontrollable or unavoidable [45, 64]. In his thesis on anxiety in extreme mountaineering, Breckner [7] refers to stress as a relational concept [41] where stress is the result of an individual evaluating a situation defined by the environment. This evaluation includes inner requirements (self-imposed goals) as well as outer requirements (e.g., trainer-imposed goals). The individual evaluates the impact of the environment for its well-being and the necessary resources or skills for success.

Since the beginning of the 20th century, research has been concerned with the effects of anxiety on performance in sports. Early experimental studies led to the hypothesis that there is a point of anxiety-infused activation resulting in peak performance. Later this idea was replaced by a theory of individual zones of optimal functioning (IZOF) that
Ruiz and colleagues verified in several studies [62]. Based on Masters et al. [49] and Liao et al. [46], Bertle conducted a study to examine the stress reactions to an impending fall of climbing protection and height above ground [3]. The study investigated two conditions of height (low/high) and two climbing styles: leading and top-rope. The results of Bertle’s study show significant cognitive and somatic anxiety reactions in a high (fall) condition compared to a low (fall) condition. Further, Bertle found that well-trained subconscious routines and movements are processed more consciously under stress [3]. In several experiments with climbers Pipers et al. measured effects of anxiety induced by fear of falling while climbing. In their setups, the participants traversed (i.e., climb horizontally) an identical route at different heights above ground, while being belayed in top rope. All studies showed an increase of anxiety-related effects in the high condition and evidence for anxiety-related reduction in performance [56–58]. Hardy and Hutchinson found in their studies with climbers that less anxious climbers perform better than highly anxious climbers. The authors assume that self-confident subjects and skillful climbers are less likely to be impacted by anxiety [26].

Physiological Measures of Anxiety
Bertle [3] summarizes multiple studies with skydivers and basejumpers. Inexperienced skydivers show a raised heart rate (HR), respiratory rate (RR) and electrodermal activity (EDA) throughout a flight which do not decrease until landing [20, 74]. Breivik et al. [8] could not corroborate Eysenck’s and Calvo’s Arousal Theory nor could they validate Fenz and Epsteins [18] consistent correlations of HR to subjective feelings among experienced jumpers [20]. However, they showed that experience (number of jumps) is correlated negatively to arousal [8], an observation Hardy and Hutchinson [26] later made again for climbers. Roth et al. found similar results that showed rising physiological measures and anxiety self-ratings for parachutists in anticipation of a jump—indeed independent of the level of expertise, only at higher levels for novices [61].

Judging from the body of work examined, we based our study on the assumption that anxiety in sport climbing is a multidimensional phenomenon. We focused mainly on one dimension: fear of falling, as it is verified to have a hindering effect on performance [18, 26, 58]. Regarding terminology, this paper is primarily concerned with anxiety, and not with fear or phobia. Therefore, we use the terms fear of falling and anxiety of falling synonymously.

VR-Treatment for Stress and Anxiety
Acrophobia is an abnormal fear of heights which in turn is related to the fear of falling [3]. Therefore, research looking into a climbers’ fear of falling in VR would be of particular interest, in context of a more general question on how VR is suited for the treatment of fear in specific situations. However, no such work has been published so far, regarding acrophobia and its therapy using VRET. Several studies showed that VRET for acrophobia treatment is as effective as regular exposure therapy, producing the same fear of height in virtual environments as in vivo [40, 53, 59]. Diemer et al. ran a controlled experiment to compare height reactions of phobics and non-phobics. Both groups showed similar physiological arousal when exposed to height [14]. This implicates that not only phobics but also participants comparable to those of other climbing-related studies listed above, show respective effects.

Since we use height above ground as a stressor, the perception of height in VR is of particular concern. The extent to which the (simulated) height works as a trigger depends on the effectiveness of VR and thereby the amount of presence it evokes in its users [51]. Meehan conducted three studies that investigated the influence different aspects of VR on presence. He looked at effects of multiple exposure, passive haptics and effect of frame rate on presence. In the passive haptics experiment, Meehan showed that a plywood ledge has an influence on presence, while participants were walking towards a virtual cliff [50]. Gandy et al. summarize several studies examining effects of varying levels of detail on presence and perception—with contradicting results [22]. On the one hand, Zimmmons and Panter could not find significant differences in presence or subjective stress when running variations of “the pit” experiment [50], ranging from wireframe to radiosity [75]. However, Slater et al. found increased psycho-physiological measures, HR and skin conductivity (SC), for a higher level of realism (reflections and shadows) [68]. In alignment with Gandy et al. (2010), who summarized that a “[virtual] environment does not have to be visually perfect to make users feel present or capable of a motor task”, we recreated a degree of realism comparable to that of “the pit” experiment by Meehan et al. [51].

Augmenting Climbing with Technology
Byrne and Mueller investigated the motives for climbing and how digital systems can support sport climbing [10]. The authors identified five key motivational themes: risk, challenge, social engagement, beauty of nature and reliving the experience. These motivational themes offer guidelines when designing augmented climbing experiences. There is also a significant line of work that investigates augmented climbing. From our literature examination, we observed several recurring motives: immersive experience, training and monitoring.
Immersive climbing setups. There are VR games that provide a virtual climbing experience. For example, *The Climb* (http://www.theclimbgame.com) employs a head-mounted display (HMD) and allows players to climb in mountain environments using VR controllers. Similarly, Kosmalla et al. [39] showed a VR climbing environment which employed a 3D scanned climbing wall and hand tracking using Leap Motion. To train motor skills, Kajastila and colleagues designed *Augmented Climbing*—a bouldering setup which employs projection on the climbing wall and full-body motion tracking [34, 35]. The setup provides augmented feedback on the climbing wall to enhance sport training. From three user studies the authors conclude that augmented climbing is an effective tool for engagement, movement diversity and also lets climbers forgetting about fear of height. Liljedahl and colleagues developed *DigiWall*, an interactive climbing wall with sensors, light and sound feedback. The authors promote a set of applications, which include training, play and sound design [47]. *BetaCube* by Wiehr et al. implements a self-calibrating camera which employs a wall projection of life-size video replays and holds highlighting for route creation [73]. Kosmalla et al. build on *BetaCube* and designed an AR climbing system to teach novices climbing movements. Their application employs a recording mode using a KINECT for the trainer and three types of video playback using AR-goggles and a projection on the climbing wall. Participants performed various tasks with different modalities of video guidance. The results indicate that a hybrid visualization suits best to fit all requirements of a climbing training environment [37]. *Venga!* is a setup that allows for fully augmented climbing in VR [70]. In their setup, the authors 3D-scanned a climbing wall and applied hand and foot tracking to create an immersive climbing experience in VR.

Wearables for climbing. Mencarini et al. defined a design space of wearables for climbing as tools for augmentation and communication [52]. They identified that in sport climbing either the walls can be augmented or climbers can be equipped with wearable devices. The authors suggest instead of eliminating negative emotions, the training environment should help climbers dealing with them and promote to design wearables that support the communication of fear in sport climbing. Based on results of an online survey and a follow up user study, Kosmalla et al. [38] found that sound is suited best as a communication channel for a wrist worn wearable in climbing. *ClimbAX* is a wristband sensor which was designed to monitor the climbers’ performance. From accelerometer data, the wearable extracts various performance metrics (e.g., climbing episodes and hold transitions) and provides climbing statistics [42]. To guide beginners in improving their climbing technique in real-time while climbing.

![Figure 2: Schematic view of the climbing routes. Two identical climbing routes at different heights, each consisting of 14 hand and footholds. T_h and T_f mark the holds at the turning point.](image)

3 STUDY

In our study, we focus on the level of immersion-induced presence, required to provoke fear of falling in VR. Therefore, we create three different conditions: real physical climbing in 10 m height (*C_{real}*), virtual reality climbing with real props (*C_{props}*), at virtual height of 10 m but physically 20 cm above the ground and an entirely virtual climbing simulation with game controllers (*C_{ctrl}*), at virtual height of 10 m while standing on the ground. We used physiological measurements (heart rate and skin conductivity) together with questionnaires asking for the subjective responses to compare the effect of three conditions. In alignment with Meehan [50],
who observed a positive effect of props on presence, we derive the following hypotheses:

**Hypothesis (H₃):** Presence (Dᵢ) is measurably higher for climbers in VR when using props (C₂) compared to a purely virtual condition (C₃).

**Hypothesis (H₄):** Anxiety (D₄) (measured directly and indirectly in form of stress) is higher for climbers in VR when using props (C₂) compared to a purely virtual condition (C₃).

Our study design has two independent variables: props or game controllers and real or virtual vision resulting in three conditions. We aimed to examine their effect on the following dependent variables: physical exertion, anxiety and presence. Since anxiety is highly subjective and varies between participants, we chose a within-subject experimental study to test our hypothesis. The order of the conditions was randomized using a Latin Square [23].

For the study, we developed a multimodal VR climbing setup that employs foot and hand tracking and allows for real climbing with a HMD.

### Measurements

In alignment with the studies by Slater [66, 67] we measured stress and anxiety as indicators of presence. Our primary stress and anxiety measures were heart rate variability (HRV) [11, 31] and skin conductivity response (SCR) [9, 50]. We recorded the physiological data using Bitalino Board Kit 2018. As shown in Figure 3, we placed the ECG electrodes at the collarbone and the EDA sensors on the shoulder of the participants. Additionally, presence was measured using the widespread Igroup Presence Questionnaire (IPQ) on the **spatial presence (sp), involvement (inv)** and **experienced realism (real)** subscales [63].

Meehan [50] suggests HR, RR and EDA as indicators for presence. Regarding sport climbers, Hardy and Hutchinson [26] reckon that those signals—when elevated—may be interpreted as signs of somatic anxiety, too. However, HR and RR cannot be described as unitary functions of arousal, because they are also dependent on physical activity [12]. Hence, they can only be interpreted as indicators for somatic anxiety when comparing conditions with a similar level of physical activity. In our setup, these are only C₁ and C₂. As a control variable for the level of movement, we also asked the subjects to give a rating of their movement level (C₃) to a purely virtual condition (C₃) using props.

For the study, we developed a multimodal VR climbing setup that employs foot and hand tracking and allows for real climbing with a HMD.

### Procedure and Tasks

Our experiment procedure included the following steps: (i) briefing (ii) put on climbing harness (iii) fit with HR, RR, and EDA electrodes (iv) initial rest (C₁) and upfront questionnaires (v) randomized order of C₁, C₂, and C₃ (vi) debriefing. After each condition, the participants had time to recuperate while filling in the Igroup Presence Questionnaire (IPQ) and AT questionnaires.

The primary task for all conditions was traversing a climbing wall, from a start point to a return point and back. The subjects were only belayed in condition C₁ utilizing an auto-belay-device. To ensure that the participants are aware of their exposition to height, we introduced secondary tasks: (i) before climbing, throw a ball down to the ground and read out the number displayed next to where the ball dropped (ii) while resting at the return point, look down and read out numbers shown on the ground floor.
Climbing Routes
We conducted our experiment in a local climbing gym with a wall that fulfills the specifications for one lane of an International Federation of Sport Climbing (IFSC) speed climbing wall. The traversal route is angled at 5° along its entire length and is easily accessible on the ground floor and from a balcony at 10 m height. Figure 2 depicts the final route layout for condition C_
real (10 m) and its twin at the bottom of the wall for condition C_
props. Both routes have a launching platform made of plywood, each serving as the start and end point for the climbing task. The routes had a difficulty of 4+ on the International Climbing and Mountaineering Federation (UIAA) scale; a difficulty level that is suitable for beginners. Therefore, a precise foot tracking was implemented using existing architectural 3D model as a foundation to match back against the wall and injure the participant in case of a fall. Furthermore, an entirely virtual rope without a physical rope present would break the immersion if the subjects tried to grab it and also lead to a fall.

Participants
We followed Guo et al.’s guidelines [25] and used GLIMMnPSE to calculate an appropriate sample size. We assumed an effect size of Cohen’s d between 0.7 and 1.2 at a confidence level of \( \alpha = 0.05 \) and a power \( P > 0.8 \) similar to climbing-related experiments by Hardy and Hutchinson [26]. Our calculations required a sample size between 11 and 30 participants to show significant effects.
A total of 28 participants (13 female), mean age 30.7 years (SD=10.6), volunteered to participate in the experiment. The participants, for the greater part customers of the local climbing gym, were mainly lead climbers (23), but also top-rope climbers (4) and one boulderer.

The average UIAA climbing degree (level of difficulty) of the lead climbers was 6+ (±1 degree) and of the top-rope climbers 5+/6− (±1 degree). About three quarters of the participants climb at least once a week. 13 participants had never used a VR system prior to the experiment and 13 only once; the remaining two participants had used various systems frequently before. All subjects provided written informed consent.
We used the short version of the STAI [24] as a standard check to measure trait anxiety. The questionnaire consists of 10 items on a 8-point Likert-scale. The overall mean trait anxiety was 34.9 (SD=14.9). The results were comparable to the scores originally surveyed by Spielberg [69]. Hence, these results indicate that the participants had no extraordinary tendency to respond to situations perceived as threatening with an elevation in state anxiety [57].
To identify and eventually exclude participants who fulfill the psychiatric criteria of acrophobia, we asked them to answer the vHI questionnaire [30]. The vHI questionnaire was only presented to the 16 participants who affirmed the opening question; none of the users gave indications to suffer from acrophobia. The average severity score was M=1.4 (SD=0.16).

4 RESULTS
In this section, we report the results gathered from the physiological measurements and self-reports for the three conditions of our experiment. First, duration of climbing and the level of physical exertion is examined to ensure we only compare homogeneous conditions later on. Then, we continue with the outcomes for stress and anxiety and, finally, presence. Due to missing or corrupted data we excluded up to 7 participants for when we performed our analysis. We report the number of subjects, that were applied for each statistical test.
Figure 4: (a) Hand over a hand hold. (b) Screenshot of the resulting, masked infrared overlay. (c) Heel-mounted VIVE tracker. (d) Virtual representation of the foot in VR.

Figure 5: Duration per condition. The error bars indicate the first and fourth quartiles.

Duration
We conducted a repeated measures analysis of variance (ANOVA) to examine the effect of the condition on the time that participants took to complete the tasks. The analysis was performed with n=23. Regarding the duration per conditions (Figure 5), there was a significant effect: $F(2, 44) = 9.092$, $p < 0.01$. The effect has an observed power of 0.966 with $\eta^2_p = 0.292$. A correction for sphericity was not necessary ($\chi^2(2) = 0.872$, $p = 0.188$). Pairwise Bonferroni tests showed significant differences between $C_{\text{real}}$ and $C_{\text{props}}$ ($p < 0.01$), $C_{\text{real}}$ and $C_{\text{ctrl}}$ ($p < 0.05$), but not between $C_{\text{props}}$ and $C_{\text{ctrl}}$ ($p = 0.575$).

Physical Exertion
We recorded the vital signs HR and RR because they are indicators for physical exertion. HR was measured using an ECG in combination with a chest strap to obtain reference signal less prone to motion artifacts. We conducted a repeated measures ANOVA to examine the effect of the condition on HR for both sources. For ECG-based HR (n=23), Mauchly’s test of sphericity indicated that the assumption of sphericity had not been violated, $\chi^2(2) = 5.866$, $p = 0.320$. There was a significant effect of condition on the HR: $F(3, 63) = 22.502$, $p < 0.01$. The effect has an observed power of 1.0 with $\eta^2_p = 0.517$. For the recordings from the chest strap (n=22), Mauchly’s test of sphericity indicated a violation of sphericity, $\chi^2(2) = 12.653$, $p < 0.05$ and therefore, a Greenhouse-Geisser correction was used. There was a significant effect of condition on the HR: $F(2.192, 46.035) = 92.877$, $p < 0.01$. The effect has an observed power of 1.0 with $\eta^2_p = 0.816$. Pairwise Bonferroni tests revealed a significant difference in HR between $C_{\text{real}}$ and $C_{\text{props}}$ ($Z=-2.517$, $p = 0.012$), $C_{\text{real}}$ and $C_{\text{ctrl}}$ ($Z=-2.34$, $p = 0.019$) and $C_{\text{props}}$ and $C_{\text{ctrl}}$ ($Z=-3.817$, $p < 0.01$). The results are shown in Figure 6.

Since the RPE yields discrete, ordinal results, we used a Friedman test to examine the effect of condition on self-reported physical exertion. There was a significant effect of condition on RPE, $\chi^2=23.234$, $p < 0.01$. A pairwise Wilcoxon signed ranks test showed a significant difference for $C_{\text{real}}$ and $C_{\text{props}}$ ($Z=-2.517$, $p = 0.012$), $C_{\text{real}}$ and $C_{\text{ctrl}}$ ($Z=-2.34$, $p = 0.019$) and $C_{\text{props}}$ and $C_{\text{ctrl}}$ ($Z=-3.817$, $p < 0.01$). The results are shown in Figure 7.
Figure 7: Average rating of perceived exertion (RPE) as reported via visual analog scale (VAS) after each condition (C_real, C_props, and C_ctrl) on a scale from 6 to 20. The error bars indicate the first and fourth quartiles.

Stress and Anxiety

For the analysis of anxiety, we used the ECG signal and derived the mean R-R interval (mRRI) from it as a measure of HRV. Further, we measured EDA and derived the SCR-level from it (Figure 8). To examine the effect of the condition on each measure, we conducted a repeated measures ANOVA. For mRRI (n=21), Mauchly’s test of sphericity indicated a violation of sphericity, $\chi^2(2)=16.401, p<0.05$ and therefore, a Greenhouse-Geisser correction was used. There was a significant effect of condition on mRRI: $F(2,40.719)=4.331, p<0.05$. The effect has an observed power of 0.725 with $\eta_p^2=0.178$. A pairwise Bonferroni test revealed only a significant difference between $C_0$ and $C_{props}$ ($p<0.01$).

To examine SC, we looked at two of the non-specific responses (NSR): response-count and conductivity level [60]. The first was calculated from the rolling average of detected peaks within a 20 s time-frame. Neither the response count nor the level did show any patterns. Hence, unlike the HRV, we could not find any significant differences between the conditions for the EDA signals.

In contrast to RPE, the VAS anxiety thermometer (AT) uses a continuous scale (Figure 9). We could conduct a repeated measures ANOVA without correction ($\chi^2(2)=0.692, p=0.708$). There was a significant effect of condition on self-reported anxiety, $F(2,44)=5.364, p<0.05$. The effect has an observed power of 0.815 with $\eta_p^2=0.195$. A pairwise Bonferroni test of the conditions revealed that the mean AT response for $C_{props}$ is significantly higher than for $C_{ctrl}$ ($p<0.01$).

Presence

In the VR conditions $C_{props}$ and $C_{ctrl}$, the participants rated their sense of presence through an IPQ questionnaire (n=27); the results can be seen in Figure 10. To examine the effect of the condition on presence scores, we conducted paired t-tests. There only was a significant effect of the condition on realness scale (REAL) with $t(26)=2.621, p=0.014$. The effect has an observed power of 0.714 with $\eta_p^2=0.209$. As suggested by [50], we also examined the rolling average of the HR—using a window of two respiration cycles—with no significant differences between the conditions.

The authors of the IPQ offer a database with results from multiple studies [32]. This dataset includes 37 records for experiments with 3D graphics displayed via HMD. Considering these, we found higher scores than those in the database on all scales. However, the differences were significant only for spatial awareness (SP) and realness (REAL) in condition $C_{props}$ ($\delta_{SP}=1.007, p<0.01; \delta_{REAL}=1.596, p<0.01$) and $C_{ctrl}$ ($\delta_{SP}=1, p<0.01; \delta_{REAL}=1.328, p<0.01$).
Figure 10: IPQ scores for the conditions \( C_{\text{props}} \) and \( C_{\text{ctrl}} \) on the scales general presence (PRES), spatial awareness (SP), involvement (INV) and realness (REAL). The error bars indicate the first and fourth quartiles.

### Qualitative Feedback

After each run, we asked the subjects to comment on what they had just experienced. Here is a list of statements ordered by frequency: "It felt like being up there [on the high platform]" (\( C_{\text{props}} \), 6 times), "Climbing [in VR] felt more realistic [than using the controllers]" (\( C_{\text{ctrl}} \), 6 times), "The scene was very realistic" (\( C_{\text{props}} \), 5 times), "My feet were off a bit" (\( C_{\text{props}} \), 4 times), "I was not sure if I can trust the [VR] technology" (\( C_{\text{props}} \), 3 times), "I was more anxious, because I was not fixed to a rope [in VR]" (\( C_{\text{props}} \), twice), "Tracking was lost" (\( C_{\text{props}} \), twice), "I had a good feeling [when standing on the high platform]" (\( C_{\text{props}} \), once), "I could not see my hands, now and then" (\( C_{\text{props}} \), once), "I didn’t feel like being up there [on the high platform] at all" (\( C_{\text{props}} \), once).

### 5 DISCUSSION

Our findings confirm the positive effect of props on presence [50], in particular for a VR sport climbing environment in which the participants performed complex whole body movements. As expected, both climbing conditions \( C_{\text{real}} \) and \( C_{\text{props}} \) were equally exerting. That is fully supported by HR and partially by RPE since the latter shows significantly higher scores for \( C_{\text{real}} \) and \( C_{\text{props}} \), compared to \( C_{\text{ctrl}} \). We hypothesize, a VR sport climbing environment is a useful tool to train climbing situation, e.g., to overcome fear of falling. This is also supported by the self-reports of our participants, as the IPQ scale for realness showed a higher score for climbing with props compared to climbing with controllers.

Regarding \( H_\alpha \), we assumed anxiety is a consequence of the increased realness, measured on multiple scales: HRV, SCR and AT. The AT score (\( M=2.43, SD=2.87 \)) for real climbing (\( C_{\text{real}} \)) was consistent with that reported by Pijpers et al. [56]. Further, in alignment with our hypothesis \( H_\alpha \), AT was significantly higher for \( C_{\text{props}} \) than \( C_{\text{ctrl}} \), although we could not find evidence in the physical measurements. For HRV, only mRRI was significantly lower in \( C_{\text{props}} \) compared to \( C_0 \) (initial resting), i.e., there is no significant difference between the climbing conditions.

In contrast to HR, the RPE scores are higher in \( C_{\text{props}} \) than in \( C_{\text{real}} \). That difference may be attributed to increased anxiety in VR (not necessarily climbing-related) as observed by Hardy and Hutchinson [26]. One subject (ID=4), who reported slight offsets of tracking in \( C_{\text{props}} \), mentioned that this experience was similar to climbing on brittle rock faces outside the climbing gym, which requires extra care and focus. The levels of exertions we demanded from our subjects were lower than those of Pijpers et al. [58]. In their experiment, they measured HRs ranging from 133 bpm to 176 bpm compared to 98 bpm to 108 bpm in our study.

From the qualitative feedback, we identified some noticeable glitches and occasional tracking issues. Our participants reported that particularly in a climbing scenario, a precise tracking is needed as, e.g., the foot placement is a crucial task in the climbing process.

### 6 LIMITATIONS & CHALLENGES

Most of our participants reported high immersion and enjoyment during the VR climbing condition. However, experiments in complex VR climbing environments are challenging since climbers are used to very accurate hand and foot placements on the holds. Therefore, a precise and robust tracking of the climber’s movements is required to avoid a break of immersion. Nevertheless, most comments were positive, and we collected only a few statements regarding glitches and mismatch problems. This assures us that our prototype worked very well and provided a sufficient trust in our setup. Also, the reports on high immersion indicate a small influence of the glitches. For security reasons and to avoid a break of immersion, we deliberately decided not to show a virtual delay rope in the VR conditions (\( C_{\text{props}} \), \( C_{\text{ctrl}} \)). This potentially affects the comparability between the two physical climbing conditions (\( C_{\text{real}} \) and \( C_{\text{props}} \)). Nevertheless, we argue that only two subjects mentioned a missing rope in \( C_{\text{props}} \) and the IPQ scores are above average, indicating that the presence of the rope has a minor effect on the overall experience and therefore, all three conditions are comparable.

Surprisingly, the SCRs measures, response-count and conductivity level, appeared to be utterly unaffected by the conditions. There are several possible reasons for this: First, we only looked at non-specific response (NSR) and not a specific event like [9, 33]. Since the majority of our subjects climbs regularly, they may be accustomed to height. Second, the EDA electrodes on the shoulder could be affected by the physical measurements. For HRV, only mRRI was significantly lower in \( C_{\text{props}} \) compared to \( C_0 \) (initial resting), i.e., there is no significant difference between the climbing conditions.
complex physical activity such as climbing and encourage the development of sensors that are suited for complex whole body movements.

7 CONCLUSION & FUTURE WORK

We conducted an experiment to investigate whether realistic physical sensations are necessary to create an immersive climbing experience (a complex whole body movement), which often comes with a natural anxiety induced by a feeling of height. We applied multiple measuring instruments, including objective bio-signal measures as well as subjective self-reports. This multimodal analysis showed a comprehensive picture of the experience which would not be captured with one of the instruments only. Our results are in alignment with earlier work by Insko [33] and Meeman [50], who also found increased self-reported and measured presence when using props. We extended the related work as climbing is a "complex whole-body task" [56]. Thereby, we showed that climbing free solo in VR results in increased stress and anxiety as well as presence. Hence, we were able to reproduce height-induced anxiety in VR, as described by [26, 57] for real climbing, and therefore our VR climbing environment could be a safe and effective training environment for sport climbing. In the future, we aim to evaluate the VR climbing environment in training with professional sport climbers over a longer time period to get insights how such a technology can be incorporated into an effective training schedule.

In this paper, we tested novel approaches for interaction mechanics throughout this experiment, which could be refined in future work. On the one hand, the heel-mounted trackers helped to work around loss-of-tracking-issues, but, on the other hand, they turned out to be sensitive to shocks and faulty calibration. Further, by employing a combination of video-overlay and tracked hands, based on the Leap Motion system, we could create a realistic representation of hands that allowed precise grasping and could compensate for the occasionally incorrectly visualized feet.

Similar to the findings of Muender et al. [54], our results show a strong connection between physical activity in VR and tactile sensation. When designing training environments, it is valuable to identify which components of the setup are most effective. We conclude that physical props can substantially increase the degree of realism and significantly strengthen the experienced presence. We argue that a VR setup that employs full body interaction and simple physical props can be used for safe and effective training environments. These findings support theories on embodied interaction [2, 16] and are of particular interest for designing immersive VR experiences and VR-based training applications.

8 ACKNOWLEDGMENTS

We thank the DAV climbing hall Bremen for their support and the opportunity to conduct the study in their facilities. This project has received partly funding by the German Federal Ministry of Education and Research in the grant program “Erfahrbares Lernen” (experienceable learning) and by the Volkswagen Foundation through a Lichtenberg Professorship.

REFERENCES

[14] Julia Dietz, Nora Lohkamp, Andreas Mühliberger, and Peter


Tim Woodman, Michelle Huggins, Christine Le Scannf, and Nicolas
VRBox: A Virtual Reality Augmented Sandbox for Immersive Playfulness, Creativity and Exploration

Thomas Fröhlich
Digital Media Lab, TZI
University of Bremen
Bremen, Germany
tfr@uni-bremen.de

Dmitry Alexandrovsky
Digital Media Lab, TZI
University of Bremen
Bremen, Germany
dimi@uni-bremen.de

Timo Stabbert
University of Bremen
Bremen, Germany
stabbert@uni-bremen.de

Tanja Döring
Digital Media Lab, TZI
University of Bremen
Bremen, Germany
doering@uni-bremen.de

Rainer Malaka
Digital Media Lab, TZI
University of Bremen
Bremen, Germany
malaka@uni-bremen.de

ABSTRACT
Augmented sandboxes have been used as playful and educational tools to create, explore and understand complex models. However, current solutions lack interactive capabilities, missing more immersive experiences such as exploring the sand landscape from a first person perspective. We extend the interaction space of augmented sandboxes into virtual reality (VR) to offer a VR-environment that contains a landscape, which the user designs via interacting with real sand while wearing a virtual reality head-mounted display (HMD). In this paper, we present our current VR-sandbox system consisting of a box with sand, triple Kinect depth sensing, a virtual reality HMD, and hand tracking, as well as an interactive world simulation use case for exploration and evaluation. Our work explores the important and timely topics how to integrate rich haptic interaction with natural materials into VR and how to track and present real physical materials in VR. In a qualitative evaluation with nine experts from computer graphics, game design, and didactics we identified potentials, limitations as well as future application scenarios.

CCS Concepts
• Human-centered computing → Virtual reality; Empirical studies in interaction design;

Author Keywords
Augmented sandbox; virtual reality; gestural interaction; playful interaction; natural materials for interaction; tangible interaction.

INTRODUCTION
Playing with sand is something most people relate to from childhood days, where this unstructured play gave room for experimentation, exploration, or cooperation. As children we develop important skills in the sandbox like proprioceptive sensing, body and space awareness, or social skills and can engage in creative behavior and open-ended play [16]. These fond memories last a lifetime and, when given the opportunity, many adults enjoy playing with sand and rediscovering the sensations of this activity.

Interactive and augmented sandboxes offer these kind of experiences but further add an interactive element to the sand, which allows for increased interaction and application possibilities. Here, people can play with the sand, making use of their natural ways of expression while at the same time perceiving interactive visual feedback that is projected onto the sand. Through these visuals, changes of the sand’s surface can be immediately experienced and reacted upon. In the literature, there is a considerable body of research that has explored augmented sandboxes and its tangible affordances (e.g., [26, 24, 2, 8, 25]). In many application scenarios, sandboxes have been applied to offer a better understanding of spatial and geographical phenomena [13, 22], or to experience the terrain
Most setups use a projector mounted above the sandbox, projecting an image on the surface of the sand and a depth camera for tracking the sand’s shape. For example, Reed et. al’s project on heat map-like coloring of the sand to indicate different types of landscapes [24].

While setups like these already offer a nice direct interaction with sand as natural material to change the visual output, we move a step further and integrate real sand interaction into virtual environments (VEs) with mid-air gestural interaction. This offers new ways to use and explore the sand landscape, such as by placing virtual objects into the scene or by teleporting oneself into the shaped environment. Moreover, our work explores the important and timely topics how to integrate rich haptic interactions into VR and how to track and present real physical materials in VR.

In this paper, we present VRBox, an interactive sandbox for playful and immersive terraforming that combines the approach of augmented sandboxes with modern Virtual Reality technology and mid-air gestures. The system extends the existing approaches with an additional interaction and visualization layer to interact with the the sand and virtual objects. Furthermore, in VRBox users can switch perspectives from a tabletop mode into a first person mode by teleporting directly into the own creation, allowing for an immersive experience of the landscape in full-size. We provide technical details of the construction as well as releasing all material and source code to the public. Furthermore, we evaluated VRBox in a qualitative user study with nine experts from computer graphics, education, and game design. The evaluation focused on questions regarding the user experience, technical aspects, and possible application use cases. In particular, we were interested in exploring research questions such as:

- Is Virtual Reality a suited extension for sandbox interaction?
- Is free hand interaction beneficial considering the discrepancy of visual and tactile experience?
- Can first-person experiences like teleporting into the 3D world increase immersion and user experience?
- How do users rate the playful and creative aspects of VR-Box?
- Can the sand surface be visually abstracted (e.g. block world) without compromising user experience?

Our results show a very positive attitude towards VRBox, highlighting the immersive and creative experience that the system offers due to the novel first person experience and extended interaction capabilities using virtual objects in VR. We contribute application scenarios in productive and entertainment contexts such as gaming and collaboration.

RELATED WORK
In 1997 Ishii and Ullmer envisioned to use the affordances of physical objects and to bring them together with digital technology as part of tangible user interfaces [14]. Since then, natural materials have also been explored as part of interactive systems (c.f. the material probe study on natural elements as interaction materials by Häkkila et al. [10]). These natural materials are used because they naturally offer rich multimodal feedback, come along with deeper meanings and often provide emotional associations. Moreover, the natural qualities of these materials have the potential to define, guide and constrain the interaction, such as in ephemeral user interfaces [9] that last for a limited time only or change over time, depending on the properties of the materials used for interaction. In our setting, we use sand as natural element that comes along with great properties for sculpting and shaping, affords moving the hands through and often naturally triggers associations of playfulness, creativity, exploration and early childhood.

Real sand or similar material as interaction material has been used in several applications and installations. In 2002, Piper, Ratti, and Ishii presented their Illuminating Clay system for the analysis of landscapes in the landscape design process [22]. Their system consists of a clay model that users manipulate, a laser scanner for capturing of depth information, and a projector that closes the feedback loop for visualization of different landscape analysis functions. The authors point out the simplicity and effectiveness of displaying 3D data directly onto the surface of a 3D model. Sand Garden [24] is a game, where a digital landscape must be shaped, providing valleys, water, and mountains in order to make the virtual villagers happy. The system uses a projection and a second screen to display the game world. The sand itself is augmented and functions as a controller for the game. An educative use case for augmented sandboxes is presented by Sanchez et al. [26], engaging students to experiment with topographical maps exploring importance of water, erosion, or mountains in the evolution of the landscape. The advantages are the better understanding of abstract concepts, better involvement, and more efficiently experimentation using a sand model. Inner Garden by Sol Roo et al. [25] is a tool to support mindfulness practices, which is defined as "the act of playing a deliberate and non-judgmental attention to the present moment" where "positive impact of a person’s health and subjective well-being" can be achieved. Users create a miniature world by playing with the sand and the natural elements of the world are connected to physiological measurements like respiration, helping to stay focused on the body. This system also offers a VR mode where users can explore their creation virtually. Another interesting use case has been developed by Audi [1] where users can first form their own racing track in a not augmented sandbox. The resulting race track is then imported into a VR driving simulator, offering direct perception of the user’s creation by driving through the digital model. However, no direct interaction with the sand in VR is supported. To the best of our knowledge there is no augmented sandbox setup, in which users can directly interact with physical sand in virtual reality.

Among the rare example works that have explored integrating the interaction with natural materials in VR is a system by Pier and Goldberg [21], in which users can directly interact with water in VR. The authors conclude that water is an extremely immersive medium that fosters curiosity and children-like play. The experience of real sand around the feet while wearing a VR head-mounted display has been recently explored by [11], who created a multi-sensory VR environment in which users
are in a beach scenario. The authors advocate to incorporate non digital sensory experiences in virtual reality, on the one hand expanding the design space of virtual environments and on the other allowing users to adjust the virtual environments at their taste. Nevertheless, they do not use the sand for interaction in VR.

Some other works investigated how to realize interaction with physical objects in VR and how physical or tangible interaction impacts the immersion. For example, Hoffman investigated in a comparative study how users anticipate the physical qualities of objects in VR when subjects can physically touch the objects [12]. The results of the study show that participants could better anticipate the physical properties of the objects in VR and also the sense of presence was enhanced. TurkDeck [6] uses human controlled walls to provide haptic feedback to a subject that navigates through a environment in VR. The authors found out that, when physical walls are present in VR, the enjoyment and the sense of realism is raised.

In our work, we provide a novel approach for the integration and presentation of a physical sand landscape in real time in VR, thus providing novel haptic feedback experiences. Furthermore, we combine this with a teleporting feature in order to allow the exploration of the self-designed landscape from a macro-perspective. For exploring small architectural models at scale in VR, a similar approach has recently been applied by MacroScope [27]. In this system, a wide-angle camera is placed on a tabletop and captures objects that are placed on the table. The camera streams the image directly to a HMD, allowing users to experience the tabletop setting from the perspective of the camera. Oasis [28] and Kintinuous [15] are systems that use a Google Tango or a Kinect respectively first to record a 3D environment and then to allow users to navigate through it in VR. Sra and colleagues [28] implemented in their system Oasis object tracking and detection of walkable areas. Using these techniques, the system allows for virtual world generation, where the shape and the boundaries of the physical space are remapped to different visual styles.

In our work, we combine real sandboxes with VR interaction and thus provide a novel and original system that explores natural physical materials in VR, gestural interaction and teleporting in a terraforming use case.

**PROTOTYPE**

**Setup**

The box has a volume of $140 \times 80 \times 30$ cm. The ground and the inner walls are covered with pond foil and an additional layer of non-reflective cloth. This size provides a good balance for single user as well as for collaborative scenarios in the future. When working with sand (especially with wet sand) there can be occlusions in the surface structure that one single sensor cannot recognize. Therefore, to minimize occlusion of the user we use three Kinect v1 sensors that are mounted on a railing at a height of 120 cm above the sand surface. The tracking volume is captured from three angles of an isosceles triangle reconstructing of a full 3D point cloud of the sand model. For virtual reality, an HTC Vive is used, opening the opportunity to track additional objects. To enhance the immersion, a Leap Motion sensor, attached to the HMD, provides information about the hand and arm positions of the users. The application runs on a Windows 10 System with an Intel i7 3,4 GHz and 16 GB of RAM.

**Volume Tracking and Calibration**

The system is implemented as a client-server architecture. The server application captures the depth images from the Kinect devices at 30fps and sends them, after preprocessing, to the VR application (client) via UDP. Figure 2 shows an overview of the incorporated components.

To fuse the Kinect sensors, the relative orientation of each depth camera was computed using OpenCV [29, 3]. In the next step, the point cloud data is generated using the Kinect Fusion toolkit3. The fused point cloud has a resolution of $512 \times 512 \times 256$ points. The VR environment is implemented in Unity3D4. We designed a 3D model of the sandbox as a reference for the calibration between the sandbox and the head mounted display. To match the positions of the real and the virtual sandbox, all components need to be calibrated precisely. We calibrate the position of the sandbox relative to the Vive tracking system by placing a Vive controller on the sandbox, so the virtual representation of the interaction environment matches. To provide a satisfying VR experience, it is crucial, that the application runs at a steady frame rate of 60 fps. Hence, we had to accept some compromises about the level of detail of the terrain. In the VR application the terrain is represented in a blocky Minecraft[18] like style on a $98 \times 59$ mesh with 32 height levels. This resolution provides a good compromise between level of detail and latency. To avoid the hands being mapped on to the terrain, we added a threshold at the top rim of the box. Everything above the rim is ignored by the Kinect sensors and the area underneath the

1. https://www.vive.com/
2. https://www.leapmotion.com
4. https://unity3d.com

Figure 2. The VRBox system. The Box is filled with sand for the interaction (see also Figure 4).
hands is marked as occluded. To increase performance the terrain is split into chunks of 8 by 8 blocks. Only those blocks, that have been occluded (by the hands) for a duration of at least 6 frames (100 ms) are updated. Since the incoming point cloud stream from the depth sensors is noisy, to avoid jittering, the height of each block is updated gradually at a rate of 10 Hz. When a new height value for a block is provided by the sensors, the block is repositioned to that value, which takes up to 200 ms. The total latency from the moment the sand was formed, until the changes are visualized in VR accounts 1.5 - 2 seconds. On top of the tracked terrain the hands provided by the Leap Motion are rendered.

The Terraforming Use Case

Our use case is a terraforming task where users form a 3D world of their liking with the sand and extend their creation by adding a water level to create rivers and other waters, add different lighting moods and decoration objects like trees, flowers, and castles. In VR, users see the virtual representation of the tracked sand-box and sand surface and form a terrain with their bare hands. Furthermore, they can interact with virtual menu using their hands. The menu items are placed to the left and right of the user in VR and are represented as bubbles that users can interact with using their hands, see Figure 5. The bubbles serve different functions. First, a set of them can be used to decorate the sand surface with virtual objects like trees, castles, and other assets. These can be used in the sense of a pouring glass, where users grab a bubble using a pinch gesture (see Figure 3) (e.g. containing a tree asset) and "pour" the trees onto the sand surface, see Figure 4. Single objects can also be placed in the same way where we concentrated on a small set of items that can be put onto the sand individually. On the left side of the user, one bubble can be used to manipulate the water level in the form of a vertical slider. Another bubble serves as a teleport item. Users can grab the teleport bubble and put it on the sand surface to mark the location where they would like to be teleported to. Then, users have to perform a "look up" gesture where they have to look up straight in order to perform the teleport. When the head reaches the highest point, the teleport is performed and when looking down, users see from a first person perspective, standing directly in the virtual world. Looking up again teleports back into tabletop mode. Further, a delete bubble and lighting bubbles (create sunset or sunrise) were implemented.

USER STUDY

In order to evaluate VRBox, we conducted a qualitative user study with nine experts from computer graphics (4), education (1), game design (3) and HCI (1). We chose these areas to gather feedback from a technical perspective and from two relevant application areas that we envision to profit from such a technology. All participants were between 30 and 50 minutes in VR. On average, the subjects formed terrains for 20 minutes and explored the environment from first person perspective for 10 minutes.

In the following, we introduce our nine participants in more detail. P1 works as a computer graphics (CG) researcher working on collision detection and development with the Unreal engine where he creates virtual testbeds for various research projects in the context of VR. P2 is a 3D environment artist working on in a local game studio where she models various types of 3D assets and environments. P3 works as a researcher developing tools for VR in relation to semantic support in the operating room. P4 is a researcher in CG working on collision detection. P5 is also a researcher with a CG background and game developer working on VR and multi user VR environments and telepresence. P6 is a game designer with a local game studio working on gameplay and the development of game mechanics. P7 is a game designer creating virtual environments and levels for a racing simulation in a local game studio.
game studio, working on the topic since 2011. P8 is a special needs teacher for mathematics and biology currently working in youth welfare services. Participant 9 is an HCI researcher in the area of educational user interfaces and a hobbyist game developer. Our participants are between 20 and 39 years of age, with eight males and one female subject. The demographics are summarized in Table 1.

Procedure
We tried to keep the humidity and form stability of the sand as constant as possible. Therefore, before each interview day, the sand was watered with 1.5 to 2.0 liter. The sand was then stirred up, so that the sand could soak in the water. We prepared the sand by smoothing the surface in order to provide a “blank canvas” for each participant. After a general introduction, participants were introduced to the system, explaining the overall functions of the system including the use of VR and Leap based hand tracking. We pointed out and demonstrated that for optimal hand tracking the hands should always be in front of the user so that they can be seen by the Leap sensor. After this, we started the tracking systems and again smoothed the surface. We assisted in putting on the VR headset and made sure that hand tracking worked as expected. While wearing the headset, we asked to try to form the sand surface building hills and valleys in order to get a better feeling for the system. After up to five minutes, we introduced the virtual controls next to the sandbox.

Users were now asked to create an environment of their liking, again introducing the different tools and interaction possibilities. The task was to at least create one hill or mountain, one water body and one lowland area. Further, users should add a lighting mood of their choice and place some virtual objects into the scene order to extend their creation.

After the scene had been created, we took various screenshots in VR, exterior view of the VRBox, and insight view in VR (could be multiple). In the next step, users filled in the System Usability Scale (SUS) questionnaire [4] and subsequently performed the interview as the last part of the study.

Interview
As we were interested in evaluating technical aspects and the possible application areas learning and design, we structured the interview in three parts: general questions, technical aspects, and application. In the general questions we asked what was especially fun about using the system, and which motivational aspects users anticipate for using the system. Also, we were interested how users perceived the teleporting into their own creation and how they felt about it, if users could think of other ways of placing objects and adjusting the lighting, what other features they felt missing in order to extend it as a creative platform, and if they thought that VRBox supported creativity in a playful way. We also asked which playful aspects of the system were generally appreciated, and if the users thought that using VRBox more creative results could be achieved compared to existing software solutions (e.g. Cities Skylines [7], Minecraft [18]).

In relation to the technical aspects, we asked how users evaluated the discrepancy between the direct haptic sensation when working with the sand and the decoupled visual feedback, how users perceived the visual representation and resolution of the block world, and what they thought about the time lag in refreshing the block world. Further, we were interested in the perception of sand as a material for the task and what alternatives users could envision, which kind of tools users were missing analog to sand box toys, and how tangible user interfaces could be integrated and used in VRBox. For the latter, we introduced the concept of tangibles in the form of videos and examples based on the research of Ishii and others (c.f. [14]).

Regarding the application scenarios, we asked what application scenarios users could envision for VRBox, how VRBox could be used as a game element of core game mechanic or another application, which kinds of creative use cases users perceived as possible, and if they could think of using VRBox as an educational tool. We were interested if users could think of using VRBox in a productive context (e.g. planning tool or design tool for game worlds), how multiple users could interact with the system, and also asked for general feedback of what was especially positive or negative about interacting with VRBox.

RESULTS
To analyze the interview results, we first collected the answers and distributed them to the three main categories General Comments, Technical Aspects (e.g., Usability and Tools), and Application Scenarios. Then we summarized similar answers in order to reduce the amount of statements. In the next step, we scanned the answers for emerging topics beside the ones that we defined for the interview. We identified Playfulness, Creativity, and Material as emerging categories. Thus, we distributed the accumulated answers into the categories General Comments, UX and Usability, Tools and Useful Extensions, Application Scenarios, Sand as Material, Playfulness and Creativity. In the following, we present the results along these categories.
General Comments
We asked what our participants generally perceived as fun. The answers include exploration (P1), working with the sand (P2, P5, P8), visualization in VR (P2, P3), innovativeness (P3), teleporting (P4), immersion (P4), and decoration (P6, P8). One participant mentioned that through the 3D representation he didn’t feel like working with actual sand (P6). P9 noted the advantage of having a sense of scale. Critique was expressed regarding the tracking accuracy, that the deco tools disappear too fast, and that the placement of individual objects was too difficult (P1).

Concerning reasons that motivate people to use VRBox, we found creative application (P1, P2), relaxation (P1), self expression (P2), playfulness (P2, P7), ease of use compared to traditional tools (P2, P9), manual crafting (P3, P8), absence of any peripherals (P3), quick prototyping (P4, P9), and the use of the inner perspective (teleporting, P4, P7). Further mentioned were visualization capabilities (P5), the colorful, interactive style and versatile use (P6), the novelty (P7, P8), and sense of scale (P9).

We asked specifically how the participants perceived the teleporting as a core mechanic. The answers highlight the easy to use activation gesture (P1, P3, P5), and found, that the function adds value (P2), is exciting and provides a nice feeling to "being in" the own creation (P3, P5, P8), that it is engaging to be more involved in creating a world (P4, P6, P8), that it worked well and it’s a powerful, moving, and thrilling feeling (P5). P6 commented that it would add value to include more details (P6), that tunnels and walking options would increase the experience. Improvements could be done by overcoming some clipping issues (P7), and one participant felt insecure about the function and had a dim feeling about it, but was still curious (P7). P5 highlighted the higher immersion compared to AR sandboxes.

P9 stated that it is really interesting to create something and then see it from a first-person perspective almost immediately in 3D and that users get a really good understanding of the space (which is the biggest limitation with modeling in 2D). In his opinion, VRBox makes 3D modeling a lot easier because with traditional systems, there are a lot more iterations on assessing the 3D scale and how the environment feels when being in there.

VRBox could be an alternative to real models (P2) and one comment highlighted the transient property as beneficial (P6). P3 said that it would be interesting to compare VRBox to current terrain generation techniques. The Minecraft [18] like look was appreciated (P1), while the menu too far away (P1), the box was too small (P2), and sound could be added for a stronger sense of immersion (P2). P3 mentioned that VRBox is an interesting work with a lot of potential. P4 said the systems is pleasing and intuitive.

UX and Usability
Data regarding UX and usability aspects were gathered with the SUS questionnaire and by open interview questions that
targeted technical aspects, which may negatively affect the immersion and the workflow.

On the SUS scale, the participants rated VRBox in the top quartile with \( \mu = 78.6 \) and \( \sigma = 9.45 \) Interestingly, P8, who has never tried out VR before is rather an outlier (z-value = 2.23) with a SUS-score of 57.5. The other participants, who are more familiar with game design and computer science rated the setup with an average score of 82.5, with ranging values from 70 up to 85.

All participants perceived the scanning resolution as grainy (P7) allowing to capture "big differences" while small details get lost (P4), and wished for a higher resolution of the terrain. The level of detail was nevertheless well accepted as "part of the rules" (P7) and did not disturb the immersion (P2, P3, P4, P5, P7, P8), some even highlighting the usefulness of the abstraction for improved imagination (P8). P3 said that without the decorations the displacement is negligible but the virtual objects cause a discrepancy, because they cannot be touched. P7 and P9 noticed a discrepancy, but it wasn’t an obstacle and that they relied on the visual feedback than on the tactile sensation. Quite often, the subjects noticed glitches or artifacts when the tracking system falsely interprets the user’s hands as part of the terrain (P1, P3, P4, P6). The latency of the terrain updates was perceived by P3, P5, and P9 as critical. Particularly P3 felt latency more important than the level of detail. The other subjects found the delay acceptable. P6 and P4 didn’t realize much delay or were used to it from other applications (P1). P4 said: "When you are in the process of moving it, it doesn’t bother that much." However, P1 pointed out, that when working with the sandbox regularly, the latency can become tedious. P9 noted that the hands were in the way and that he felt a disconnection between the sand and the display, missing a one to one relationship.

Tools and Useful Extensions

Most subjects thought of physical tools to help them shape the sand, such as shovels, rakes, and molds (P1, P3, P5, P6, P7, P8, P9). P1 and P2 imagined virtual tools for water sources or extended UI elements. Options for colors adjustments and adding night scenes, moon, stars, and clouds would be beneficial (P2, P4, P5). Tangible objects could be used as points of interest or in-game objects like spawn points, water sources or physics simulations (P1, P3, P5, P6, P7, P8). Additionally, they can be placed as modifiers to shape the surface of the terrain like spline curves (P2). P1 proposed a control panel at the edge of the sandbox, and P9 imagined a tip-jar metaphor.

Generally, the pouring metaphor for object placement was accepted (P1, P2, P3, P4). Alternatively, objects could be placed with the hands (P3, P4) or with scattering or sprinkling gestures (P1). It was criticized that the placement lacks precision (P2, P3, P4, P5, P6, P7) and that undo/redo or re-placing weren’t implemented. Options such as spawn radius adjustment, object scale, color, and different spawn area shapes were proposed. Similarly, subjects P1-P6 and P9 asked for more control over the light settings, including color, direction, and intensity. P3 also suggested a sun object, that can be placed in the scene as a main source of light.

Application Scenarios

All participants suggested either prototyping, landscape modeling, or city planning as main application scenarios. The system could also be used as an installation in museums or festivals (P6). VRBox could be integrated in the content creation pipeline in game design (P1), for idea generation, architecture visualization, and presentation in general (P4). P1 states that VRBox is not ready for production due to the low accuracy, and P6 saw it more as a toy than a tool.

For games, VRBox could be used in simulations and planning games (P5, P6), physics simulations, table tops (P5), exergames (using the weight of the sand, P1), tower defense (P1, P9), or god games (like Black&White [20] P1, P5). Multi user scenarios include co-op and player vs. player games (P1, P2, P3, P4, P5, P6, P8) with synchronous playing using multiple HMDs or distributed playing (external devices). Co-op was suggested with one user modeling and the other acting on the game play (P1, P6, P8) or giving instructions from outside (P7). Here, co-presence visualization would be needed (P5, P6). City planning simulations, landscape shaping (P2, P8), game world and map creation, 3D paintings, and Zen garden games have further been proposed (P1).

Furthermore, participants mentioned general application in education (P4, P6, P7), weather and climate simulation (P1, P2, P9), city transport simulation, agriculture planning (P1, P2), urban and rural infrastructure (P1), terrain modeling (P2), tech and science teaching (P2, P3), and 3D modeling (P3). Also, teaching physical and mechanical principles (P4, P8), creativity support in education (P6), art classes (P6), experiments and creativity (P2), relaxation and mind wandering (P1) were mentioned.

Sand as Material

All participants agreed that for the use case of terraforming sand is an appropriate material, especially because it is well-known from playing with (regular) sandboxes (P1). P2 and P6 found it too dirty. P3 mentioned, that sharp edges were difficult to create with sand, and he recommended using a denser material. P6 suggested to use kinetic sand, as it brings more robustness to the forms. But he also admitted that it is costly. P8 added that clay could be added for models that are impossible to form with sand.

Playfulness

P1 stated a disconnection between tactile and visual resolution and an indirect visual perception which makes VRBox not playful. Working with the hands made P2 feel like in childhood days, but the participant also stated that the menu interaction disconnected her from the immersion. Teleporting was playful for P4 while modeling was particularly nice for P5. Object placement was very good to bring the world to life. P6 was unsure about playful characteristics and stated, more detail would be better and more interaction could be added. Decoration and the interactivity using the own body was playful for P7 and creating an own scene was a huge attraction.

Creativity

All nine participants agreed that VRBox playfully supports creativity. P1 was ambivalent if a beautified or plain model
supports imagination and thus creativity better. Simultaneous modeling in the physical and virtual medium is a creative factor for P2 and P8 because it provides power and supports playful experimentation. VRBox enables art creation by facilitating manual work without peripherals, and VR adds unlimited possibilities to extend creations (P3). The model creation is translated to something more using visuals and applying force increases the liveliness and fun (P4). Similar models could be created in 3D tools but VRBox is more interactive, fun, and playful. P5 said that a lot of things can be created and there is no pre-determination of results. Users can decide what world they want to be in as a major factor for creativity. VRBox would be useful for kids and younger people as modeling is a form of creative expression (P6). The spatial perception is supported by modeling landscape (P7).

P1 says that VRBox offers more creative results than standard applications and that sand interaction is something more than building a landscape in block by block. P2 states an increased freedom using the own hands and no peripherals, a good connection to the task without learning an interface, and the very low learning curve. In line, P8 mentions the direct access to the many parameters of the model, and the benefit of direct connection between action and result. VRBox is faster than traditional modeling and useful for non-technical people like artists compared to Blender and the like, enabling 3D content creation for everyone (P3). P4 was unsure if VRBox is more fun or if VRBox stimulates creativity. P5 was unsure if VRBox is more creative because of VR itself, and task dependency. P7 highlights creativity because of immersion and the many possibilities. P9 denied creativity because of the limitations of what can be modeling with sand.

**DISCUSSION**

In the following subsections, we discuss the accumulated interview data related to our research questions.

**VR as an Extension for Sandbox Interaction**

The use of VR in combination with a sandbox was well received by our participants and there is supporting evidence that this combination works well and adds benefit to the technology. In this sense, our participants positively emphasized the overall visualization capabilities that VR offers, teleporting and inner perspective, higher immersion, sense of scale and presence, found it engaging to be more involved in the own creation, stressed that VR offers a good understanding of space and location (also objects in relation to each other), and pointed out that VR adds unlimited possibilities to extend the own creations. Based on these finding, we are confident that VR is a beneficial extension to augmented sandboxes, considering our group of participants, being mainly familiar with VR applications.

**Free-Hand Interaction and Visual/Tactile Discrepancy**

The free hand interaction itself was perceived ambiguously. We noted several comments about the Leap sensor losing the tracking, disorientation issues of the hand model, and the problem of keeping the hands in the field of view while rotating the body. On the positive side, our free hand interaction metaphors were received well. Participants positively mentioned the pouring-style interaction for object spawning using the bubbles and the absence of peripherals. From this we conclude that free hand interaction is beneficial for interacting in VR with a physical medium like sand. However, technical challenges have to be overcome regarding hand tracking, posture recognition and stability. For other prototypes and systems, more powerful hand tracking systems can be used that do not have these limitations that the Leap sensor and Leap development kit have. Regarding the discrepancy between the tactile sensation and visual perception, we also received mixed feedback. There are two technical components to this issue. First, the sensing of the sand surface and second the visualization in block style. From the user perspective, we received statements that it "feels right", that there is no big difference, the discrepancy is there but negligible/wasn’t in the way, and perceived as part of the rules. On the other hand, users told that it didn’t feel like working with actual sand, that tracking accuracy is low, that there are glitches, and that the visualization should look more like sand. Overall, most of the users were mostly satisfied with the experience of seeing virtual objects that resemble a large but not complete part of their tactile sensation. We conclude that this can be improved but was not a problem in our case.

**First-Person Experience for Immersive Feeling**

Being able to teleport into the own creation and having the sense of “being” there was perceived as one of the most appealing features of VRBox. We collected a large variety of supportive statements. According to the participants, the function adds value, is exciting, engaging, powerful, moving, provides immersion, provides a good understanding of the own creation, and is a thrilling experience. One participant felt insecure in the beginning, but then also said that he was still curious about the function. After overcoming some minor technical issues like clipping, this feature adds great value to augmented sandboxes and is very well perceived due to the immersive characteristic.

**Playfulness and Creativity**

We found different playful aspects in our study. Working with the sand as in childhood days was mentioned several times, the teleporting feature, object placement, bringing the world to life, and that using the own body were perceived playful. For one participant the disconnection between visual and tactile made VRBox less playful. As there are mainly positive attributions for playfulness, we assess VRBox as a playful environment. Regarding creativity, all participants agreed that VRBox supports creative expression. The reasons were different: beautification, modeling of sand, endless possibilities, no predetermination, expressiveness, freedom, and low learning curve. We collected a lot of positive feedback regarding the creative capabilities and assess that with VRBox, users feel creative.

**Visualization and Visual Abstraction of the Sand Surface**

On the positive side, the block world look was generally appreciated, it was perceived as colorful, the level of abstraction was well accepted, "fits well", was pleasant, did not interfere
with the immersion and provided a good abstraction of the sand. It was further stated that a visual approximation of the details can lead to a more free expression compared to very high detail. However, participants also mentioned that a higher resolution would be better. From this we conclude that the overall block style is suitable for this kind of use case and people tend to like it, also because due to Minecraft, they are already used to it. However, more detail could be added by increasing the resolution of the blocks, keeping the overall style while increasing the detail.

Limitations
In general, all participants showed a positive attitude towards VRBox. However, the sample size and composition of our participants needs to be considered when interpreting our findings. With only one female participant, there is a high gender bias in our study results, which needs to be considered in future work. Further, most participants are working in computer science related fields and have prior experience with VR, having a natural tendency towards the acceptance of new technology. P8, which is the only non-technical participant with an educational background was more critical about the setup and rated the system much lower on the SUS scale. This is a finding which requires further investigation in future studies, because with VRBox we aim to support explorative learning, which should center around providing a positive user experience and usability for novice users in order to support the learning experience. Thus, the selection of participants should focus more on those users and also on users of the respective target groups, depending on the implementation (in our case for example students and teachers for the terraforming use case). One limitation regarding the interaction is a missing locomotion feature. Having the option to move through the virtual environment from a first-person perspective would increase the immersive and explorative capabilities of the system even more and will be integrated for future studies.

CONCLUSIONS AND FUTURE WORK
In this paper we presented the setup and implementation of VRBox, an augmented real sandbox that is used in a VR environment and combined with gesture interaction. With this prototype, we explored the integration of rich haptic interaction with natural materials into virtual reality and present a solution for tracking and presenting real physical materials in VR. We evaluated our system in a qualitative expert study with nine participants and found strong support for our assumption that VR increases the immersive and exploratory aspects of augmented sandboxes, leading to high levels of self-perceived creativity and playfulness. We further contribute application scenarios and possible use cases in gaming and 3D modeling. VRBox will be made public as an open source project in order to facilitate further development and replication as we see high potential in our system.

We conclude that VR is a very beneficial addition to augmented sandboxes and supports immersion, participation, and control. Using free hand interaction, users can rely on their natural ways of expression without having to interact through hardware. There is a discrepancy between the tactile sensation and the visual appearance, especially because of the blocky world style. However, as our results show, users are very positive regarding this aspect and can adapt well to a little offset between what they feel and what they see. Further, the creative capabilities of our system have been highly appreciated, making it a useful tool for different kinds of tasks like modeling and gaming.

Future work will include improvements in the tracking quality and visualization detail by adding higher resolution modes, more robust hand tracking, and increased locomotion features where users can move through the virtual world more freely. For example, a flying mode or teleporting features could be implemented as well as other, more immersive locomotion options such as magical or body-centric variants, as presented by [19]. Besides the technical and interaction aspects, we plan to further explore creative and playful application scenarios and research cognitive boundaries of redirected haptics. For application scenarios we build upon our findings as a starting point and plan to develop advanced modes for terrain design and also for sculpting where an interesting area of research could investigate how to restore previously modeled structures by augmenting the visualization with a translucent “ghost” model for fast and easy re-modeling. Regarding the redirection of haptics, which is the “warped mapping between the real and virtual worlds, such that the relation between the user and the virtual space can be different from the one between the user and the real space” [5]. With redirected haptics, virtual worlds can be experienced in a different way than the actual real-world model. For example, virtual worlds can be experienced to be much larger than the actual model or the effort of moving the arms across the model could be reduced. This can be achieved by warping the virtual dimensions so that users don’t recognize the differences, which represents an interesting line of research [5, 17, 23].

ACKNOWLEDGEMENTS
This project has received partly funding from the European Unions’ Horizon 2020 research and innovation programme under grant agreement No 688244, project first.stage and by the German Federal Ministry of Education and Research in the grant program “Erfahrbares Lernen” (experienceable learning).

REFERENCES
   http://dx.doi.org/10.1145/1347399.1347421


Examining Design Choices of Questionnaires in VR User Studies

Dmitry Alexandrovsky*1, Susanne Putze*1, Michael Bonfert1, Sebastian Höffner1, Pitt Michelmann1, Dirk Wenig1, Rainer Malaka1, Jan David Smeddinck2

1 Digital Media Lab, University of Bremen, Germany
2 Open Lab, School of Computing, Newcastle University, UK
{dimi,sputze,bonfert,shoeffner,pitt,dwenig,malaka}@uni-bremen.de Jan.Smeddinck@newcastle.ac.uk

ABSTRACT
Questionnaires are among the most common research tools in virtual reality (VR) user studies. Transitioning from virtuality to reality for giving self-reports on VR experiences can lead to systematic biases. VR allows to embed questionnaires into the virtual environment which may ease participation and avoid biases. To provide a cohesive picture of methods and design choices for questionnaires in VR (INVRQ), we discuss 15 INVRQ studies from the literature and present a survey with 67 VR experts from academia and industry. Based on the outcomes, we conducted two user studies in which we tested different presentation and interaction methods of INVRQs and evaluated the usability and practicality of our design. We observed comparable completion times between INVRQs and questionnaires outside VR (OUTVRQs) with higher enjoyment but lower usability for INVRQs. These findings advocate the application of INVRQs and provide an overview of methods and considerations that lay the groundwork for INVRQ design.

Author Keywords
Virtual reality; VR; user studies; in-VR questionnaires; inVRQs; research methods.

CCS Concepts
• Human-centered computing → Virtual reality; HCI design and evaluation methods; Empirical studies in HCI; User studies;

INTRODUCTION
The notable rise of a new generation of virtual reality (VR) systems in recent years opened up new methods and interventions for researchers across many different areas. These range from highly immersive stimulus-response studies [36, 60] over spatial navigation [147, 175] and embodied cognition [149, 154] to exposure therapies [47, 56, 146], exercising [165, 202], education [18, 104], work collaboration [7, 97], and other forms of social interaction [5, 99]. Typically, mid- and post-experience measures are collected via subjective responses on questionnaires [103]. Furthermore, the development and evaluation of VR experiences for entertainment or serious purposes also frequently relies on questionnaires. To fill out conventional paper- or computer-based forms, the subjects need to take off the head-mounted display (HMD) and change the domain from virtual to physical reality [83]. This often leads to temporal disorientation and loss of sense of control [91]. Accordingly, questionnaire results are likely biased to a degree that is difficult to quantify and likely varies from case to case. Such undetermined bias is highly problematic for many types of research and evaluations.

In contrast to the physical domain, alternate reality technologies allow for the embedding of questionnaires directly into the environment. While the transition from VR to answering, for example, paper-based questions presents a drastic change of context, embedding question-items in VR offers an opportunity to stay closer to the context of an ongoing exposure than out-of-VR research setups and avoid a break in presence (BIP) [83, 137]. Especially for measures where the self-reporting needs to be administered as soon as possible after the treatment and may be especially sensitive to differences in study setups, such as common measures for presence [155], immersion [82] or flow [35, 183], it appears crucial to give careful consideration to the interaction modalities around delivering question-items and gathering responses in order to foster the adequate interpretation of individual research outcomes and for fostering replicability.

Schwind et al. observed a higher consistency of self-reported presence when administering questionnaires in VR. The authors highlight that the effects of using questionnaires in VR are unclear, pointing out that the commonly applied measures were not validated for VR studies [159]. These considerations motivate our investigation on questionnaire practices in contemporary VR user research. We investigate whether authors employ comparable terminology and reflect their choices with regard to questionnaire presentation and response collection mechanisms. To date, VR user research does not have a shared range of common administration procedures, well-defined classification schemes, or standardized toolkits for presenting questionnaires in VR user studies that could guide
such considerations and make it easier to contrast against other work.

To build an understanding of the current practices and to lay the groundwork for future efforts around developing practical toolkits, we based this work on the following research questions:

RQ1 *How are questionnaires and individual question-items applied in contemporary VR user research?*

RQ2 *What are the considerations around – and possible advantages of – administering questionnaires in VR?*

RQ3 *What are appropriate design choices for presentation and interaction with in-VR questionnaires?*

To approach the research questions, we conducted: (i) a literature review of 123 publications on VR user studies, (ii) a survey with 74 VR experts, (iii) a preliminary design study to compare different presentation and interaction methods, (iv) a user study \( (n=38) \) of a revised iteration of our in-VR questionnaire tool, comparing it to a on-screen response system. This aggregated examination of the literature review and expert survey allows us to extract a detailed picture of proceedings of VR user research, compensating for incomplete reporting in publications and for sampling effects [123]. Based on these insights, we iteratively implemented an in-VR questionnaire tool and applied it in 2 user studies to investigate design options and effects of questionnaire administration in VR.

This work focuses on self-reporting methods. Other more objective measures (e.g. behavioural or biosignals) provide more reliable data [79] and are less likely to be affected by BIPs. We also did not take qualitative analyses into account since the procedures differ notably from quantitative studies. By providing a coherent survey of questionnaire methods in VR through a literature analysis, expert interviews, and an empirical investigation of in-VR questionnaire (iNVRQ) designs, our research can inform the discourse around questionnaire methods in VR research. Based on our research and also inform the design and implementation of practical questionnaire toolkits that are relevant to both research and industry.

**STATE OF THE ART**

Due to its immersive nature and a wide variety in technical setups, VR with HMDs requires careful deliberation by researchers aiming to conduct studies with human subjects. In this section, we review methods and practices for questionnaires in human-subject research followed by a consideration of the VR-specific technicalities around moving between worlds and their effects on question asking.

**Questionnaires in Human Subject Research**

Questionnaires are an important source of information for evidence-based research [12, 45, 103]. They embody self-reports and therefore gather the participants’ subjective experiences [45]. Question types in surveys can be divided into unstructured and structured questions [144]. Structured questions allow for a clear classification of the responses (see [152] for a more detailed discussion), while unstructured or open-ended questions allow the subjects to respond freely. This type of question is more exploratory and allows for a broader understanding of phenomena [103], while also requiring more effort from the respondents. Survey methodologies received much attention in the literature and their advantages or drawbacks are widely explored [13, 17, 163]. Reliable (consistency of the measurement) and validated (measuring the right construct) questionnaires are vital for reproducible and consistent research [17]. Choi and Pak [32] list 3 groups of potential biases: design of the question, questionnaire design, and administration. Question design covers the effects of poor wording, such as double-barreled questions, negative phrasing or wording that enforces choices [12, 32, 103]. Biases of questionnaire design are due to formatting and length of the surveys [4, 32] as well as length and structure of the questions [23]. Context-dependent forgetting [1, 58] due to environment change [136] biases the responses. We argue that especially in immersive scenarios, a series of random errors can be minimized through consistent administration of questionnaires. Notably, these considerations on questionnaire design are typically contextualized against paper-based or screen-based questionnaires not considering aspects around BIPs or switching between different realities [159].

**Moving Between Virtual and Physical Reality**

When individuals are deeply engaged with an activity or absorbed in a virtual environment (VE), they completely block out the world around them [35]. Brown and Cairns [26] identified 3 levels of immersion in games as a scale of involvement: (i) *engagement*, the lowest level of immersion, (ii) *engrossment*, when players become emotionally affected by the VR and (iii) *total immersion* where players are detached from reality. This phase is also associated with empathy for the characters in the game and transfer of consciousness [150] and is linked to the state of flow [35]. A sudden interruption or transition between realities can invoke negative feelings and affect the emotional state [91]. Accordingly, assessing presence during immersive experiences results in more reliable measurements [21, 49]. In contrast to immersion, presence is a state of mind, describing the feeling of being part of the VE [82, 201]. Presence relies on involvement and immersion [201]. When “returning” from a state of presence in VEs, a real world task is impaired to the degree of immersion and one’s ability to re-engage with the “real world” is decreased [82, 171].

Thus, leaving the VE is likely to interrupt the presence perception. Schwind et al. [159] investigated the effect of filling out a questionnaire directly in VR. They replicated their lab in VR and asked participants to fill out presence questionnaires in physical reality and in VR after exposing participants to a VE at varying degrees of realism. Schwind et al. could not find significant differences of presence between the 2 questionnaire modalities. However, the data revealed a lower variance and, thus, higher consistency of the data when the questions were answered in VR. This is in line with evidence from the literature that support the assessment of questions in VR [83, 91].

Frommel et al. observed a significant effect on presence when questionnaires are integrated in the game context [53]. Similarly, Shute discussed how to embed questionnaires into games without disturbing the game flow [169]. These considerations
can guide the choice of adequate levels of embeddedness, and a general decision whether to remain in VR or to exit VR when asking study participants to respond to questionnaires or individual question-items.

LITERATURE REVIEW

For our literature review we searched in digital libraries [10, 80] that host proceedings of high quality conference series about human-centric methods and VR applications. Namely, we scanned ACM CHI, ACM CHI PLAY, ACM VRST, IEEE VR and IEEE 3DUI for the years 2016–2019. We deliberately chose this timeframe to gain consistent insights on contemporary research methods and the release of the HTC Vive in June 2016 [78] substantially changed HMD use in VR user studies. The search included publications until July 2019, considering papers categorized with any of the keywords “virtual reality”, “head mounted display”, “virtual environment”, “user study” and “questionnaire”. Only papers with abstracts that mention a VR application with HMDs and explicitly mention some form of evaluation with users or empirical user study were added to the list.

In total we reviewed and evaluated 123 research papers, see Table 1. The papers are organized following whether they use questionnaires for measurement or not. We also categorized the different representations of questionnaires: in-VR questionnaires (INVRQ), out-VR questionnaires (OUT VRQ) and those who did not include a report of the presentation method.

Analysis

To investigate common procedures of questionnaires in VR user research, we analyzed both the system design and the study design. We paid special attention to the treatment of questionnaires or individual question-items and to the description regarding transitions between VR and physical reality. Further, we looked at the nature of the VR experience. These factors were considered as discriminatory for VR user research from classic, less immersive interventions. Each paper was examined by 2 of the authors. Disagreements were resolved in discussions.

Uses of VR Hardware

Established desktop VR devices, such as the HTC Vive (63) and Oculus Rift (42), are most commonly used in the papers reviewed followed by mobile device powered HMDs, e.g. Samsung Gear VR (9) and Google Cardboard (4). 3 publications did not report the device used. The input modalities and devices used in the VR applications are mainly native VR controllers (e.g. Oculus Touch, HTC Vive Controller) (49) followed by freehand interaction (e.g. Leap Motion or Microsoft Kinect) (35) and general purpose input devices (e.g. game controller, keyboard, mouse, stylus, smartwatch, and touch screen) (25).

Questionnaire Assessment

110 out of 123 papers report having used questionnaires in their VR user studies. Since the use of VR devices entails design decisions regarding the presentation of questionnaires and individual question-items, we surveyed the documentation of such decisions in the respective papers. 77/110 do not report how they presented the questionnaires to their users. 13 papers report that the participants filled out the questionnaire after leaving the VE but do not describe whether they used paper-or screen-based questionnaires. 15 papers report on the usage of INVRQs – either for the whole question asking procedures in the user-study (3) or in combination with OUT VRQs (12).

Cases of In-VR Questionnaires

15 papers report the use of INVRQs. Some describe the design in more detail. Figure 1 depicts 7 different realizations of INVRQs. Kang et al. (Fig. 1a) used a 2D heads-up display (HUD) overlay with a single question about the user’s motion perception between multiple trials in their VE [85]. The user interface (UI) shows a single question with a multi-line question text and 3 buttons for answering choices. As input device, they used a native VR controller [85]. Schwind et al. (Fig. 1c) included the full 32-item PQ [201]. The participants stayed in the VE for the whole duration of the study: on average 58.6min [160] and 75min [158]. The authors designed a 3D floating UI which appears in front of the subjects showing a one-line text instruction and 4 items on 7-point Likert-scales. Users select answers and navigate the questionnaire with freehand gestures using a Leap Motion [158, 160, 161]. In another study, Schwind et al. (Fig. 1f) placed single questions on presence on a virtual PC in the VE, with which a user interacts with the trackpad of a VR controller [159]. Oberdörfer et al. (Fig. 1d) presented the NASA-TLX [72] using a virtual world-referenced representation of the paper-based version. The users interacted using a VR controller with pointing [127].
Figure 1. Examples of different realizations of INVRQ: (a) and (b) present the questionnaire using a HUD, (c)-(f) use a world-referenced questionnaire, and (g) presents the questionnaire attached to the body.

Wienrich et al. (Fig. 1g) presented a body-referenced INVRQ [198]. The questionnaire was displayed on a 2D floating UI with a 20-items PANAS [96] attached to the hand of the virtual character. They combined the in-experience measurement with further out-VR measurements and oral answering of the Fast Motion Sickness Scale (FMS) [88]. Fernandes and Feiner (Fig. 1e) is the earliest example of INVRQs in our sample. The authors applied a 10-point Likert-scale slider on well-being where subjects could stop the experiment by selecting the maximum value of discomfort [44]. [28] (Fig. 1b), [6] and [66] adapted this method in their works in different realizations.

All INVRQ designs differ in their presentation (HUD, world-reference, body-referenced), their extent (single-question vs. multi-item questionnaire), question-item presentation (text-based vs. scales) and interaction modality (pointing, free-hand, trackpad). Table 2 summarizes the variation of the INVRQ designs. The applications of INVRQs cover questions about the subject’s well-being, their sense of presence and task-specific questionnaires, e.g. about task workload or affect.

Discussion
Our literature analysis shows that comprehensive reporting of questionnaire usage in VR research is frequently neglected. Only Schwind et al. [159] compare INVRQs and OUTVRQs and discuss their effects. This indicates that the field may benefit from building awareness and providing guidelines. We identified 15/123 cases of INVRQs applied in VR user studies. The realizations differ substantially in their presentations and interaction methods. The majority of cases used presentations that contextualized the questionnaires in the VE, either attached to the user, or anchored in world space in a stationary manner. The participants used predominantly native controllers to interact in VR, directly followed by free-hand interaction. In order to better contextualize these literature findings and to collect details about the design of INVRQs and their potential shortcomings, we discuss a further investigation through an online expert survey in the following section.

EXPERT SURVEY
To augment the insights gained from the literature review, we conducted an expert online survey that evaluates general proceedings of VR user studies and attitudes towards INVRQs. With this additional analysis of the state of the art, we aimed to capture an impression of the actual procedures employed by the researchers independent from possible biases present in publications with space limitations.

Survey Dissemination and Pre-Processing
We developed a custom survey to capture the general reporting of proceedings of VR user studies as well as experiences and attitudes of the study designers regarding INVRQs. The complete survey is provided in an OSF project1. Following informed-consent, it consisted of 22 questions grouped into 5 categories (demographics, general research practice, VR research practice, INVRQ experience and OUTVRQ experience). The survey was designed and distributed using Google Forms. For recruiting expert participants, we extracted a list of authors from the papers (2016–2018) we analyzed in our literature review and sent them personal invitations. We also advertised the survey via social media channels. Over 6 weeks in July and August 2018, we collected 74 replies.

Since the online survey focused on researchers and VR experts, we excluded 4 participants who were not directly involved in VR user studies in the last 24 months and 3 participants who indicated that they do not generally use questionnaires as measures in their VR user studies. Moreover, we corrected obvious spelling mistakes to facilitate accurate counts of established terms (e.g. hardware or questionnaire names). For anonymized analysis, we removed time stamps and added unique identifiers (E1–E74). The reported analysis focuses only on questions that are most relevant to our research questions.

Analysis
After data cleaning, the analyzed data set consisted of 67 full set responses. Based on participant indication, we sampled

1https://osf.io/f5qy7/
responses from 13 different countries of residence (Q3). With 20 participants from Germany, 9 responses each from the USA and Portugal and 4 responses each from Denmark and France, the majority of our sample came from northern-hemisphere countries. As a result of the sampling strategy, the expert sample stems predominantly from academia. The participants indicated to be holding the following positions (Q2): 22 Ph.D. students, 16 professors (full, associate, assistant or equivalent), 10 undergraduate students, and 8 Postdocs. The remaining 11 participants provided individual answers, such as research engineer or research fellow. On average, the participants indicated to be rather experienced with designing user studies (Q4). On a 6-point Likert-scale (0 to 5): $M=3.76$, $SD=1.18$. The majority (56) were directly involved with conducting 1–5 VR user studies in the last 24 month while 8 experts indicated involvement with between 6 and 10 studies. Only 3 indicated more than 10 VR user studies (Q9). Our sample consists of a diverse spectrum within the group of academic researchers who have a sound background on VR user research allowing us to interpret the responses as an expert evaluation.

**VR research**

The experts’ most commonly used VR devices (Q10, multiple choice) in the last 24 months were desktop VR devices (HTC Vive (55) and Oculus Rift (37)) or mobile VR devices (Samsung Gear VR (15) and Google Cardboard (11)). Other devices were mentioned occasionally. Only 4 participants used neither HTC Vive nor Oculus Rift. The most commonly used input modalities (Q11, multiple choice) reported for the general interaction with VR are mainly native VR controllers (50) and freehand interaction (e.g. Leap motion or Microsoft Kinect) (31) followed by gaze interaction (19) and general purpose input devices (e.g. game controller, keyboard, mouse, stylus or smartwatch) (34). 6 participants reported using custom controllers that were not further defined.

**Questionnaires**

We asked our participants how they usually present questionnaires in their user studies. 29 use a separate screen outside the VR, 21 paper questionnaires, 6 embed the questionnaire in VR, 2 use oral answers and 9 use mixed methods (Q13). Validated questionnaires (OUTVRQs) are very common in VR user research (Q12). Independent of the form of presentation, the experts reported they encounter some difficulties with questionnaires in VR user studies (Q16): 14 experts reported problems with the questionnaires in general, such as ambiguous question items in validated questionnaires, e.g. PQ, or the length of questionnaires, e.g. SSQ [87], NASA-TLX [72], especially for use in measurements between trials. 7 experts reported problems of their participants when moving from VR to reality for answering the questionnaires, namely the lack of immersion (4) and the temporal effort (3).

29/67 experts have tried in-VR questionnaires (INVRQ users). Most of them (25) tried fully interactive questionnaires that are both presented and answered in VR (Q18, single choice). All experts rated the usefulness of INVRQs (Q17) on a 6-point Likert scale (0–5). They rated the usefulness with $M=2.97$, $SD=1.58$ which is significantly higher than the midpoint 2.5 (one-sided independent sample t-test $t_{65}=2.41$, $p<.01$). For further analysis, the experts were split into 2 sub-groups according to whether they have tried an INVRQ in their research setups (Q18) or not ($M_{INVRQ}=3.72$, $SD_{INVRQ}=1.33$ and $M_{NoINVRQ}=2.39$, $SD_{NoINVRQ}=1.55$), cf. Figure 2. A two-sided independent sample t-test showed a significant difference between prior INVRQ experience and the usefulness rating ($t_{65}=3.69$, $p<.01$, Cohen’s $d=.45$). To appropriately weight the responses we further split the experts according to their scientific seniority. Group A: post-docs and professors, Group B: all other participants (presumably more junior researchers). The experts in group A tend to rate the usefulness higher than the other participants: With $M_{A}=3.46$, $SD_{A}=1.74$ and $M_{B}=2.70$, $SD_{B}=1.46$. But the difference was not significant $t_{65}=1.91$, $p=.$.06.

**In-VR Questionnaires**

14/29 INVRQ users presented the questions in world space as floating displays and 4 used a 2D UI (probably replacing the virtual world) (Q19). 4 experts responded that they presented only 1 question at a time. 14 experts did not observe any difficulties of their participants using INVRQs (Q20). Usability and the “amount of time for filling out the questionnaires” (E44) were major challenges experts came across (4× each). Further, “completing long questionnaires can be (...) annoying” (E18) and participants are “impatient (…)” and hence made mistakes in the scaled questions” (E5).

**Out-VR Questionnaires**

Using Q18 as a branching question, 38 OUTVRQ users were asked to provide reasons for not embedding the questionnaires into the VE (Q21). 19 from that group reported that they do not suspect to benefit from INVRQs. 9 experts reported technical difficulties and 3 exceeding time effort in setting up the embedding for INVRQs. “Time effort to embed into VR much higher than doing the same on paper and harder to change/make additions” (E48). 6 experts mentioned usability problems and 1 sees problems in all these categories. The experts explained their critical opinions of INVRQs further (Q22): They argue, it would reduce the “willingness to answer the questions” (E31) because interaction with the UI in VR is more frustrating and takes longer than on desktop. Especially text entry is seen as a major issue. The experts reason that this might bias the results “towards the choices that are easiest to make through the interface” (E31). Additionally, they mentioned the time effort for implementation and usage:
“My last study took 45 minutes including questionnaires. It gets very hot and inconvenient under a Vive for such a long time – especially if it’s your first time” (E71).

Discussion
The responses to the survey indicate that experts generally appear to have a positive attitude towards INVRQs. Both our literature review and expert survey demonstrate that there is no common standard for using questionnaires in VR user studies. While the reporting of the presentation type of the questionnaires in the literature is mostly imprecise, our expert survey did not surface a clear standard method for presentation. Thus, regarding RQ1, the state of the art for delivering questionnaires in VR user studies is quite heterogeneous. Also, we found inconsistencies between the literature analysis and the online survey: While only 15/123 papers reported on the usage of questionnaires embedded in VR, 29/67 experts reported that they have tried INVRQs before. Possible reasons for the contrast between the positive attitude of the INVRQ user group and the rare reporting of INVRQs are not intuitively clear. The difference could indicate that a shift may be taking place in the community with many applications of INVRQs not having been published yet. There is, however, the possibility of increased variance due to publication bias that warrants control in future work. Alternatively, the authors of the analyzed corpus may have considered the presentation type to be irrelevant. In contrast to these outcomes, there are experts who have a strong opinion against embedding questionnaires and argue that INVRQs could bias responses. In response to RQ2, the broad range of reasons provided for or against using INVRQs indicate technical challenges, implementation effort, and fear of biases and participant overload.

In the following section we investigate these objections from the users’ viewpoint and design and evaluate an INVRQ tool.

DESIGN STUDY
Although many guidelines from “traditional” HCI, cf. [41, 168], can be applied to VR, the third dimension brings its own challenges for UI designers. LaViola et al. [102] point out that there are no standards for 3D UIs yet, such as the ISO 9241-110 [81] standards for 2D GUIs. The results from the literature review and from the expert surveys are heterogeneous and do not allow for a definite answer of the usefulness and realization of INVRQs. With regard to RQ3, the goal of this design study was to identify the most usable interface design.

In-VR Questionnaire Design
Based on the results of our expert survey and general guidelines on traditional UIs [41, 168] as well as UIs in VR [40, 62, 128, 203], we implemented 4 variants of INVRQs. The interfaces differ in 2 dimensions: anchoring (world-referenced and body-referenced) and interaction modality (pointer and trackpad). The world-referenced UI is anchored in the VE and users can freely move around the interface. A body-referenced UI is attached to a hand-held controller.

To enhance readability, we applied guidelines from research and industry which recommend signed distance field fonts [61, 128]. In line with Dingler et al. [40], we used light glyphs on a dark background. Oculus best practices recommend to avoid HUDs (cf. Fig. 1a, 1b) in favor of UI elements that are settled in the VE, as it overcomes the binocular disparity and allows to contextualize the UI in VR [102, 128]. This is in line with the statements from the experts who applied INVRQs, as the majority (32/67) implemented questionnaires in the world space (see Fig. 1c–1f). In contrast to world-referenced UIs, body-referenced interfaces, as in Fig. 1g, take advantage of the user’s proprioceptive sense and can significantly enhance the interaction with the UI [102, 118].

The interaction with the UI varies between a laser pointer – as Oberdörfer et al. [127] applied for their INVRQs – and a clicking interaction where the trackpad of the controller is used to navigate through the questionnaires, similar to Schwind et al. [159]). Oculus guidelines [128] suggest a laser pointer with a visible ray-cast and a cursor projected on the UI as an appropriate and intuitive method to interact with UI in VR. This is reflected in general guidelines which suggest better performance in terms of speed, accuracy and cognitive demand [112, 113, 119]. In contrast, navigating a UI using a trackpad promises to be more efficient when the UI is close to the users since they are not required to twist their arms for aiming.

The 4 designs that emerge from the 2 dimensions anchoring (world, body) and interaction (pointer, trackpad) cover a wide range of designs that are applicable to INVRQs. We developed all 4 designs iteratively following the same usability guidelines and paying particularly attention for comparability. Our interface supports continuous values (slider), checklists, radio lists, drop downs; and switches. The prototype is implemented in Unity3D with OpenVR. Their interaction and design is demonstrated in the accompanying video figure.

Study Design, Procedure and Tasks
The study has a 2×2 within-subject design where users filled out INVRQs using the 4 different versions of the interface: world-pointer (WP), world-trackpad (WT), body-pointer (BP) and body-trackpad (BT). The order of the condition was counterbalanced using Latin Square. First, the subjects were welcomed and informed about the study. The experimenter fit the HMD (HTC Vive) and explained the interaction in the current condition with the native controllers. To provide a context to the VE, we used a sci-fi scene.

To investigate the usability of all response types, we developed a questionnaire which asks for common knowledge facts to ask subjects easy-to-answer but objective questions to calculate correctness. Each condition consisted of a questionnaire which comprised all 5 question types once. Each participant answered 20 questions in total with each question type once per condition. After each condition in VR, the participants took off the HMD and filled out a paper-based System Usability Scale (SUS) [25]. When the participants finished all 4 conditions, the experimenter asked them to put the 4 interfaces into a ranked order and conducted an interview.

Participants
10 male (age M = 29.9, SD = 2.9) subjects from a game jam at the campus participated in the study. For their experience in VR and as VR developer or researcher, the participants gave...
a score on a Likert scale from 1 (no experience) to 5 (high experience). All participants had prior experience in VR and game development ($M=3.6$, $SD=.84$). 6 participants developed VR applications or conducted studies in VR ($M=2.3$, $SD=1.33$). Although this sample lacks representative diversity, it allows for a deep discussion of the design space while avoiding general issues that could have come up with novices.

### Results and Discussion

For the quantitative metrics (SUS [25] and duration), we conducted a repeated measures ANOVA with condition as factor. The descriptive statistics and the results of the analysis are in Table 3. Bonferroni corrected pairwise comparisons of the SUS [25] scores revealed significant differences between BP and WT ($p<.01$), between BP and WP ($p<.01$), and between WP and WT ($p<.01$). For the required time, Bonferroni corrected pairwise comparisons revealed significant differences between BP and WT ($p<.01$), BT and WP ($p<.01$), and between WP and WT ($p<.05$). On the subjective rankings, the participants rated the world-pointer setup significantly higher ($F_{3,36}=13.27$, $p<.001$).

6/10 subjects stated that the laser pointer is easy and intuitive to use. In contrast, the majority perceived the trackpad as tedious and confusing. 3/10 participants liked being able to move the questionnaire. But 4/10 participants reported they were confused by the body-referenced interface and stated the movement as unpleasant. The world-anchored UI was stated as less demanding than the body-referenced (3/10).

Completion time, SUS scores [25] and the interviews clearly show the world-referenced anchoring with the laser pointer interaction is easiest to use and therefore the best candidate to investigate the concerns raised by the experts.

### USER STUDY

To assess whether usability and duration concerns about INVRQs raised by the experts hold true and to provide guidance on question-asking methods in VR, we conducted a user study in which the participants shot balloons with bow and arrow in a VR archery game and then filled out INVRQs as well as questionnaires presented on a notebook (OUT VRQ). The aim of this study was to evaluate adequate design choices of INVRQs we identified in the design study and to capture the users’ perspectives on INVRQs. In contrast to previous work by Schwind et al. [159], our user study applies the questionnaires in a realistic study setting without replicating the lab space virtually.

### The Questionnaire Tools

We implemented the INVRQ with world-based anchoring and laser pointer interaction (Fig. 3). We refined the overall readability, i.e. font size, contrast, spacing and positioning of the UI in the VE. The tool supports sliders, radio lists, radio grids and check lists. We omitted drop downs and free-text fields since these elements rarely appear in standardized questionnaires, as the literature review shows. If needed, drop downs can be represented using a radio list, and free-text input can be approximated with voice recording.

The OUTVRQs were realized using the questionnaire tool LimeSurvey [57]. It was presented on a 15” notebook with external keyboard and mouse with default speed.

### Measurements

After the archery game, we measured presence in the VE using IPQ [155] on a Likert scale with the subscales general presence (GP), spatial presence (SP), involvement (INV) and experienced realism (REAL). Furthermore, we asked the participants to rate the game and the perceived control over the bow on a 10-ticks slider. For demonstrating a greater variety of question types, we additionally included questions about the VE (1x numerical, 4x single choice with 2–5 items, 4x multiple choice with 5 and 16 items).

To evaluate the workload and the usability of the INVRQ tool, we employed the raw NASA-TLX [72], using a 20-ticks slider to measure physical, cognitive and temporal demand, as well as performance, effort, and frustration. Usability was measured with the UMUX [46], a four-item questionnaire providing comparability with the SUS [25] ($r=.96$). As an objective performance metric, we logged the exact time. To get detailed insights from the users’ perspective, we conducted a semi-structured interview at the end of a session.

### Study Design and Procedure

The study compares the usage of INVRQs to digital OUTVRQs. To provide a realistic study setting, we designed a

---

**Table 3. Descriptive statistics and RM-ANOVA of the design study for the 4 conditions world-pointer (WP), world-trackpad (WT), body-pointer (BP) and body-trackpad (BT) on SUS [25] and time.** $^*p<.01$.
balloon archery game as an immersive VR experience (see Fig. 3). We chose this task because (i) it is engaging and requires the participants to focus, promising an immersive VR experience, (ii) the interaction is easy to learn but differs from the pointing interaction used for answering the INVRQs; thus, the task is less likely to produce carry-over learning effects. For the archery task, we used free Unity3D assets and implementations from the SteamVR Interaction System. The VE consists of a round platform with 3 pillars that display instructions that guide through the experiment. The platform is surrounded by 12 spawn points for the balloon targets and it is situated in a realistic environment with mountains, trees, a river and high resolution textures. The questionnaire interaction builds on the laser pointer by SteamVR and the GUI interaction by HTC Vive.

The study followed a within-subject design with the conditions INVRQ and OUTVRQ in randomized order. After the participants were informed about the procedure, they signed a consent form. The experimenter fitted the HTC Vive Pro HMD. Then the participants played a tutorial round in which they had to hit 5 balloons followed by a 90 s round where they should hit as many balloons as possible. After the archery task, the participants filled out questionnaires using the corresponding tool for their first condition (INVRQ, OUTVRQ). Afterwards, the participants repeated the game and questionnaires using the other questionnaire method. We encouraged them to take a 2 min break in between. After both conditions, the experimenter conducted a semi-structured interview followed by a paper-based demographics questionnaire. Finally, the participants were orally debriefed. The study, including game, questionnaires and interview, was conducted in German and took around 45 min (∼11 min in VR).

Participants
We advertised the study on campus, social media and in lectures and conducted it in July 2019. In total, 38 participants (age: \(M=27, SD=10.8\); 16 f, 22 m, 0 other; 20 started with the INVRQ condition) volunteered for our study. Most participants were students. 21 participants used vision aids in VR, 1 participant has a light dyschromatopsia. The sample has a broad range of prior VR experience: 6 participants use VR regularly, 27 occasionally and 5 never used VR before. 19/38 participated in other VR user studies previously and 3 participants used INVRQs before. We detected no outliers regarding demographics and task completion times.

![Figure 4. Ratings of usability on UMUX [155] (left) and of completion times (right) for both conditions.](image)

**Results**

**Performance, Presence and Rating of the VR Experience**

On average, participants shot 24.71 balloons \((SD=10.33)\) in the first round and improved by 6.34 \((SD=6.01)\) balloons in the second round. We obtained presence on the IPQ \([155]\) on a 7-point Likert scale (0–6). To determine if the measures deviate from neutral, we performed a two-sided one-sample t-tests against the midpoint 3. The results show a positive difference for GP \((t_{37}=14.93, p<.001)\), SP \((t_{37}=22.53, p<.001)\) and INV \((t_{37}=5.99, p<.001)\), but no difference for REAL \((t_{37}=-.48, p=.63)\). There was no significant effect of condition on any IPQ subscales (see Fig. 5a). Like Schwind et al. \([159]\), we performed a t-test to compare the variances of both conditions on all IPQ subscales; the differences were not significant \((GP_{var}: t_{37}=-.52, p=.60, SP_{var}: t_{37}=-.05, p=.96, INV_{var}: t_{37}=1.47, p=.15, REAL_{var}: t_{37}=.17, p=.87)\).

Further, the participants rated the interaction with the bow \((M=7.89, SD=1.70)\) and how they liked the game \((M=8.0, SD=1.73)\) on a 10-ticks analog scale from 1 (not at all) to 10 (very much). A one-sample t-test revealed significant differences against the midpoint \((5.5)\) for bow control \((t_{37}=15.01, p<.001)\) and game ratings \((t_{37}=14.86, p<.001)\), but no effect of conditions neither for bow control nor game rating.

**Duration and Self-reports on Usage of the Questionnaires**

Fig. 4 shows plots of UMUX scores and answering time. On average, participants required 6.77 min \((SD=2.69)\) to fill out the questionnaires in VR and 6.30 min \((SD=2.26)\) on the notebook (without the time for taking off the HMD). There was no significant difference of condition on duration of filling out the questionnaires \((t_{37}=1.05, p=.29)\). Nevertheless, in the post-experiment interviews, 5 participants perceived filling out the questionnaires in VR as faster. 2 had the impression the questionnaires in VR would be shorter. The IPQ completion times did not differ significantly \((INVRQ IPQ: M=146.39 s, SD=65.46)\) in comparison to the ones reported by Schwind et al. \((VR IPQ: M=146.94 s, SD=63.2)\) \([159]\): \(t_{37}=0.4, p=.97\). On the UMUX, the participants rated the questionnaire tools’ usability. Both systems were rated positively \((INVRQ: M=77.35, SD=18.35, OUTVRQ: M=86.21, SD=9.4)\). The difference between the conditions was significant \((t_{37}=2.82, p<.01)\). In accordance with Grier \([65]\), the participants experienced low to medium workload on the NASA-TLX \([72]\) \((INVRQ: M=18.64, SD=11.38, OUTVRQ: M=14.40, SD=9.42)\). Paired t-tests revealed significant differences on physical demand \((t_{37}=4.14, p<.01)\) and effort \((t_{37}=3.00, p<.01)\) subscales but not on mental demand \((t_{37}=1.42, p>.05)\), temporal demand \((t_{37}=1.15, p>.05)\), performance \((t_{37}=1.48, p>.05)\) or frustration \((t_{37}=1.76, p>.05)\). The results are depicted in Figure 5b. We contrasted the mean TLX-scores against Schwind et al.’s values \((VR: M=33.16, SD=20.96, PC: M=37.77, SD=19.26)\) \([159]\). Our data show a significantly lower workload for both corresponding comparisons \((INVRQ (t_{37}=3.50, p<.01)\) and \((OUTVRQ (t_{37}=6.57, p<.01)\) conditions.

**Qualitative Results**

We collected relevant statements from the interviews and dynamically generated categories emerging from the material.
We indicate how many participants agree with the central statements made by others. Exact quotes were translated by the authors and are labeled with an ID in parentheses.

The majority of the participants (27/38) stated their VR experience as “fun” or “enjoyable”. 17 found the inVRQs easy to use and the interaction intuitive. 4 participants preferred the directness of input with the VR controllers. However, 16 reported that sometimes the UI did not respond to their input; 4 described this as frustrating. 26 referred to the outVRQs as the “common” and “normal” questionnaires. 4 participants reported higher effort using the mouse and 12 participants would have preferred a faster mouse movement in OUT VRQ.

Although the participants highlighted the advantages of familiarity using the mouse, we observed that all users instinctively understood the pointing interaction.

17 users criticized changing from VR to physical reality and stated that not having to switch the system feels more fluent and “better integrated” (P12). 4 users addressed the BIP due to the change of medium and how it might affect data quality.

Regarding presentation, 31 participants confirmed that readability and font size of the inVRQs were good and the questionnaire canvas was positioned well. P34 disliked not having a full overview of the displayed content in contrast to the notebook screen. 13 users mentioned that blurred edges in the HMD required hand movement for reading. 8 users found it strenuous to fill out the questionnaires in VR. 5 of them suggested to provide a chair. P38 sat on the floor. Concerning data validity, 4 participants hypothesized that maintaining the immersion is “better for the results” (P11). When answering how they feel, 14 participants found it beneficial to do so in VR, because they still were in the situation: “You don’t need to recall how you just felt.” (P9). Correspondingly, 8 valued the immediacy of the surveying in VR.

12 stated the inVRQs were entertaining: “I didn’t know filling out questionnaires can be fun!” (P33). Accordingly, outVRQs were referred to as “drier” (P4) or boring by 12 users, and P35 stated that a pleasant setting can be motivating to fill out the survey. At the same time, such positive excitement may influence certain measures (e.g. affect).

Finally, we asked the participants how they would like to answer questionnaires if they would have to repeat the study. A majority of 31/38 would prefer to do it in VR, 5 on a computer and 2 had ambivalent opinions.

Discussion

The high ratings on the IPQ [155] with positive differences against neutral on all subscales but REAL, the ratings of the game and the control over the virtual bow, as well as the qualitative statements indicate that the game provided a high sense of presence. This indicates that our experiment design, which intended to simulate a realistic VR user study scenario, was successful. In alignment with results by Schwind et al. [159], we could not find any differences between the conditions on presence. However, our data show no differences in consistency and we cannot confirm their findings of lower variances of presence when surveying in VR. This supports prior findings that presence questionnaires are inadequate to assess BIPs [171]. Further, literature suggests that presence should be assessed behaviourally [48, 166, 172] or physiologically [24, 39, 116, 170].

With the UMUX, we measured high usability scores for both questionnaire tools (inVRQ: $M=77.35$, $SD=18.35$, outVRQ: $M=86.21$, $SD=9.4$). However, the outVRQs were rated higher with a medium effect size (Cohen’s $d=.60$). A possible explanation was given in the interviews: The participants sometimes had to repeatedly click on a UI element of the inVRQ for selection. The UMUX score allows a comparison with the SUS [46]. According to Bangor the scores are in a highly acceptable range [14], discarding the concerns of the experts. Similarly on the TLX, we measured a comparatively low workload [65]. However, the physical demand and effort are significantly higher in the inVRQ condition. As 5 participants stated, this could be attributed to the fact that they were standing and using the VR controller in mid-air rather than sitting on a chair and using the mouse with a resting hand on
the desk. The comparable duration is consistent with the TLX ratings for temporal demand and allows to infer that answering questions in VR does not affect the duration of a user study.

Limitations and Future Work
Our research only investigates the design aspects of the anchoring and interaction modality with INVRQs. This first step was essential to establish a convenient INVRQ design that can be applied in VR user research. Due to small sample size and non-diverse participants in the pre-study our design recommendations are not conclusive. However, the high UMUX scores in the user study confirm the prior results. In future work, we aim to systematically examine further design options (e.g. HUDs, freehand interaction) at different degrees of contextualization and embeddedness as previous work showed for games [53, 169] and VR [52, 198].

We only considered a short assessment of an INVRQ (=6.5 min) at the end of 1 exposure. Future work should examine the effects of long VR exposures and the effect of INVRQs for repeated between-trials measurements. We also did not evaluate open-ended questions which allow the subjects to respond freely, but require sophisticated methods for text-entry in VR. Although such methods exist [92, 129], they are often less accurate and efficient than out-VR settings, especially for untrained users [92]. In future work, we aim to investigate text-entry and oral assessment of open-ended questions in VR as [37], [121] and 4 surveyed experts suggested. The slightly lower UMUX rating suggests room for improvements for the in-VR questionnaire components and the increased TLX physical demand and effort, together with participant comments and behaviour suggest that investigations into “middle-ground” approaches (e.g. remaining in VR but being seated) could lead to more practical solutions.

CONCLUSION
Subjective self-reports are frequently used in VR user studies and administered in the physical domain. This can lead to a break in presence [83], disrupt the immersive experience [91] and bias the responses [159]. Embedded questionnaires in the VE reinforce the association of VR and the subjective responses. Although different presentation methods of the questionnaires may affect the results, contemporary research has no shared agreement or validated assessment methods of self-reports in VR user studies. This work aggregates the contemporary body of research, VR expert perspectives and the user experience of INVRQs.

Our first research question (RQ1) investigates current applications of questionnaires in VR user research. From the literature analysis, we identified 15 instances of INVRQs. These few examples differ substantially in visualization and interaction, emphasizing the lack of validated surveying procedures in VR user research. For a comprehensive understanding of the advantages and challenges of INVRQs (RQ2), we conducted an online survey with 67 VR experts. 43/67 of researchers see the importance of embedding questionnaires directly into the VE. To explore presentation and interaction modalities suitable for INVRQs (RQ3), we conducted 2 user studies, in which we first identified world-anchoring and pointing as most adequate design choices to administer questionnaires in VR and then contrasted an INVRQ against a common screen-based OUTVRQ. Although the results show lower usability and higher physical demand and effort of INVRQs, the ratings are within tolerable range and the majority of participants stated a positive attitude towards INVRQ.

24/67 of the experts rated the usefulness of INVRQs below neutral and raised concerns regarding usability (10) and the required time for answering (5). The high UMUX-scores and comparable completion time between INVRQs and OUTVRQs defy these objections. Moreover, the majority of the participants (31/38) would choose INVRQs over OUTVRQs. This result may be partially attributed to a novelty effect of VR and might weaken with a wider dissemination of VR technology. Similarly, our online survey showed that 29/67 of the experts have already applied INVRQs in user studies and mainly consider them as useful and effective. Based on our findings from the literature review, the survey and the user studies, we advocate that presenting questionnaires in the VE helps the participants to report their experience in a convenient, non-interruptive manner.

Based on the results from the 4 presented studies we conclude: (i) researchers should to apply INVRQs in their user studies, (ii) pointing and world-anchoring are usable ways to realise INVRQs, (iii) participants prefer using INVRQs over OUTVRQs and (iv) researcher should be aware of slightly raised physical and mental demands when using INVRQs.

Similar to the establishment of standardized questionnaires that are empirically validated, we propose moving towards standardizing questionnaire implementation and presentation methods in VR, similar to a quasi-standardization that has already occurred in more traditional screen-based questionnaires due to the prevalence of selected survey tools, such as Google Forms or LimeSurvey [57]. Future work will need to seek for comparative experimental evidence on how the questionnaire modality affects the reliability of the measurements. As with most design choices, there is no absolute right or wrong. However, researchers should be aware how their measurement methods influence the data. This research lays the groundwork for a design theory of INVRQs to provide validated and standardized methods of question-asking in VR.

ACKNOWLEDGEMENTS
This project was partly funded by the European Unions’ Horizon 2020 research and innovation program (No 688244), project first.stage, Klaus Tschira Stiftung gGmbH, and by the German Research Foundation DFG, as part of Collaborative Research Center (Sonderforschungsbereich) 1320 “EASE – Everyday Activity Science and Engineering”, University of Bremen (https://ease-crc.org/). The research was conducted in subproject P01 Embodied semantics for the language of action and change.

REFERENCES


http://dx.doi.org/10.1089/cpb.2006.9.207


http://dx.doi.org/10.1145/3170427.3188629

http://dx.doi.org/10.1037/0276-4913.48.2.85

http://dx.doi.org/10.1177/1541931215591010

http://dx.doi.org/10.1145/3170427.3188629

http://dx.doi.org/10.1109/VR.2018.8395994

http://dx.doi.org/10.1109/VR.2018.8395994

http://dx.doi.org/10.1145/3025453.3025683

http://dx.doi.org/10.1145/3173574.3173628

http://dx.doi.org/10.1145/3330530.3330649


http://dx.doi.org/10.1145/3330530.3330634

[74] Tianyu He, Xiaoming Chen, Zhibo Chen, Ye Li, Sen Liu, Junhui Hou, and Ying He. 2017. Immersive and Collaborative Tai Chi Motion Learning in Various VR Environments. IEEE, 307–308. DOI:
http://dx.doi.org/10.1109/VR.2017.7892299

http://dx.doi.org/10.1145/3170427.3188629

http://dx.doi.org/10.1145/3025453.3025665

http://dx.doi.org/10.1145/3330530.3330667


[156] Peter Schulz, Dmitry Alexandrovsky, Felix Putze, Rainer Malaka, and Johannes Schöning. 2019. The


Breaking The Experience: Effects of Questionnaires in VR User Studies

Susanne Putze∗,1, Dmitry Alexandrovsky∗,1, Felix Putze2, Sebastian Höffner1, Jan David Smeddinck3, Rainer Malaka1
1 Digital Media Lab, University of Bremen, Germany
2 Cognitive Systems Lab, University of Bremen, Germany
3 Open Lab, School of Computing, Newcastle University, UK
{sputze, dimi, fputze, shoefchner, malaka}@uni-bremen.de Jan.Smeddinck@newcastle.ac.uk

ABSTRACT
Questionnaires are among the most common research tools in virtual reality (VR) evaluations and user studies. However, transitioning from virtual worlds to the physical world to respond to VR experience questionnaires can potentially lead to systematic biases. Administering questionnaires in VR (inVRQs) is becoming more common in contemporary research. This is based on the intuitive notion that inVRQs may ease participation, reduce the Break in Presence (BIP) and avoid biases. In this paper, we perform a systematic investigation into the effects of interrupting the VR experience through questionnaires using physiological data as a continuous and objective measure of presence. In a user study (n=50), we evaluated question-asking procedures using a VR shooter with two different levels of immersion. The users rated their player experience with a questionnaire either inside or outside of VR. Our results indicate a reduced BIP for the employed inVRQ without affecting the self-reported player experience.

Author Keywords
Virtual reality; VR; user studies; in-VR questionnaires; inVRQs; break in presence; surveys; biosignals; research methods.

CCS Concepts
•Human-centered computing → Virtual reality; HCI design and evaluation methods; Empirical studies in HCI; User studies;

INTRODUCTION
Recent advances of VR technology have enabled new research methods and interventions across various fields and allow for the design of highly immersive applications that evoke a strong sense of presence. Subjective responses through questionnaires remain a widely applied method for administering mid- and post-experience measures [52]. Research with – as well as the development and evaluation of – VR experiences often relies on questionnaires which are completed on paper or through computer-based forms. Thereby, participants are required to leave the VR, i.e., take off the HMD, to return to the physical domain [44] and experience a break in presence (BIP) [89]. The BIP is associated with physiological effects [85, 86], disorientation and loss of sense of control [48]. Accordingly, questionnaire results are likely biased to a degree that is difficult to quantify and may vary across individuals. Such undetermined bias is highly problematic for many types of research that are based on subjective measurements.

In contrast to the physical domain, alternate reality technologies allow for the embedding of questionnaires directly into the virtual environment (VE). Embedding question-items in the VE offers an opportunity to stay closer to the context of an ongoing experience [32] than out-VR data acquisition allows, as it avoids a drastic change of context – BIP [3, 44, 74]. Schwind et al. have shown that administering questionnaires directly in VR contributes positively to the consistency of self-reports on presence. They advocate for surveying participants directly in the VE [74]. However, these results were only validated for self-reports on presence and might differ for other constructs, especially since presence is a primary candidate for being affected by BIP. The publication also does not report measured BIP in relation to the outcomes.

Recent research has already started to embed questionnaires in the VE for various applications [45, 63, 76]. However, related work does not offer considerations about the extent and influence of BIPs on the VR experience, as triggered by posing questionnaires [3]. It appears crucial to investigate the side-effects of question-asking in VR user studies as researchers have to be aware of biases that may exist for their research methods of choice. In this paper, we address this gap and investigate the effects of BIPs evoked through question-asking in VR via a laboratory user study. We imitate a representative VR user study with repeated self-report measures and administer questionnaires both embedded into the VE (inVRQ) and outside of the VE (outVRQ). For our investigation we developed the following hypotheses:

H1: Switching from virtual experiences to completing questionnaires produces a physiologically measurable BIP.
H2: There is a measurable difference of BIP between the administration of INVRQ and OUTVRQ.

H3: Completing questionnaires outside VR and in VR has a measurable effect on performance in successive trials.

H4: There are measurable (uncontrolled) biases in self-reports triggered by BIP between INVRQs and OUTVRQs.

We conducted a mixed-design user study (n=50) and employed physiological measures of BIP as well as subjective ratings of presence and player experience (PX). Our results show a clear difference of BIPs between INVRQs and OUTVRQs and support the evidence from prior research [31, 74, 89, 85] that less-invasive methods of self-reports increase the validity of the measures. Our findings highlight the effects of self-reporting in VR and help both researchers and developers to consider them in their study design.

RELATED WORK

The sense of presence is considered to be the key feature of VR [23, 41] and is often treated as a metric of effectiveness of VEs [67, 102]. It is commonly defined as the subjective experience of “being there” [60] – in the virtual environment rather than in the physical space [25, 81, 84, 102]. Similarly, Lombard and Ditton identify presence as the disappearance of the communication medium – a “perceptual illusion of nonmediation” [55], which is consistent with Bystrom et al.’s suspension of disbelief [13]. Slater et al. extend this definition and describe presence as “how well a person’s behaviour in the VE matches their behaviour in similar circumstances in real life” [88] and argue that how data is displayed – and how the participants are able to interact in VR – is more important for presence than the level of realism [67, 84]. Zahorik and Jenison also claim that plausible responses of the VE to the user’s actions engage the sense of presence [103]. IJsselsteijn et al. adapt Lombard’s and Ditton’s definition and distinguish two subcategories of physical and social presence. Physical presence refers to the sense of a physical location and social presence describes the feeling of “being together” [42]. The authors also provide four determinants of presence: extent and fidelity of sensory information, match between sensors and display, content factors and user characteristics. Likewise, Sherdian proposed three determinants of presence: extent of sensory information, control of relation of sensors to environment and ability to modify the physical environment [78] and Witmer and Singer defined four factors that contribute to the sense of presence: control, sensory, distraction and realism [102]. The immersion, presence, performance model [13] follows Slater’s definition [88] and proposes a feedback loop between attention, engagement and presence which also underlies the flow theory [18].

While there are diverse definitions of presence across the literature (cf. [38, 55, 82, 102]), there is a consensus that presence is a multidimensional construct that is driven by media characteristics, such as technological factors [99, 56, 68, 81], and personal characteristics [4]. Skarbez et al. [81] provide a comprehensive review of literature on presence. The authors identify similarities between the definitions and aggregate common variables and related constructs that contribute to the sense of presence into a conceptual model of Place Illusion, Plausibility Illusion and Social Presence Illusion.

In contrast to presence, immersion is recurrently described as the objective properties of the VE and the applied technology that induce the sense of presence [9, 88, 92]. Immersion and presence are two logically separated constructs, however they are directly related [84] as presence is the outcome of immersion [70]. Brown and Cairns [12] identified engagement, engrossment and total immersion as three levels of immersion in games on a scale of involvement. Immersion includes the software and the hardware components that produce stimuli to the user’s senses and affect how the user perceives the VE [65]. Stereoscopic rendering, resolution, frame rate, field of view, levels of user tracking and fidelity of sensory input are considered as driving factors that enable immersion and thus, facilitate the sense of presence [9, 19, 36, 39, 58, 91, 99]. From a content analysis of 83 studies, Cummings and Bailenson infer that technological immersion has a medium-sized effect on presence [19]. Witmer and Singer conclude that both involvement and immersion are vital to invoke the sense of presence [102].

The relationship between presence and performance is ambiguous. While some literature showed a positive correlation between presence and performance [80, 88], others could not find a significant relationship [61, 47, 104]. Welch argues that presence does not facilitate performance, arguing that positive effects are most likely due to the increased immersive properties of the VE (e.g. frame rate, latency, resolution) [98]. Similarly, Bystrom et al. contend, that the relationship between presence and performance is task-specific and performance only increases with greater presence if the latter is relevant for the task [13]. However, a substantial body of research shows evidence that the immersive characteristics of a VE can promote task performance [62, 71, 94], learning outcomes [1, 16, 17, 57] and therapy effects [15, 23, 41, 31].

Measuring Presence

Self-reports

Post-experience presence questionnaires are the predominant method applied in the literature [81, 89]. Subjective measurements of presence should be relevant, sensitive, convenient, non-intrusive and reliable [39]. We are aware that a significant body of work developed questionnaires to assess subjective sense of presence. However, in this review we focus only on the three most commonly used questionnaires in contemporary research [74]. Witmer and Singer developed a 32-item presence questionnaire (PQ) with the subscales Involvement/Control, Natural, Auditory, Haptic, Resolution and Interface Quality [102]. The Slater-Usch-Steed (SUS) questionnaire consists of 6 items [91] on a 7-point Likert scale with no separate subscales. The authors argue that any measuring instrument of presence needs to be validated against responses in a real environment. To verify their questionnaire, the authors contrasted the SUS and the PQ between a task in VE and real environment. The results show marginal differences between real and virtual results on the SUS but not on the PQ [96]. Based on items from previous work [15, 26, 40, 65, 91, 95, 102], Schubert et al. developed the igroup...
**Presence Questionnaire (IPQ)** which consists of 14 items on a 5-point Likert scale with the subscales *General Presence, Spatial Presence, Involvement* and *Realism* [70].

The major advantage of questionnaires is that they are easy to administer and generally don’t require modifications of the VE [81]. However, since questionnaires on presence are most commonly conducted post-experience, they make for intrusive and not continuous experiences. Therefore, they are inherently unreliable for assessing presence [31, 81, 83] and they are not sensitive to state changes during the ongoing experience [54, 85]. Apart from questionnaires, research proposed relative measures using a number rating [99], a continuous scale [93] or physical sliders [31] for perceived presence. Nonetheless, these assessment methods share the same deficiency as questionnaires [81]. Accordingly, Freeman et al. suggest the application of objective measurements of presence [31].

**Behavioral Assessment**

Several approaches have been proposed for behavioral measures of presence, including responses to social or threatening stimuli [78] (e.g. ducking from a flying to object), measures of discrepancy between stimuli inside and outside the VE [90], or magnitude of postural responses [30]. The rationale behind this approach is that a higher degree of presence should result in stronger behavioral responses to the stimuli in the VE [30]. Skarbez et al. point out that behavioral measures are objective, contemporaneous and non-intrusive and thus, they overcome some of the shortcomings of the subjective measures. However, in order to trigger specific behavioral responses the VE and any ongoing study or evaluation procedure requires specific manipulations, which are not always applicable [81].

**Physiological Measures**

Physiological responses provide information about specific episodes of the experience [54] and allow a better interpretation of subjective ratings and task performance [11]. Highly immersive experiences are expected to facilitate specific reaction patterns from the autonomous nervous system [20]. Meehan reported a significant correlation between self-reported presence and skin conductance in a stressful VE, while neither skin temperature, nor heart rate (HR) [58, 59] showed significant effects. Slater et al. used the same (stressfull) VE to investigate the effects of visual realism on presence. The authors found a significant increase of electrodermal activity (EDA) and HR over baseline for the high fidelity variant, while the low fidelity VE didn’t differ from baseline [87]. Dillon et al. correlated EDA and electrocardiogram (ECG) signals to the ITC-SOPI [53] presence questionnaires and could not find a correlation between physiological signals and subjective self-reports on presence [21, 20]. In contrast, Wiederhold et al. found significant negative correlations between subjective self-reports of presence/realism and EDA/HR. Contrary to their expectation, HR was also negatively correlated with realism and presence [100]. Brogni et al. found that physiological stress increases in immersive VEs and decreases in natural-looking and engaging VEs [10, 11]. Baumgartner et al. attributed specific activity of the dorsolateral prefrontal cortex and dorsal stream in fMRI signals modulating the experience of presence in adult subjects [5]. These regions are responsive to control of attention, orientation and control of egocentric orientation [43] which is in line with many definitions of presence discussed above. Similarly, Bouchard et al. identified activation of the parahippocampus regions corresponding to the sense of presence [7]. Skarbez et al. conclude that physiological measures are “truly objective”, since they are contemporaneous and continuous. Yet again, they are also cumbersome to administer [81]. Slater et al. argue that physiological measures of presence can only be applied in anxious scenarios (e.g. a response to a threat) but that they are ineffective in mundane situations [84].

**Break in Presence**

An alternative approach to measuring presence is based on the assessment of “breaks in presence” (BIPs) [89]. BIP describes the moment when the illusion generated by the VE collapses and the user switches awareness from the VE to the physical environment [85], or when the users experience inconsistencies between their mental model and the VE [54]. BIPs can be caused, e.g. by loss of tracking, glitches, or noises outside the VE [44] and may provoke negative emotions [69]. Slater and Steed proposed a stochastic model of presence over time that relies on self-reports of transitions between “presence in the VE” to “presence in the real world”, rather than on post-experience assessment of presence. Based on BIP, a Markov chain was induced that continuously modeled the user’s state of presence. In contrast to other self-report measures e.g. questionnaires, this method assesses presence after it has been disturbed and thus is minimally invasive [67, 81, 83]. Further, it allows for investigating the cause of a particular BIP, since the moment of the BIP can be determined precisely [89]. A BIP causes a (moderately) shocking experience which invokes a physiological reaction [85] similar to the responses in physiological measures of presence, cf. [58]. Slater et al. examined the physiological response to BIPs in six different VEs and found changes in EDA and HR as indicators of BIP [85]. Liebold et al. used ECG, EDA and muscle activity (EMG) together with behavioral data measures to differentiate between ten most frequent and intense types of BIPs in a commercial PC game. The results show that gameplay interruptions produce the strongest EDA and thus the most intense BIP [54], an effect Slater et al. observed earlier when measuring HR, HRV and EDA as indicators of BIP in a “sudden whiteout” scenario [86].

**Question-asking in VR**

Recent research started taking interest in administering questionnaires in VR. Schwind et al. investigated the effect of filling out questionnaires on presence in VR. 36 participants played a VR shooter with two degrees of realism and filled out 3 presence questionnaires (IPQ [70], SUS [96], and PQ [102]) on a desktop PC or on a virtual PC in a replicated lab in VR. The authors found no differences on the mean scores, but the responses in VR showed a significantly lower variance and therefore, higher consistency of the results [74]. These results support the aim of finding noninvasive measures of presence [31, 93, 99, 89]. Alexandrovsky and Putze et al. examined the contemporary usage of INVRQs and evaluated
the design of tINVRQs in two user studies. The authors could confirm Schwind et al.’s comparable means in and out VR but could not find higher consistency of tINVRQs. Further, they report slightly lower—but still high—usability rating for tINVRQs in contrast to oUTVRQs with higher user enjoyment of tINVRQs [3]. Kang et al. [45] included a single item full screen overlay (i.e. HUD) between multiple trials in their VR environment. Participants interacted with a native VR controller to select between two or more unordered answers [45]. Schwind et al. [73, 75, 76] presented the full 32-item PQ [102] on a 3D floating UI showing one-line text instructions and 4 items on Likert-scales. Users selected answers and navigated through the questionnaire with freehand gestures. Oberdörfer et al. [63] presented a paper-like, world-referenced NASA-TLX [37] in VR. The subjects interacted with the questionnaire using a native VR controller as a pointer. Wienrich et al. [101] applied a body-referenced tINVRQ in their user study. The questionnaire is displayed on a 2D floating UI with a 20-items PANAS [50] attached to the hand of the virtual character. Fernandes and Feiner [29] applied a 10-point slider in VR as a measure of discomfort; this method was further adapted in recent research [2, 14, 35].

All INVRQ designs differ in their presentation (2D overlay, world-reference, body-referenced), their extent (single-question vs. multi-item questionnaire), question-item presentation (text-based vs. scales) and interaction modalities (pointing, free-hand, gamepad). These differences highlight a lack of standardized tools to assess self-reports in VR.

Background Summary

Previous research points out that presence is a crucial aspect of VR [23, 41, 67, 102]. Further, a break in presence threatens the VR experience and should be minimized [44]. As the ongoing debate on questionnaires for measuring presence highlights, they likely cause a break in presence [89, 81] and such sudden interruptions or transitions between realities can affect the emotional state negatively [48]. Accordingly, assessing presence within immersive experiences using biosignals suggests more reliable measurements [8, 31, 89]. However, although research has shown that INVRQs can provide more consistent results [74] for measures of presence, related work does not offer insights about the degree of BIP caused by filling out a questionnaire and thus offers limited insights into potential systematic influences, which would be valuable insights, since any such systematic influences could statistically be accounted for.

STUDY

In our study, we focus on BIP caused by discontinuation of the VR experience to fill out questionnaires. Therefore, in alignment with Schwind et al. [74], we created an immersive VE in which participants were required to engage with a playful task repeatedly at different levels of realism and respond to questionnaires inside and outside VR. We recorded physiological signals during the whole session, since they are more sensitive to assess BIPs than subjective self-reports [89].

The study design employs mixed methods and contains two independent variables: questionnaire modality and fidelity. The questionnaire modality describes the presentation of a questionnaire, either as tINVRQ or oUTVRQ. Since physiological responses are highly specific and can vary drastically between participants we chose a within-subjects design with repeated measures for the questionnaire modality. This also allows the participants to compare both assessment methods. The order of the questionnaire conditions was randomized. Fidelity is operationalized as degree of immersion [10, 46, 74] with the levels low (LoFi) and high (HiFi). Based on prior research which showed evidence that visual fidelity fosters immersion and therefore affects the sense of presence [9, 87, 104], we hypothesize that a switch from LoFi to physical reality would cause a smaller BIP compared to switching from HiFi. To avoid transfer effects, fidelity was administered between-subjects.

The Virtual Environment

In the VE, the player is located on a plateau in open space surrounded by three crystals which are attacked by drone enemies. The task of the game is to shoot the drones with a pistol using the VR controller. To eliminate a drone, the player is required to hit it twice. This is to decelerate the body movement of the player and thus reduce artefacts in the biosignals. We aimed to provide a sustained medium-intensity engagement with the game; players should feel connected to the task while not being overly aroused or stressed. Therefore, we balanced the game so that the player would lose the round within the last seconds if they did not eliminate any drones. Since we attached biosensors to the non-dominant hand, we deliberately designed the game to be playable with one hand only. To operationalize different levels of presence, we altered the visual and aural fidelity of the VE. For the HiFi variant (stylized), we...
used high resolution textures, sound, physics simulation and particle effects (Fig. 1a). The LoFi environment (abstract) only consisted of primitive mesh objects which approximated the HiFi objects without particle effects nor sounds (Fig. 1b). We paid attention to design both fidelity variants with comparable difficulty and only altered the visuals and sounds with the same hidden models for collision-detection. To avoid learning effects as confounds in the biosignals and performance, the drones spawned randomly at two fixed locations. Besides the altered fidelity and the randomized spawn-points, there were no differences in the VE between the trials. The environment was implemented in Unity3D and run on a high-end PC with the Mind Media NeXus 10 MKII biofeedback device

![Image of Mind Media NeXus 10 MKII biofeedback device](https://www.mindmedia.com/en/products/nexus-10-mkii/)

We employed the anVRQ tool, which was previously designed and evaluated by Alexandrovsky and Putze et al. [3]. The design follows general guidelines from traditional UIs [24, 79] and VR interface design [22, 34, 33, 64] and received high usability scores [3]. The questionnaire is anchored in world space and users interact with a controller using a common laser pointer metaphor (cf. Fig. 1c, 1d). We kept inVRQs constant and in the same position in both variants of the VE.

### Recording of Physiological Signals

We used a Mind Media NeXus 10 MKII biofeedback device1 with the BioTrace+ V2018A software for recording of the physiological signals with a sampling rate of 128 Hz. The NeXus 10 was connected using a 5 m USB cable allowing the participants to move around freely in the tracking space. We used a chest strap respiration sensor, a blood volume pressure (BVP) sensor and a skin conductance (SC) sensor to measure electrodermal activity (EDA) on the non-dominant hand of the participant. These biosignals are in alignment with physiological measures of presence [53, 58, 59] and BIP [54, 85, 86].

Figure 3 illustrates the recording setup: We attached the SC sensors using adjustable velcro straps to the middle phalanx of ring and little finger which have the highest SC responsiveness [77, 97]. To synchronize the recordings of the signals with the game, we used audio signals and manual triggers. Before placement of the electrodes, the participants cleaned their non-dominant hand with a wet wipe. To get a clear signal quality and reduce artifacts due to movement, we briefed the participants not to use their non-dominant hand with the sensors and to let it hang down during the whole study. A conductor helped the participants with fitting the HMD.

### Subjective Measures

To assess player experience after each game round, we applied the Player Experience of Need of Satisfaction (PENS) [66] questionnaire either using an inVRQ or outVRQ. It consists of 21 items on a 7-point Likert scale with the subscales of autonomy, competence, relatedness, presence, and intuitive control. With 21 items PENS is similar in length to other questionnaires (e.g. PANAS, IPQ, NASA-TLX) used in previous user studies with inVRQs [45, 63, 76]. To assess potential differences in perceived sense of presence due to different questionnaire modalities and to validate the “breakable experience”, we operated the igroup presence questionnaire (IPQ) [70] after the third game rounds in each condition (2 × for each participant). The IPQ consists of 14 items on a 5-point Likert scale with the subscales General Presence (GP), Spatial Presence (SP), Involvement (INV) and Realism (REAL). We did not assess the IPQ between trials because it is not sensitive for measuring a BIP [3, 74] and the game rounds did not differ. We also collected self reports about game experience and usability with both questionnaire modalities. Finally, the users ranked the BIP events by the degree of distraction.

### Procedure and Tasks

Our study flow is depicted in Figure 2 and consisted of the following states: 1. Study preparation: Briefing and complete consent form. Random assignment to a condition (HiFi or LoFi) and order of the questionnaire modality. Attach physiological sensors and synchronize biosignals with the game. Put on the HMD. 2. First questionnaire condition, steps A.--H. (inVRQ or outVRQ). 3. Break (optional). 4. Second questionnaire condition, steps A.--H. (outVRQ or inVRQ).
5. Conclusive questionnaire on a PC with demographics, ranking questionnaire modalities and debriefing.

Each questionnaire condition contained the following steps: A. Put on the HMD and play a tutorial round (60s). B. Initial blackout BIP. C. Game round #1 (90s). D. PENS #1 using INVRQ or OUTVRQ depending on the condition. E. Game round #2 (90s). F. PENS #2 using same questionnaire modality as PENS #1. G. Game round #3 (90s). H. Take off the HMD and fill out an IPQ on the PC. The whole procedure required around 45 min with an average in-VR time of 17.35 min (SD=1.45). The study took place in a lab room without irregular light, climate or noise conditions.

Participants
53 participants, mostly students from computer science and related areas from the local university, volunteered to participate in our study. All participants qualified for an Amazon voucher lottery. Due to technical problems we excluded three participants from the analysis. In the following, we consider n=50 participants (8 female [self-identified], mean age \(M=26.08, SD=4.39\)). One participant suffers from red-green color blindness, 45 participants are right-handed, 3 left-handed and 2 bi-manual, who decided to interact with their right hand. 20 participants used vision correction during the study. The four participant groups were balanced for gender (cf. Figure 2). The mean game experience was \(M=2.76, SD=2.17\) (1–8 scale, 1 max) and the mean VR experience was \(M=6.48, SD=2.22\) (1–8 scale, 1 max). One-way ANOVA.s showed no significant differences across conditions for both.

Pre-processing
We used Python 3.7 and R with a Jupyter notebook for the data analysis. The recording software BioTrace+ V2018A features pre-processing of the biosignals: From the BVP sensor, it calculates heart rate (HR) and the heart rate variability (HRV), from the signals of the respiration sensor it calculates respiration rate (RR). The signals were exported in raw format from the recording software with a sampling rate of 128Hz. The HR and the EDA signal were cleaned up using a Butterworth lowpass filter of order 5.

RESULTS
In the following section we focus on significant results regarding our hypotheses.²

Physiological Signals
In this subsection, we inspect two aspects of the physiological signals: First, we compare the magnitudes of physiological response to the INVRQ and OUTVRQ events with the signals resulting from the blackout event. Second, we analyze the influence of an INVRQ or OUTVRQ on the subsequent tasks to study any potential physiological aftereffects following such report-related BIPs.

Detection and Comparison of BIP events
In alignment with work by Liebold et al. [54] and Slater et al. [86], we employed 3-second blackouts during a first tutorial round of the VR game as a reference BIP event, which we compared to the INVRQ and OUTVRQ events. We expect a reaction in the biosignals to the blackout and the INVRQ event directly after the event. The analysis of OUTVRQ events is less straightforward: At \(t=0s\) the participants were asked to contact the experimenter for taking off the HMD, requiring \(M=13.78\)s (SD=3.85). Thus, we analyzed both post-BIP events, the occurrence of the “contact the experimenter” (red line, Fig. 4c) notification and the HMD-off event (black line). Figure 4 shows the physiological response to the blackouts (Fig. 4a), the the INVRQs (Fig. 4b) and OUTVRQs (Fig. 4c), each averaged over all participants. For EDA, we followed the procedure of Liebold et al. [54] and analyzed the phasic skin conductance response [6] as an indicator for event-related activity. We excluded the signals of one participant from the EDA analysis due to an unusually high baseline signal.

For every BIP event, we compared the measure from a time interval of seconds 3 to 6 after the event with a 3 s baseline window just before the task using paired samples t-tests. The intervals investigated are illustrated in Figure 4 with dashed green lines. As indicated in Figure 2, each questionnaire condition contains three different BIP events, leading to a total of six BIP events per participant. Table 1 lists the results of the Benjamini-Hochberg corrected paired t-tests for all six BIP events for both HiFi and LoFi conditions. All BIP events in both conditions showed a significant increase of EDA values between the investigated time frames. We analyzed HR and HRV derived from the BVP signal. To detect the effect on the HR after an event, we took a window of 3 s after the event and compared it with a 3 s baseline window before the event. For the OUTVRQ event we also compared both events (notification and HMD-off event) separately. In both fidelity conditions, neither the first blackout event nor the second BIP event showed a significant effect. Thus, the HR is not a reliable predictor for BIP events and we do not take the HR not into account in the further analysis of the BIP events. The same holds for the HRV. For the RR, we averaged and compared 10 s before and after the event as suggested by Liebold et al. [54]. However, the Benjamini-Hochberg corrected paired t-tests showed no significant differences for both blackout events in both fidelities.

Table 1 lists the results of the BIP investigations for EDA, HR, HRV and RR for every BIP event with the mean differences between the tested time frame (dependent on the signal) after the event and the baseline window before the event as well as the results of the corresponding paired samples t-test. The p-values for the t-test of each signal are Benjamini-Hochberg corrected. In line with [54] and [86], the detection of the BIP events was robust for the EDA signals for both fidelities. Thus, we base the further analysis on the EDA. Further, we investigated which of the questionnaire modalities lead to the highest BIP as indicated by the strongest EDA response. We analyzed the mean difference between the two time frames before and after the BIP events, as used for BIP detection (c.f. Figure 4 and Table 1). Figure 5 shows barplots of the results. For both OUTVRQs we took the maximum of the response to the notification and HMD off event to gain a

²The (complete) analysis and data are available at open science framework: https://osf.io/cgsqa
Figure 4. Normalized EDA reactions to all BIPs events. The x-axis represents time relative to the BIP event (red vertical line). The green dotted lines illustrate the windows used for event detection and the black line the “HMD off” event for the OUTVRQ. Note: Plot 4c has different scale-windows.

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>Blackout 1</th>
<th>Blackout 2</th>
<th>inVRQ1</th>
<th>inVRQ2</th>
<th>outVRQ1 (notification)</th>
<th>outVRQ1 (HMD off)</th>
<th>outVRQ2 (notification)</th>
<th>outVRQ2 (HMD off)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EDA</strong></td>
<td></td>
<td>23</td>
<td>0.05(2.23)*</td>
<td>0.12(2.88)*</td>
<td>0.04(2.21)*</td>
<td>0.04(2.18)*</td>
<td>0.05(2.28)*</td>
<td>0.16(2.32)*</td>
<td>0.09(2.57)*</td>
</tr>
<tr>
<td><strong>HiFi HR</strong></td>
<td></td>
<td>24</td>
<td>-7.46(-2.38)</td>
<td>-4.09(-1.31)</td>
<td>-5.91(-2.38)</td>
<td>0.64(0.23)</td>
<td>-2.65(-0.58)</td>
<td>-10.55(-2.15)</td>
<td>-1.78(-0.65)</td>
</tr>
<tr>
<td><strong>HiFi HRV</strong></td>
<td></td>
<td>24</td>
<td>-5.38(-1.68)</td>
<td>-1.16(-0.43)</td>
<td>-4.94(-1.72)</td>
<td>-1.63(-0.66)</td>
<td>-3.98(-1.22)</td>
<td>5.65(0.96)</td>
<td>1.85(0.89)</td>
</tr>
<tr>
<td><strong>HiFi RR</strong></td>
<td></td>
<td>24</td>
<td>-3.58(-3.02)**</td>
<td>-2.13(-1.93)</td>
<td>-2.51(-1.58)</td>
<td>-2.86(-1.76)</td>
<td>-0.12(-0.08)</td>
<td>-0.10(-0.07)</td>
<td>-1.84(-1.21)</td>
</tr>
<tr>
<td><strong>LoFi EDA</strong></td>
<td></td>
<td>24</td>
<td>0.06(2.49)*</td>
<td>0.15(3.01)*</td>
<td>0.15(3.08)*</td>
<td>0.08(2.56)*</td>
<td>0.11(2.12)*</td>
<td>0.28(2.39)*</td>
<td>0.07(1.66)*</td>
</tr>
<tr>
<td><strong>LoFi HR</strong></td>
<td></td>
<td>24</td>
<td>-2.23(-0.52)</td>
<td>0.67(0.23)</td>
<td>1.58(0.46)</td>
<td>-2.21(-0.67)</td>
<td>1.69(0.63)</td>
<td>-1.88(-0.37)</td>
<td>2.89(0.87)</td>
</tr>
<tr>
<td><strong>LoFi HRV</strong></td>
<td></td>
<td>24</td>
<td>-3.80(-1.46)</td>
<td>-2.27(-1.28)</td>
<td>-4.46(-1.44)</td>
<td>-1.43(-0.78)</td>
<td>1.49(0.47)</td>
<td>14.37(2.47)*</td>
<td>-2.21(-1.03)</td>
</tr>
<tr>
<td><strong>LoFi RR</strong></td>
<td></td>
<td>24</td>
<td>0.55(0.40)</td>
<td>-2.27(-1.28)</td>
<td>-4.02(-4.10)**</td>
<td>-3.45(-4.01)**</td>
<td>-1.72(-1.31)</td>
<td>2.56(1.94)</td>
<td>1.07(0.99)</td>
</tr>
</tbody>
</table>

Table 1. Responses to every BIP event for EDA, HR, HRV, RR signals split by the fidelity condition. The OUTVRQ event is split into the occurrence of the notification and the HMD take off event. For each event the mean difference between the baseline window before the event and the test window after the event and the result of a paired-sampled t-test are given. P-values are corrected with Benjamini-Hochberg correction for each signal (row). For each value: Mean difference (t-statistic), *p<0.01, **p<0.05

Figure 5. EDA differences (M,SD) of all BIP events for both fidelities.

observed power $1-\beta=.89$. Fidelity did not show a main effect and we did not observe additional interactions. For questionnaire modality and questionnaire order*fidelity we performed post-hoc tests with Benjamini-Hochberg correction. For order*fidelity none of the post-hoc tests led to a significant difference. For the questionnaire modality the post-hoc test revealed that the reaction to the OUTVRQs led to a higher BIP than the INVRQ events ($t(97)=-4.14, p<.001$, Cohen’s $d=-.58$). Further we compared the reaction to the blackout events with the reaction to the INVRQ events, running a RM-ANOVA with the four tIVRQ and blackout events as within-subject factor and fidelity as a between-subject factor. Mauchly’s test indicated that the assumption of sphericity is violated ($\chi^2(5)=15.16, p=.01$), therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon=.83$). There was no significant difference between tIVRQs and blackout events ($F(2.48, 116.64)=2.74, p=.06, \eta_p^2=.055, 1-\beta=.61$). The between factor fidelity ($F(1.47)=2.33, p=.13, \eta_p^2=.047, 1-\beta=.53$) and interaction ($F(2.48, 116.63)=1.14, p=.35, \eta_p^2=.024 1-\beta=.28$) also did not reveal an effect.

Further, we examined the EDA recovery after BIP events. To this end, we applied a 3 s sliding window for a timeframe up to 50 s after the BIP until the signal sunk back to baseline. For the blackout events and for INVRQ2, the signal stabilized after 3 to 6 seconds. The recovery after INVRQ1 took longer. We observed significant differences to baseline for 3 – 9 s.
Recovery from the OUT VRQs took by far the longest. Except for two windows of 6–9 s in LoFi and 9–12 s in HiFi (which we consider as outliers) the EDA signal was significantly above the baseline window for over 45 s. The effect can also be observed in the mean signal (cf., Fig. 4c).

Physiological evaluation of tasks
To investigate how the question asking method influences the subsequent tasks, we compared the average physiological signals during a task directly after the assessment of the subjective measurement for the previous task using either IN VRQs or OUTVRQs. We conducted a two-way ANOVA with questionnaire modality and fidelity as factors to investigate the differences in physiological responses dependent on the preceded questionnaire modality. We found a main effect on questionnaire modality $F(1,191)=28.30, p<.001$, $\eta_p^2=.99, 1-\beta=.99$ but no main effect on fidelity and no interactions. The post-hoc test confirmed this difference for tasks after an OUT VRQ ($M=40.40, SD=63$) and an IN VRQ ($M=40.40, SD=41$) with $t(193)=5.35, p<.001$ with a medium effect size of Cohen’s $d=.77$. The other biosignals did not lead to consistent effects. A comparison of the biosignals during the actual filling out of INVRQs and OUTVRQs was not possible, as the activity of both tasks is structurally and behaviorally too different [49].

Subjective measurements
Figure 6 shows the results of all PENS takes for both questionnaires (INVRQ and OUTVRQ) and both fidelities (HiFi and LoFi). One sample t-tests against a neutral score of 4 revealed positive ratings on competence ($t(199)=13.25, p<.001$) and intuitive control ($t(199)=101.88, p<.001$), and negative scores on autonomy ($t(199)=-4.43, p<.001$) and relatedness ($t(199)=-14.40, p<.001$) within both fidelities. For all PENS subscales we conducted a mixed factorial ANOVA with the order of the repeated questionnaire (Q#1, Q#2) and the questionnaire modality (INVRQ, OUTVRQ) as within-subject factors and fidelity (HiFi, LoFi) as between-subject factor. The analysis revealed significant differences between Q#1 and Q#2 for autonomy ($F(1,48)=5.89, p=.02, \eta_p^2=.11$, observed power $1-\beta=.68$), competence ($F(1,48)=4.41, p=.04, \eta_p^2=.08, 1-\beta=.56$) and intuitive control ($F(1,48)=8.90, p=.004, \eta_p^2=.15, 1-\beta=.85$). However, we did not find any significant effects/interactions on questionnaire modality or fidelity. Bonferroni-Holm corrected post-hoc t-tests between OUT VRQ1 and OUT VRQ2 showed a significant decrease of autonomy ($t(49)=2.33, p=.02, Cohen’s d=.34$) and significant increases for intuitive control ($t(49)=2.33, p=.002, Cohen’s d=.18$). As suggested by the ANOVA, a trend was indicated for competence ($t(48)=1.95, p=.06$). To investigate random biases of the assessment method (H4), we calculated the Cronbach’s alphas of the PENS measures for both questionnaire modalities (Table 2). We compared the Cronbach’s alphas using Feldt’s test [27, 28] which showed no significant differences between INVRQs and OUTVRQs on any subscale.

For the post-experience IPQs, we conducted a mixed-factorial ANOVA with questionnaire modality as withing factor and fidelity as between factor. The analysis showed no significant main effects or interaction effects. However, one sample t-tests against neutral score (3) revealed positive ratings of GP ($t(99)=8.16, p<.001$), SP ($t(99)=15.59, p<.001$) and IN ($t(99)=4.13, p<.001$) and a negative score on REAL ($t(99)=6.96, p<.001$) for both fidelities. The results are depicted in Figure 7.

Times & Performance
The completion time of the game experience including the time for filling out the PENS questionnaires was on average 742.78 s (SD=78.31) for the OUT VRQ condition and 680.90 s (SD=87.00) for the INVRQ condition with a significant difference ($t(99)=3.71, p<.001$). To exam-

---

**Table 2. Results of the Feldt reliability comparison** [27, 28] between INVRQs and OUTVRQs for the PENS subscales.

<table>
<thead>
<tr>
<th>Cronbach’s α</th>
<th>diff.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>Autonomy</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>Relatedness</td>
<td>0.69</td>
<td>0.70</td>
</tr>
<tr>
<td>Presence</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>Intuitive Controls</td>
<td>0.75</td>
<td>0.73</td>
</tr>
</tbody>
</table>
The influence of the questionnaire modality on the performance in the subsequent task, we analyzed the number of eliminated drones for game rounds 2 and 3 (Figure 8). Anderson-Darling and Levene tests assured that normality and equal variance assumptions were not violated for the performance metrics ($\alpha=.05$). A mixed-factorial ANOVA with fidelity as a between-subject variable and the within-subject variables questionnaire modality and scores in game round 2 and 3 showed no main effect for the questionnaire modality, but for fidelity ($F(1,48)=29.22, p<.001$, $\eta^2_p=.38, 1-\beta=.99$) and game round ($F(1,48)=6.14, p=.017$, $\eta^2_p=.11, 1-\beta=.70$). Post-hoc tests between HiFi and LoFi ($t(99)=-10.13, p<.001$, Cohen’s $d=.20$), and game round 2 and 3 ($t(99)=-3.02, p=.003$, Cohen’s $d=1.28$) confirm this difference. Additionally, we found an interaction effect for questionnaire modality*fidelity ($F(1,48)=5.39, p=.02$, $\eta^2_p=.10, 1-\beta=.64$). Yet, Benjamini-Hochberg corrected post-hoc tests, revealed no differences on questionnaire modality for HiFi ($t(98)=-2.01, p=.06$), nor for LoFi ($t(98)=-.48, p=.63$).

Concluding questionnaire

At the end of each session, the participants ranked the interruptions during gameplay by their level of disturbance beginning with the strongest: In the LoFi condition, the users rated the blackout event as most disturbing factor ($M=2.04, SD=1.25$), second “Leaving the VR environment” ($M=2.36, SD=1.44$) followed by “Answering questionnaires in general” ($M=2.96, SD=1.25$), “The recording of the biosignals” ($M=4.12, SD=1.14$) and “Noise from outside” ($M=4.20, SD=.98$) with no significant difference between the three investigated BIPs (blackout, questionnaires, leaving VR). Thus, the most interrupting factors where the planned interruptions and not the recording setup or external noise. In the HiFi condition, the general order of the average ranking was the same, however a paired samples t-test revealed that leaving the VE was ranked significantly higher than answering questionnaires in general ($t(49)=4.19, p<.01$).

**DISCUSSION**

An essential precondition for evaluating the effects of breaking a virtual experience is to provide the participants with a virtual experience which can be broken. The IPQ scores are high on General Presence, Spatial Presence and Involvement, but surprisingly below neutral on the Realness subscale. Remarkably, as in the results by Schwind et al. [74], the IPQ ratings were unaffected by the fidelity. The PENS outcomes revealed high scores on competence and intuitive control as it was intended in our design. Autonomy and relatedness were rated with lower scores, which was expected, since the game offered only one simple mechanic and no interaction with other players or virtual characters.

Neither game experience, nor prior experience in VR affected the ratings on IPQ and PENS. This asserts that the random assignment was successful. Further, the participants perceived the planned interruptions as much more disturbing than the uncontrolled factors, which indicates a valid study flow. Plausible ratings on the IPQ and PENS confirm that our design of the environment and the study are valid. Overall, we can conclude that the simple shooting game fulfilled its purpose and generated an engaging VR experience.

Our analysis of the physiological data (cf. Table 1 and Figure 4) show that the blackout events during gameplay and both questionnaire modalities induced significant BIPs expressed by phasic skin responses. With this, we confirm H1 and support evidence of previous work [54, 58, 85, 86] that EDA is an effective predictor for BIPs. Countering intuition, HR failed to respond reliably to the blackout event, which is in line with related work on physiological BIP detection [58], but in contrast to other results [54, 85, 86]. We explain this missing effect with the physically demanding task: To protect all of the three crystals on the plateau and kill as many drones as possible, the participants had to turn around frequently. All BIP events lead to stopping of movement, i.e. the movement effect on the signal should not differ between conditions and thus not influence any comparisons. The measured drop in HR after stopping movement was higher than the expected amplitude after the BIP [54]. Thus, it negates a BIP effect on HR. Stopping movement influences all signals including the EDA, but not equally. So the expected effect of stopped movement on the EDA signal is a drop in amplitude [72]. Instead, we observe an increase in amplitude which is smaller than the BIP-triggered increase reported by Liebold et al. [54] (stationary). Thus, we conclude that the effect of BIP is dominant and counteracts the effect of stopping movement.

Regarding H2, the comparisons of the BIPs induced by filling out questionnaires shows evidence that OUTVRQs elicit a stronger BIP than INVRQs. As depicted in Figure 5, OUTVRQs produced higher EDA responses than INVRQs with an effect size of Cohen’s $d=.58$. Although this is only a medium effect size, we measured it in a realistic setup not forcing for exaggerated BIPs. These differences might become more prominent as participants are immersed in longer VR sessions or with more demanding tasks.

The EDA response to the OUTVRQ events differ significantly from the responses to the blackouts and INVRQ events. Moreover, the ranking of disturbances indicate that the higher visual fidelity gives room for a stronger break in presence, while breaks in VE with a weaker sense of presence are less promi-
nent. In terms of temporal dynamics, the BIP appears more sudden in the blackout and INVRQ. While the EDA signals after INVRQs show similar patterns to responses after the blackouts and pacified within a few seconds, OUTVRQs stayed significantly above baseline for over 55s (Fig. 4). These outcomes are in line with observations on the experience of exiting VR by Knibbe et al. [48].

There was no measurable difference on performance (number of killed drones) in successive game rounds. Thus, we cannot find direct evidence to support H3. However, we measured a higher and longer sustained EDA for OUTVRQs. This might affect the performance for more complex or demanding tasks.

We could not find conclusive evidence to support H4. The differences only between OUTVRQ1 and OUTVRQ2 on PENS indicate that OUTVRQs appear to be more sensitive to BIPs than INVRQ. This is also supported by a raised EDA in a successive game round after an OUTVRQ. However, the analysis of the reliability of PENS does not further substantiate these results. For both modalities the reliability was high and did not differ significantly.

LIMITATIONS AND FUTURE WORK

Most of our participants were students and only 16% self-identified as female. Although we payed attention to distribute self-identifying females equally across the conditions, this indicates room for further validation. Our study design aimed to resemble a representative VR user study; therefore we asked our participants to sit at the PC when filling out questionnaires. In VR however, the participants were standing and moving around which is common in many VR applications and games [51] and was kept as such to retain ecological validity. Thus, we did not compare the physiological responses during filling out phase between the INVRQs and OUTVRQs. Such an investigation would be beneficial to examine the influence of the questionnaire modality on the attention or concentration during filling out the questionnaire.

For this study, we employed a shooting game which only consisted of a simple mechanic. Therefore, the setup might not be sensitive to effects of BIP on performance. Future work should investigate how BIPs affect performance of tasks with different degrees of complexity and workload. Likewise, we investigated the effects of BIP with only one specific application. Although the VE (particularly LoFi) was designed generically, it is unclear how presence can be affected in other types of VEs (e.g., learning, productivity, therapy) and should be investigated in the future. Furthermore, the participants interacted only with questionnaires containing Likert scales. Other types of questions (e.g., open-ended, single or multiple choice), other questionnaire types or extents might reveal different results. Unexpectedly, despite only altering the visuals and the sounds between LoFi and HiFi while keeping the game mechanics consistent, subjects in LoFi performed significantly better and hit more targets. A possible explanation is that missing aural feedback and less visual richness in LoFi allowed for stronger focus on the task.

Our results show that INVRQs can reduce BIP, but not eliminate it entirely. Due to the design and length of the embedded questionnaire this is not surprising, since the participants perceived the answering of the questionnaires as interrupting. In future work we will examine different contextualisation methods of INVRQs that foster a stronger connection between the questionnaires and the VR tasks and therefore, might reduce the question-asking related BIPs even further.

CONCLUSION

Questionnaires are widespread measurement instruments to assess subjective responses on a particular experience in VR user studies. However, research on presence shows evidence that switching between VR and physical reality leads to a break in presence [48, 89] that might alter the outcomes of the self-reports [74], especially when assessing constructs that are sensitive to disturbance and should be assessed with the least possible invasion. Researchers started to administer questionnaires in VR, which most likely reduce the BIP [29]. Yet, related work offered no clear evidence whether INVRQs diminish the BIP and to what extent filling out questionnaires contributes to a BIP.

In this paper, we investigated how question-asking itself breaks the VR experience (H1) and if INVRQs can minimize the break in presence (H2) and therefore, reduce uncontrolled biases in performance (H3) and self-reports (H4). To investigate our hypotheses, we conducted a user study (n=50) where we recorded biosignals of subjects while they played a simple VR shooter either with low or high visual fidelity and filled out questionnaires regarding their PX both in VR and on PC in physical reality. Our results clearly show that both questionnaire modalities produce BIPs. Moreover, the physiological responses in the EDA for INVRQs are significantly lower and shorter than to OUTVRQs. These results show evidence that INVRQs are less invasive than OUTVRQs and provide more reliable self-reports. Our findings suggest an influence of BIPs on performance of subsequent tasks. Moreover, these effects might become even more prominent in VEs high-quality VR experiences, such as AAA games as there is “more immersion to break”. Our findings can help researchers and designers to apply the appropriate instruments for their study design and lay groundwork for the design of INVRQs to provide validated and standardized methods of question-asking in VR.

ACKNOWLEDGMENTS

This research was partially funded by the European Unions’ Horizon 2020 research and innovation program (No 688244), project first stage, and by the German Research Foundation DFG through the project “PALMS” (no. 402779631), as well as as part of Collaborative Research Center (Sonderforschungsbereich) 1320 “EASE – Everyday Activity Science and Engineering”, University of Bremen (https://ease-crc.org/). The research was conducted in subproject P01 Embodied semantics for the language of action and change.

REFERENCES


[27] Leonard S. Feldt. 1969. A Test of the Hypothesis That Cronbach’s Alpha or Kuder-Richardson Coefficient Twenty Is the Same for Two Tests. Psychometrika 34, 3 (Sept. 1969), 363–373. DOI: http://dx.doi.org/10/d9br8k


Virtual Environments 8, 1 (Feb. 1999), 1–13. DOI: http://dx.doi.org/10/d7Fn6


Exploring Interactive Systems for Algebra Learning in School

Dmitry Alexandrovsky
Digital Media Lab
University of Bremen
dmi@uni-bremen.de

Anke Reinschüssel
Digital Media Lab
University of Bremen
anke.reinschuessel@uni-bremen.de

Tanja Döring
Digital Media Lab
University of Bremen
doering@uni-bremen.de

Angelika Bikner-Ahsbahs
University of Bremen
bikner@math.uni-bremen.de

Stefan Bollen
University of Bremen
sbollen@uni-bremen.de

Rainer Malaka
Digital Media Lab
University of Bremen
malaka@tzi.de

Abstract
Interactive Systems for learning offer great potentials for the use in school classes as they allow supporting teachers in addressing students at different knowledge levels at the same time, the systems can provide valuable and individual feedback and are motivating and entertaining for the students using it due to gamification elements. Nevertheless, developing learning systems for schools is demanding, as many different stakeholders need to be involved.

In this workshop paper, we report from a transdisciplinary research project with computer scientists, mathematics educators and a textbook publisher, in which we investigate the potentials of using tangible user interfaces for algebra learning. In this paper, we particularly present insights from a comparative study in school, in which 22 students (grade 7) used either a touch-based or a tangible-based algebra learning system on a tablet. Our results show that both system versions generally work well and that the tangible system received higher user experience.

Author Keywords
Tangible user interface; multimodal interaction; algebra learning.

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: User Interfaces
Introduction

In mathematics education, simple passive manipulatives provide valuable "hands-on" approaches to teach students abstract concepts when they start learning a new field, e.g., arithmetic, geometry, or algebra. These approaches are in-line with models from mathematics education like Bruner’s concrete-representational-abstract approach [1] or the constructionist objects-to-think-with approach [8] that suggest to use physical objects for abstract concepts, especially for beginners. While a considerable body of research on using tangible user interfaces (TUIs) for learning has been conducted, more research efforts are needed to investigate how tangible user interfaces can be applied in learning environments to support the learners.

Algebra tiles are manipulatives, which are widely used in algebra learning in North America. They consist of two types of objects with positive and negative values: square tiles as constant units $\equiv 1$, $\equiv -1$ and variables represented by rectangular objects $x$, $-x$. The red and blue color stand for the positive and the negative value of the object respectively. The interaction canvas consists of two areas for the left and right side of the equation. To solve an equation students rearrange the tiles on the surface.

Figure 1: An equation laid out with algebra tiles.

Multimodal Algebra Learning

Our design follows the guidelines from educational game design model (EGDM) proposed by Ibrahim and Jafaar [5]. This framework decomposes (gamified) learning applications into learning content modeling, pedagogy, and game design. In this section we discuss our ongoing development of the tangible algebra learning system with a focus on EGDM.

Learning Content Modeling

This project collaborates with the department of mathematics didactic and a text book publisher. This partnership provides a body of well designed task and exercises, which are already applied in the curriculum. So far, we focus on linear algebra at seven grade level of the curricula. At this stage, the students should get an understanding of variables for solving linear equations.

Pedagogy

The platform builds on algebra tiles [9] (Fig. 1). To solve equations, the students arrange tile shaped objects on a canvas. Our design considers to represent entities as objects and algebraic operations as actions. The learning environment allows for different types of operations with the tiles (e.g., grouping objects to units, addition, subtraction, multiplication, subtraction). We implemented several levels of feedback and hints, to adjust it to personal needs. In collaboration with mathematics didactics we designed exercises that progress in difficulty and cover different aspects or lenses on linear equations. New operation are introduced after the student mastered a certain skill, For example, students unlock automatic sign flip as a short hand operation: when moving a tile from one side of the equation on to the other, the object is multiplied by $-1$, as shown the transformation from steps 1 (Fig. 2) and step 3 (Fig. 3). Further we employed several single and multi user scenarios including gamified tasks (e.g., hide and seek, distribution of responsibilities).

Game Design

Our first citizen game design elements are multimodality of interaction [5] and immediacy of feedback [10] as they afford for action and exploration [3]. The student’s progress is embedded into a achievements system with further gamification elements to gratify for performance (e.g., three stars for a solution with least steps). As a good balance of challenge is required for a motivational learning experience (c.f. [2, 6]), we design the tasks based on the curriculum according to the student’s knowledge. However, we aim for
dynamic difficulty adjustment. Therefore, based on machine learning techniques we are developing a model to identify the student’s lack of knowledge for certain concepts from errors while solving equations using algebra tiles.

The system is designed for multiple target groups in mind and therefore, supports different setups: 1) An app version for tablets, which does not employ tangible interaction. 2) We employed the passive indirect tangibles with a standing tablet (Fig. 6) as was used in products like CETA [7] and OSMO 1. 3) A hybrid multitouch/tangibles version for a 55” touch screen.

User Study
We conducted a study in a local high school with seven graders to examine user experience using the indirect and the tablet versions of MAL. The goal of this study was to get insight on differences in learning behavior and user experience when interacting with the two setups. The two variants of the system were randomly assigned between subjects. We designed 5 algebraic tasks, which required only addition and subtraction for solving. The tasks were a mix of verbal and symbolic exercises, presented in the same order across all runs.

Procedure
The participants started with a tutorial. By means of examples the experiment introduces algebra tiles and the interaction to the participants. Before the subjects start the tasks, a first exercise is solved together with the experimenter. Next, the participants solve a set of 5 algebraic exercises in a row. The participants had 15 minutes to complete the tasks. The next task was only presented, when the current was solved correctly. We logged the steps and the time for each exercise. Finally, the experimenter verbally acquired answers from UMUX [4] questionnaire and asks further qualitative questions how did the participants solve the tasks, if the algebra tiles helped them solving and for the tangible group, how the subjects experienced the indirectness of the tangibles. A single session took about 20 minutes.

Results
The study was conducted with 22 seven graders from one class one week after an exam in algebra. Therefore, the students were already familiar with the mathematics presented in our application. Based on the exam’s grading the teacher categorized the students into four groups (weak, average, strong and very strong). The groups were evenly balanced between the conditions. Most participants solved all five tasks (86.4%, p < 0.001). The time for solving the tasks was significantly different between the groups ($F(3, 65) = 5.86, p < .01 \eta^2_p = .21$, showing that weaker students needed more time for solving. There were no significant differences between the conditions on the performance measures. Based on the UMUX score the interaction with tangibles received significantly higher user experience than the interaction solely with touch (See Fig. 5, $t(20) = 2.149, p < .05, \text{Cohen’s } d = 0.916$).

We asked the participants how they proceeded when they worked with algebra tiles and if the tiles supported the solving process. Most participants replied that first, they tried to solve the equation mentally and then validate their steps of calculations with the tiles. 8 subject in the tangible condition and 4 participants in the touch condition pointed out, that the tangibles helped them to visualize their steps and that they allowed for exploration. The others mentioned that algebra tiles were not helpful at all and some of them preferred the symbolic representation of the expressions. Regarding the latency and the indirectness of the physical
tangibles only one student mentioned that it was an issue. All remaining 10 subjects in the tangible condition didn’t see the technical limitations as an issue.

Discussion
As stated in our study design, the participants already covered some linear algebra in their curriculum and therefore, they were no beginners in that field. This might reflect, why most of the students tried to solve the problems mentally before trying it out with the tiles. Nevertheless, the high proportion of solved tasks and the significant difference in time for solving the equation with the weaker group performing slower, are in line with our task design and allow us to suggest that the tasks were designed with the right level of challenge. As we assumed from our design guidelines, we observed on the UMUX scale that, interacting with tangibles increased the user experience and fostered exploration. This is also reflected in the answers from the interview. Further, the technical limitations of the comparatively simple passive tangibles does not seem to negatively affect the user experience. These findings let us assume that this setup is a good candidate to be applied for use in schools.

Conclusion and Future Work
In this position paper we discuss how tangible interaction can be applied in learning environments for algebra. We presented the design and ongoing development on our hybrid tangible algebra learning platform which we employ passive indirect tangibles. Our preliminary results from a study with 22 pupils reveal, that tangible interaction raises the overall user experience of the learning application and engages for exploration. However, a comprehensive analysis of the interviews is still due. Further, we aim to look at the video recordings of the sessions for a deeper analysis on what patterns of interaction and behaviour emerges when students use the different interfaces. So far, our system is still in an early stage of development. It is planned to integrate an elaborated task management system to the teachers, where the have access to task design and adjustments for individuals. We also foresee and further game elements, such as score systems, avatars, narratives. Also, we explore other technical setups which employ smart objects as tangibles with feedback modalities of the tangibles (e.g., light, vibration). At the moment we are designing a cardboard version of the indirect tangibles, the stand for the tablet and the mirror to be added supplementary to the textbook.

REFERENCES


Vita

Dmitry Alexandrovsky is a Ph.D student in computer science at the Digital Media Lab at the University of Bremen. He is lead developer in the multimodal algebra learning project. His work focuses specifically on reality based interaction in learning environments and game user research.
ABSTRACT

We present VRBox—an interactive sandbox for playful and immersive terraforming that combines the approach of augmented sandboxes with virtual reality technology and mid-air gestures. Our interactive demonstration offers a virtual reality (VR) environment containing a landscape, which the user designs via interacting with real sand while wearing a VR head-mounted display (HMD). Whereas real sandboxes have been used with augmented reality before, our approach using sand in...
VR offers novel and original interactive features such as exploring the sand landscape from a first person perspective. In this demo, users can experience our VR-sandbox system consisting of a box with sand, multiple Kinect depth sensing, an HMD, and hand tracking, as well as an interactive world simulation.

KEYWORDS
Augmented Sandbox; Virtual Reality; Gestural Interaction; Playful Interaction; Natural Materials; Tangible Interaction.

ACM Reference Format:

INTRODUCTION
Most people relate to playing with sand from their childhood days. This unstructured play opens room for experimentation, exploration, or cooperation and helps developing important skills like creativity, proprioceptive sensing, body and space awareness [5]. Natural materials like sand have also been investigated as part of interactive systems (c.f. [2]). These materials are used because they naturally offer rich multimodal feedback and often provoke emotional associations.

Whereas several research projects have explored sandboxes and its tangible affordances in augmented reality (AR) environments (e.g., [1, 8, 9]), using real sand as interaction material in VR is a novel and original approach. Similar to our VR approach, many AR sandboxes have been used for interaction with geographical phenomena [4, 7] and for playful terrain exploration and design [1, 8].

In our work, we combine real sandboxes with VR interaction and thus provide a novel and original system that explores natural physical materials in VR, gestural interaction, and teleporting in a terraforming use case. The VRBox system has been recently published as a full paper without system demonstration at CHIPlay 2018 [3], where it received an Honorable Mention, and, in an earlier version, in 2017 as a Workshop presentation at a national workshop [10]. VRBox has not been presented at a demonstrations track before, so CHI 2019 is the first scientific event where this work can be experienced hands-on.

THE VRBOX SYSTEM
Setup
The original sandbox, which was used in a previously conducted expert evaluation [3], has a volume of 140 × 80 × 30 cm. The ground and the inner walls are covered with non-reflective cloth. Three Kinect
v1 sensors are mounted on a railing 120 cm above the sand surface and capture the tracking volume from three angles of an isosceles triangle. We use an HTC Vive as our VR platform. In addition, a Leap Motion sensor is attached to the HMD to track and visualize the user’s hands. The setup runs in Unity 3D on a desktop PC. The entire setup is shown in Figure 1.

In the VR application, the terrain is represented in a blocky Minecraft [6] like style on a 98 × 59 mesh with 32 height levels. This resolution provides a good compromise between level of detail and latency. To avoid the hands being mapped onto the terrain, we added a threshold at the top rim of the box. Everything above the rim is ignored by the Kinect sensors, and the area underneath the hands is marked as occluded. To increase performance, the terrain is split into chunks of 8 by 8 blocks. Only those blocks that have been occluded by the hands for a duration of at least 6 frames (100 ms) are updated. Since the incoming point cloud stream from the depth sensors is noisy, the height of each block is updated gradually at a rate of 10 Hz to avoid jittering. When a new height value for a block is provided by the sensors, the block is repositioned to that value, which takes up to 200 ms.

Application Scenario and Interaction

Our use case is a terraforming task where users form a 3D world of their liking with the sand and extend their creation by adding a water level to create rivers and other waters, add different lighting moods and decoration objects like trees, flowers, and castles. While standing in front of the sandbox and wearing an HTC Vive HMD, users can enjoy the haptic interaction in the sand and explore the corresponding landscape in virtual reality. Overall, they can interact with the sandbox in two modes: In tabletop perspective (Figure 2) and in first person perspective (Figure 4). In tabletop perspective, similar to AR sandboxes, users can sculpt the surface of the real sand and shape the terrain. Users can select virtual objects (e.g., plants or buildings) via mid-air gestures from a menu (see Figure 3, top) and place them into the scene using a pouring gesture above the sand. Also, users have control over properties of the environment, such as water level and illumination via moving virtual objects (see Figure 3, bottom). Moreover, users can place virtual spawn points into the environment. By locating a dedicated virtual object to the sand surface and by performing a “looking up” gesture (moving the head up), users can change to the first person perspective (Figure 4). In this mode, they can directly enter the self-created virtual world and explore it from first-person perspective on a 1:1 scale by relocating and looking around. Looking up again teleports them back into tabletop mode.

Presentation at the Demo Session

Currently the setup allows for single-user in-VR interaction. However, at the same time, multiple users can form the sand and observe their modifications on a display located next to the sandbox setup. Thus, the demonstration will be also interesting and comprehensible for bystanders. One exploration session in VR should take approximately 3 to 5 minutes on average.
CONTRIBUTIONS
With this prototype, we explore the integration of rich haptic interaction with natural materials into virtual reality and present a solution for tracking and presenting real physical materials in VR. In our informal user tests and in a previous expert evaluation we experienced that VR increases the immersive and exploratory aspects of augmented sandboxes, leading to high levels of self-perceived creativity and playfulness.

ACKNOWLEDGMENTS
This project has received funding from the European Unions’ Horizon 2020 research and innovation programm under grant agreement No 688244, project first.stage and by the German Federal Ministry of Education and Research in the grant program “Erfahrbares Lernen” (experienceable learning).

REFERENCES
Playful User-Generated Treatment:
Expert Perspectives on Opportunities and Design Challenges

Georg Volkmar
University of Bremen
Bremen, Germany
gvolkmar@uni-bremen.de

Dmitry Alexandrovsky
University of Bremen
Bremen, Germany
dimi@uni-bremen.de

Jan David Smeddinck
Newcastle University
Newcastle, United Kingdom
jan.smeddinck@newcastle.ac.uk

Marc Herrlich
University of Kaiserslautern
Kaiserslautern, Germany
herrlich@eit.uni-kl.de

Rainer Malaka
University of Bremen
Bremen, Germany
malaka@tzi.de

1 INTRODUCTION

The immersive characteristics of reality-altering technologies such as virtual reality (VR) open avenues for novel modes of treatment and facilitate the democratization of therapy [6, 18, 22, 28]. For mental health – e.g., treating phobias –, a growing body of work has shown VR to be valuable [8, 10, 11, 27], enjoyable [6, 16, 17] and sometimes even more effective [6, 20, 21] than traditional therapies. As with many other uncomfortable activities, undergoing and adhering to a therapy is difficult and many patients avoid treatment [2, 4, 5, 16]. To this end, motivational strategies from game design are frequently recommended. While a large portion of the literature on game design builds on functional challenges, which address physical or cognitive skills of the players [7], these insights may not be applicable for therapy games. Therefore our approach leans on emotional challenges where the gratification results from resolutions of tension or overcoming negative emotions [3, 7, 12, 19].

For acrophobia therapy in VR, Alexandrovsky et al. [1] developed Playful User-generated Treatment (PUT) – a two-step approach, where users first engage with a design phase, in which they can shape and design a terrain in table-top mode with top-down view and then enter an exposure phase, in which they experience the very same terrain at realistic scale from a first-person perspective (see Figure 1). Enabling users to design their exposure in a simulation (top-down view of a miniature map) before they undergo the exposure with the terrain at full-scale is the key concept of PUT, as it enables playful interaction in the first phase without impacting any desired characteristics of the second phase. The approach of Alexandrovsky et al. [1] was based on related literature on game design for mental health [13, 14], motivation [29], behavioral theories [30] and informed by interviews with practicing therapists. The concept was evaluated in a user study and showed positive effects on player experience. After showing that the game design principle can be effective, we conducted a second round of interviews with expert therapists to begin to further consider ecological validity. The outcomes confirmed the value of the approach and also pointed towards valuable design recommendation for VR games.

This work augments the previously reported evaluation of the PUT concept and discusses the design approach from the perspective of a larger group of therapists based on outcome-oriented qualitative content analysis [23]. The analysis is guided by two main areas

ABSTRACT

Virtual reality exposure therapy (VRET) is a promising approach in treating phobias such as fear of heights (acrophobia). VRET provides an effective, cost-efficient, scalable and individually adaptable alternative to traditional exposure therapy. To further foster the potential of VRET, a novel concept called Playful User-generated Treatment (PUT) was derived from expert interviews and literature review. In this paper, we provide additional insights regarding the applicability of PUT in real therapy scenarios. For that purpose, practicing psychotherapists (n=13) participated in an online survey and shared their assessments regarding PUT. By conducting qualitative content analysis (inductive category formation), we identified opportunities and challenges that should be considered for the design of playful VRET systems. Opportunities were seen for preparatory habituation, increased control and self-efficacy, improved interaction, economic usage and a realistic display of anxiety-inducing environments. Challenges included lack of direct communication and realism as well as pseudo-habitation to virtual environments.

CCS CONCEPTS
• Human-centered computing → Human computer interaction (HCI). Virtual reality.

KEYWORDS
virtual reality; exposure therapy; user-generated content; game design

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI PLAY ’20 EA, November 2–4, 2020, Virtual Event, Canada
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7587-0/20/11 $15.00
https://doi.org/10.1145/3383668.3419895
of consideration:
Opportunities: Where do professional therapists see potential in using PUT for VRET?
Challenges: What concerns need to be considered when employing PUT in VRET?
We build on – and extend – the results from the initial interview study and provide additional guiding insights for the playful design of VRET.

2 METHODS
In order to address the challenges and opportunities, a survey targeting professional psychotherapists was implemented as an online study. This survey extends the expert evaluation reported in [1] and therefore followed the same structure and procedure. However, whereas the previous evaluation served only to gain insight regarding technology acceptance and the general applicability of PUT [1], this survey was aimed at deriving specific strengths and weaknesses regarding the concept. Therefore, it included a larger group of therapists (n=13) together with a deeper analysis procedure.

2.1 Material
The survey was delivered using a Google Form consisting of an introductory page, a consent form, an extensive description of the concept (composed of a text, images and a 3 minute explanatory video) and 12 questionnaire items. The embedded video provided a short explanation of the possibilities of VR in the context of exposure therapy and it illustrated the core functionalities of PUT by displaying short clips of the terrain editor application. Next to structured and free-form responses to the questionnaire items, demographic data on the therapists’ age, gender and professional background (13 additional question items) was collected as well.

2.2 Characterizing the Expert Interviewees
In total, 13 professional psychotherapists (9 self-identified as female, 4 as male) took part in the evaluation. The reported age ranged between 28 and 67 years (M=47.69, SD=12.45). 12 participants held a professional approbation whereas the remaining participant held a master’s degree in psychology as highest qualification. In terms of work experience, participants stated to have performed their occupation as psychotherapists for a period between 1 and 40 years (M=14.00, SD=10.50). Being asked about their job specialization, 9 therapists reported to use methods from the domain of cognitive behavioral therapy (CBT) [26] most frequently, whereas the other 4 primarily used psychoanalytic methods from the domain of depth analysis / depth psychology [9]. The frequency of engaging with acrophobia therapy was assessed with “Once a Week” (n=1), “Multiple Times a Year” (n=5), “Once a Year” (n=5) and “Never” (n=2).

Regarding experience with VR in general, participants responded on a 5-point Likert-scale ranging from "No Experience (1)" to "Expert (5)" resulting in a minimum score of 1 and a maximum of 3 (M=1.46, SD=0.63). None of the therapists indicated that they had ever used VR in a therapy setting before.

2.3 Procedure
The link to the survey was distributed via social media and several networks of therapists that shared it in their newsletters and mailing lists. The first part of the survey gathered informed consent. Upon agreeing to the terms, the concept of PUT was laid out with a descriptive text accompanied by images and the 3-minute explanatory video. After the experts were informed about the concept, they responded to the items of the survey (consisting of qualitative and quantitative measures) and a demographic questionnaire. The entire procedure took between 15 and 20 minutes.

3 OUTCOMES
As part of the online survey, quantitative and qualitative measures were collected that will be reported separately.

3.1 Expert Ratings
For the first three questions, participants were asked to respond on 5-point Likert-scales ranging from "Not Useful"(1) to “Very Useful”(5). Question one asked how the therapists would rate VR in general in terms of applicability in exposure therapy. On average, this item received a score of M=4.15 (SD=.77). Question two was concerned with the applicability of playful software in exposure therapy and was assessed with an average rating of M=3.85 (SD=.86). Finally, the third question addressed the applicability of the PUT design approach specifically which received an average score of M=3.69 (SD=.91).

The majority of therapists (n=11) stated that giving patients the ability to create (or take an active part in designing) the anxiety-inducing environment themselves would be a valuable approach. Accordingly, most therapists (n=10) agreed that separating therapy into two phases of creation and actual exposure may have a positive...
Table 1: Codes and text passages of the inductive qualitative content analysis. Categories O1-O5 are concerned with opportunities of PUT whereas C1-C3 cover challenges. T1-T13 represent respective therapists. Text excerpts are translated from German.

<table>
<thead>
<tr>
<th>Coding</th>
<th>Category</th>
<th>Text Example(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>Habituation to anxiety-inducing situations</td>
<td>&quot;Playful (not as threatening), as a preparation and habituation for anxiety-inducing thoughts.&quot; (T1) &quot;This allows a graduated approach employing one's own design elements [...].&quot; (T4) &quot;By employing a playful approach, exposure therapy becomes more accessible for patients, it also facilitates the eventual real exposure in vivo.&quot; (T4) &quot;Deep cognitive processing of anxiety-inducing situations can lead to reassessment and facilitate curiosity/exploratory behavior in vivo.&quot; (T9)</td>
</tr>
<tr>
<td>O2</td>
<td>Perceived control and self-efficacy</td>
<td>&quot;[...] which increases one's own perceived control and with it one's perceived self-efficacy.&quot; (T4)</td>
</tr>
<tr>
<td>O3</td>
<td>Improved interaction of patients and therapists</td>
<td>&quot;Additionally, it enables an easier interaction with the therapist.&quot; (T4)</td>
</tr>
<tr>
<td>O4</td>
<td>Economic usage of VR</td>
<td>&quot;In a therapist’s everyday life, the HMD is more practical as it does not require the therapist to go somewhere with the patient but allows them to stay in the facility.&quot; (T5) &quot;In some regions there simply is not enough 'material' for exposure.&quot; (T10) &quot;A realistic emotional response in VR can (somewhat) replace a challenging exposure planning/execution outside the therapeutic facility and thus, save time for travelling long distances.&quot; (T11)</td>
</tr>
<tr>
<td>O5</td>
<td>Realistic environment</td>
<td>&quot;[...] very realistic and capable of addressing situational anxiety triggers of patients with fear of heights.&quot; (T7) &quot;realistic projection&quot; (T8)</td>
</tr>
<tr>
<td>C1</td>
<td>No replacement for actual communication</td>
<td>&quot;It is hard to say to what extent the software is applicable as its own therapeutic approach.&quot; (T1) &quot;[...]the therapeutic relationship would be missing which I think is essential.&quot; (T6) &quot;Direct communication with the therapist is very important.&quot; (T3) &quot;How about the communication between patients and therapists?&quot; (T3)</td>
</tr>
<tr>
<td>C2</td>
<td>Lacking realism</td>
<td>&quot;According to the video, the environment (situation) was not displayed in a very realistic way.&quot; (T1) &quot;The expo-scenario showing the mountains was poorly done, too artificial, virtual&quot; (T7) &quot;Buildings and the environment seemed rather unreal.&quot; (T12) &quot;It is fairly obvious that it is not real&quot; (T13)</td>
</tr>
<tr>
<td>C3</td>
<td>Pseudo-habituation</td>
<td>&quot;It is rather simple to expose patients to heights in real life which is preferable to a virtual version since certain thoughts such as 'this is not real', which may increase the feeling of security, do not appear in a real scenario.&quot; (T5) &quot;There might be a false sense of security which in turn prevents a therapeutic effect when actual exposure happens.&quot; (T5) &quot;In addition, it can become a cognitive avoidance-mechanism.&quot; (T13)</td>
</tr>
</tbody>
</table>

impact on the course of therapy. The remaining (n=3) therapists stated that the approach may have no effect at all.

In the design phase, patients view the terrain they are editing from a top-down perspective and at miniature scale. We asked the therapists if this may have an impact on reducing the patient’s anxiety level. Responses included "Yes, a Positive Impact" (n=8), "No Impact" (n=4) and "Yes, a Negative Impact" (n=1).

We asked the participants whether the design phase of the PUT concept may form too much of a distraction from the actual therapy. On a 5-point Likert-scale ranging from "No Distraction"(1) to "Full Distraction"(5) the mean response was a rating of $M=2.31$ ($SD=1.20$).

3.2 Opportunities and Challenges

Accompanying the quantitative items, the survey contained open-ended qualitative questions that were phrased to address the two areas of investigation. As described in the previous section, the survey participants were asked to rate the applicability of PUT on a 5-point Likert-scale. In the following question we asked the therapists to explain their reasoning for this rating in a free-text field. Additionally, another item of the survey asked for any further remarks regarding the PUT concept. Responses to these two items were subjected to a structured qualitative content analysis performed by two independent researchers. More precisely, we employed inductive category formation [23, 24] to work out specific opportunities and challenges of the concept that were expressed by
the experts. The steps reported in this section are in line with the standard procedure of inductive qualitative content analysis [25]. The content-analytical units were defined as follows: A coding unit was defined as distinct semantic elements in the text. This could be a sentence or a bullet point that was entered into the online form. The context unit was composed of two open-ended questions of the online survey which specifically targeted opportunities and challenges of the playful user-generated treatment PUT design concept. The recording unit entailed the summarized data of the online survey from all 13 participants. For the analysis, a category was defined as a property of PUT design which was emphasized by the therapists to be an opportunity or a challenge in a real therapy setting. Hereby, the level of abstraction was specified to be concrete properties of PUT design that impact its applicability for actual usage in therapy. With the preparations for a structured content analysis finished, we worked through the material and derived 5 categories of opportunities (O) and 3 categories of challenges (C) which are depicted in Table 1.

3.3 Suggestions for Improvements
In the survey, one item asked for particular suggestions that the experts may have for future implementations of PUT. One expert proposed to “enter the virtual world together” (T3) to enhance the interaction between patients and therapists. Another therapist expressed the wish to mirror the patient’s view on their device. This way, they could “encourage the patient to look around, stand still, face the anxiety-trigger consciously, to really look at it without evading the situation” (T7). One suggestion included the option to “integrate real buildings that relate to the patient’s [personal experience] as a first step to exposure” (T13). Other suggestions included the “option to enter unknown terrain” (T5) and a way to “create potentially phobic stimuli while being able to adjust the level of difficulty” (T9).

4 DISCUSSION
The quantitative ratings confirm preliminary findings of the previous study in which PUT was assessed to be well applicable in therapy and deemed capable of raising interest and enjoyment [1]. Accordingly, in this study we found that therapists were rather fond of VR and playful applications in terms of applicability in a real therapy setting. Similar responses were recorded for the PUT concept which received high ratings regarding applicability and was attributed potential positive effects on the patients’ health according to the experts. Although most participants assessed PUT to be a valuable approach, it received mixed results regarding the scaled-down virtual scene and possible distraction from the actual therapy. To obtain more nuanced findings on the experts’ reasoning for their assessment, we included open-ended questions and employed qualitative content analysis to categorize distinct opportunities and challenges of the concept.

The therapists stated that PUT allows for a graduated habituation to anxiety-inducing situations (O1) and identified this property to be a core feature of the concept. They stated that by using PUT as an element of therapy, it can serve as a preparation for actual exposure in-vivo and ease the early stages of therapeutic procedure. Additionally, according to the therapists, PUT may also increase the level of perceived control and self-efficacy (O2), which can be relevant mediators of motivation and adherence. In terms of patient-therapist communication, the approach may improve the interaction between both (O3) but should not be seen as a replacement for real communication or in-vivo exposure therapy as a whole (C1), since the relationship between patients and therapists is clearly seen as an essential element of therapy. Another opportunity that therapists noted is the relatively low cost of VR when used in a therapy setting (O4). Especially for treating certain phobias that require seeking extraordinary anxiety-triggers (e.g. treating fear of flying), VR may serve as an economic and efficient alternative. However, the experts also pointed out that VR exposure alone might lead to a kind of pseudo-habituation (C3) which means that patients could become accustomed to the virtual scene but remain anxious regarding real exposure. This concern is in line with the lack of realism (C2) that was expressed to be a potential weakness that could lead to pseudo-habituation. As some therapists rated the virtual environment to be realistic (O5), there seems to be disagreement between the experts regarding this point. This is understandable since the therapists had only sparse prior experience with VR and thus different, highly subjective standards in rating a scene to be realistic or not. Nonetheless, realism is considered to be a relevant factor in allowing a graduated preparation to real exposure while preventing pseudo-habituation to the virtual scene.

5 CONCLUSION & FUTURE WORK
This work provides an extended expert-perspective on the applicability of a novel VRET concept - PUT (Playful User-generated Treatment). This research provides an extension of previous findings [1] with a more nuanced view regarding the potential of the concept as well as concerns from professional therapists. The therapist considerations indicate that VR applications in general and the PUT concept specifically, bear great potential to be used as effective treatments in exposure therapy. Additionally, we identified points of concern that should be considered for future implementations of the concept. From the responses reported in the previous section, we can already derive expert suggestions for improving the PUT application. As therapists pointed out, communication between patients and therapists is vital for a successful treatment and should be incorporated into the concept. Entering the virtual scene together could be implemented through avatar-based projection, which is a promising trend in VRET [15]. On top of improving the interaction between patients and therapists, the participants of the survey suggested that there may be value in increasing the level of visual realism and thus, potentially prevent pseudo-habituation. These propositions can be seen as design implications to inform future developments of the PUT concept. Future work will need to investigate the applicability of PUT in a long-term study including phobic patients and CBT therapists in a real therapy setting. Moreover, we will consider other use cases in addition to acrophobia where PUT may form a valuable addition to traditional exposure therapy and assess the impact on motivation as a potential mediator.

ACKNOWLEDGMENTS
We want to thank all participants of the online expert evaluation for taking the time and sharing their expertise.
REFERENCES


Evaluating User Experiences in Mixed Reality

Dmitry Alexandrovsky∗
dimi@uni-bremen.de
DMLab, University of Bremen
Bremen, Germany

Susanne Putze∗
sputze@uni-bremen.de
DMLab, University of Bremen
Bremen, Germany

Valentin Schwind
valentin.schwind@fb2.fra-uas.de
Frankfurt University of Applied Sciences
Frankfurt, Germany

Elisa D. Mekler
elisa.mekler@aalto.fi
Aalto University
Helsinki, Finnnland

Jan David Smeddinck
jan.smeddinck@newcastle.ac.uk
Newcastle University
UK

Denise Kahl
denise.kahl@dfki.de
German Research Center for Artificial Intelligence (DFKI), Saarland
Informatics Campus
Saarbrücken, Germany

Antonio Krüger
antonio.krueger@dfki.de
German Research Center for Artificial Intelligence (DFKI), Saarland
Informatics Campus
Saarbrücken, Germany

Rainer Malaka
malaka@fzj.de
DMLab, University of Bremen
Germany

ACM Reference Format:

1 INTRODUCTION
Recent advances of Mixed Reality (MR) technology have enabled new research methods and interventions across various fields and allow for the design of highly immersive user experiences. By this, Virtual Reality (VR) and Augmented Reality (AR) research have become central topics in HCI. To measure these experiences, researchers apply a wide range of research methods using objective or subjective metrics [2]. Objective measures include behavioural metrics (e.g., gaze direction, movement amplitude), physiological measures (e.g., ECG, EDA, EEG), and performance measures (e.g., time logging, success rates). Subjective self-reports through standardized or custom questionnaires remain a widely applied method for administering mid- and post-experience measures, such as the sense of presence [30] or being embodied using virtual avatars [29]. Alternatively, VR offers a wide range of opportunities for non-obstructive assessment methods of user experience, like objective measurements using biosignals [26, 27] or behavioural measures [32, 36]. Many of these measurement methods were adapted from use-cases outside of MR, in which interactions are often less immersive, and their validity of usage in MR experiments has not yet been validated. However, researchers are faced with various challenges and design alternatives when measuring immersive experiences. These challenges become even more diverse when running out-of-the-lab studies [20, 39]. Measurement methods for VR experience recently received much attention, and research has already started to embed questionnaires in the Virtual Environment (VE) for various applications (e.g., [14, 23]) as this allows to stay closer to the ongoing experience while filling out the survey [2, 7, 12, 27, 30]. However, there is a diversity in the interaction methods and practices on how the assessment procedure is conducted. This diversity in methods shows that there is no shared agreement on standardized methods of assessing the experience of being in the VR. Moreover, research pointed towards a multitude of open questions around methodological [2, 30], technical [26], social [41], and other challenges that require a focused investigation [20]. It appears crucial to work towards a shared agreement on assessment methods of VR user studies as researchers in the HCI community have to be aware of biases that may exist for their research methods of choice. AR research strongly orients on the research methods from VR, e.g., using the same type of subjective questionnaires. However, there are some crucial technical differences that require deliberate considerations during the evaluation. In this workshop, we exchange experiences with research methods in MR (i.e., AR/VR) user studies and examine the particular challenges of the different research methods. By this, our workshop launches a discussion of research methods that should lead towards standardizing assessment methods in MR user studies. The outcomes of the workshop will be aggregated into a collective special issue journal article.

∗Both authors contributed equally to this research.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI ’21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8095-9/21/05
https://doi.org/10.1145/3411763.3441337
2 BACKGROUND
Due to its immersive nature and a wide variety of technical setups, MR requires careful deliberation of the assessment methods when aiming to conduct immersive studies with human subjects. While MR allows for the implementation of diverse research settings, the technology itself affects the research results [40]. The research tries to counteract the disengaging and tedious qualities of (VR) user studies by making the tasks more appealing [42, 43]. The assessment of User Experience (UX) falls into two categories of subjective and objective metrics [24]. Most research attributes a sense of presence [32] and immersion as the central characteristic of UX in VR. There is a variety of standardized questionnaires to assess the presence, c.f., [30]. The major advantage of questionnaires is that they are easy to administer and generally don’t require modifications of the VE [32]. However, post-experience questionnaires are not sensitive to state changes during the ongoing experience [16, 34]. Moreover, the existing scales (on presence) are often long, and the items are not always fit well to the experiences. Further, it remains open for discussion if presence is actually a good candidate to describe the quality of a VR experience since a) it is difficult to measure and b) its relationship with user performance [15, 19, 44] or the fidelity [4, 35, 44] of the environment is ambiguous. Particularly while looking at applications in the mixed reality, using a construct such as presence requires critical discussion. Yet, post-experience presence questionnaires remain the predominant method applied in the literature [32, 36]. Surveying UX within the VR experience received recent attention in the literature. Schwind et al. [30] contrasted the screen-based questionnaires against VR-embedded questionnaires and found that with embedded assessment, the subjective responses in VR are more consistent. In contrast, others have shown that in-VR questionnaires may lead to inconsistencies [11].

To counteract such inconsistencies, Alexandrovsky et al. [2] presented important usability criteria for in-VR questionnaires. Other tools that allow administering questionnaires in VR are the VR Questionnaire Toolkit [7], VRate [28]; Similarly, MRAT [21] is a toolkit for AR studies. These tools aim for a less-disruptive study flow and target problems of context-dependent forgetting [1, 10] due to environment change [25] which may bias responses.

Several approaches have been proposed for behavioural measures of UX, including gaze direction [22] responses to social [38], or threatening events [31], perception of discrepancy between VR and the physical space [37], or magnitude of postural responses [8]. Skarbez et al. point out that behavioural measures are objective, contemporaneous, and non-intrusive, and thus, they overcome some of the shortcomings of the subjective measures. However, in order to trigger specific behavioural responses, the VE or evaluation procedure of the ongoing study requires specific manipulations, which are not always applicable [32]. Highly immersive experiences are expected to facilitate specific reaction patterns from the autonomous nervous system [6]. Physiological responses provide information about specific episodes of the experience [3, 16, 18, 26] and allow for a better interpretation of subjective ratings and task performance [5]. However, these physiological signals are challenging to administer in MR scenarios. For example, assessing brain activity using Electroencephalography (EEG) sensor with Head Mounted Displays (HMDs) is cumbersome for both participants and researcher, as they may be uncomfortable to wear together, and the electrical signals from the HMD can interfere with the EEG sensors [26]. Although research has shown that physiological measures are well applicable, Slater and Steed argue that physiological measures of presence can only be applied in anxious scenarios (e.g., a response to a threat) but that they are ineffective in mundane situations [33]. While measuring VR experience in the lab is diverse, measuring becomes even more technically and methodology challenging when running out-of-the-lab studies. Out-of-the-lab VR studies allow for larger variations in the settings [20] and require researchers for complex technical solutions [38, 39]. Ma et al. investigated how to enable telemetric web VR studies and to address the technical obstacles [17].

While AR research strongly orients on the research methods from VR (i.e., presence as a quality outcome of an experience), there are some crucial differences that require careful consideration. Especially optical see-through AR includes a high degree of interaction with physical reality. Therefore, a strong focus on AR content might be disturbing [13] and a balanced fusion of reality with the virtual information is desired, which should be ideally indistinguishable for the users. Therefore, measurement methods of immersive technology should account for both AR and VR. While a significant body of work developed standardized scales for measuring presence in VR (c.f., [32]), little research has been done on the development and adoption of the questionnaires for AR experiences. Georgiou and Kyza [9] developed the Augmented Reality Immersion (ARI) questionnaire, which conceptualizes immersion in AR applications on the three levels of engagement, engrossment, and total immersion, including subscales of interest, usability, emotional attachment, attention, presence, and flow.

The presented literature outlines a series of challenges and possible pitfalls HCI faces in the context around measuring UX in immersive environments. Various toolkits and frameworks exist which address some of those challenges. However, there is still no agreement on assessing methods for UX in MR applications. This workshop targets general and specific problems of UX research methods in MR and opens a critical discussion of existing research methods aiming to retain valid results when evaluating immersive technologies. The objectives of the workshops are to find a common ground of research practices and lay out a research agenda towards standardized research methods of MR experiences.

3 ORGANIZERS
The organizers are all experienced researchers in the area of MR, evaluation of immersive experience, and the development of research methods. The co-organizers bring multiple perspectives from computer science, interaction design, psychology, and user engagement.

Dmitry Alexandrovsky is a final-year doctoral student at the Digital Media Lab, University of Bremen, Germany. His research interests are immersive interaction, user engagement, and game design research. He works on interface designs for questionnaires in VR and developed an in-VR questionnaire toolkit. His research was awarded with ‘Honorable Mentions’ at CHI PLAY conferences.

Susanne Putze is a final-year doctoral student at the Digital Media Lab at the University of Bremen. Her research interests are in HCI,
improvement of research workflows, and research communication methods. She works on measuring VR experiences using subjective questionnaires and physiological signals.

Valentin Schwind is a professor for human-computer interaction at the Frankfurt University of Applied Sciences. His work explores immersive and multimodal user experiences in virtual and augmented reality. He is an expert in the research of quantifying immersion and presence. Valentin has received multiple awards at CHI and other HCI conferences for his research of avatars and virtual characters.

Elisa D. Mekler is an assistant professor at the Aalto University Department of Computer Science. Her research interests include the applications of psychological theories and methods in HCI, as well as the development and validation of UX questionnaires. Elisa’s work has garnered multiple awards at CHI and CHI PLAY.

Jan David Smeddick is an assistant professor at Open Lab and the School of Computing at Newcastle University in the UK. Building on his background in interaction design, serious games, web technologies, human computing, machine learning, and visual effects, his research interests include virtual-, mixed- and augmented reality with a focus on applications in digital health and education.

Denise Kahl is a doctoral student at the German Research Center for Artificial Intelligence (DFKI). In her research, she explores the relationship between virtual objects and their physical representations for tangible interaction in optical see-through Augmented Reality. She evaluates AR visualizations by measuring presence using subjective questionnaires.

Antonio Krüger is the CEO of the German Research Center for Artificial Intelligence (DFKI) and a professor of computer science at Saarland University, heading the Ubiquitous Media Technology Lab (UMTL). He is an internationally renowned expert on human-machine interaction and artificial intelligence. His research focuses on Mobile and Ubiquitous Spatial Assistance Systems, combining the research areas of Intelligent User Interfaces, User Modeling, Cognitive Sciences, and Ubiquitous Computing.

Rainer Malaka is a professor for Digital Media at the University of Bremen. He is managing Director of the Center for Computing Technologies (Technologiezentren Informatik und Informationstechnik, TZI) and Director of the Ph.D. program Empowering Digital Media that is funded by the Klaus Tschira Foundation. His research focus is on multimodal interaction in MR, language understanding, entertainment computing, and artificial intelligence. Rainer is the councillor of IFIP (International Federation for Information Processing) and chair of IFIP’s technical committee on Entertainment Computing. He has an extensive experience in VR research and evaluation of VR applications from various research projects, including H2020s “first stage”.

4 WEBSITE
To advertise the workshop, we will make our workshop website (http://evaluating-mr-ws.com/) available upon the workshop acceptance, which features organizational aspects such as a Call for Participation, information about organizers, paper submission instructions, and a workshop agenda, as well as later all contributions, presentations, and discussion outcomes (included the annotated Miro boards) of the workshop.

5 PRE-WORKSHOP PLANS
We plan to broadly advertise our Call for Participation via distribution lists and on social media (e.g., Twitter, Facebook). Meanwhile, we will also send personal invitations to potential researchers and practitioners from our research community network. The submission of workshop papers will be handled through a conference management system. All submitted workshop papers will be reviewed and selected by the workshop organizers (juryed selection). We will share accepted workshop papers with the participants in advance of the conference. Participants are encouraged to publish a pre-print of their work, e.g., on arXiv or OSF.

6 WORKSHOP STRUCTURE
The workshop is planned as a single-day workshop. The schedule consists of a mixture of prerecorded talks by the participants as well as active discussions and a breakout session. We expect around 15–20 participants, where 20 is the maximum. This group size allows, on the one side, for a versatile perspective on research methods and their challenges. On the other side, it enables intensive discussions with the active participation of all participants. Preliminary schedule (CET):

Welcome (15:00 – 15:15): Opening presentation to outline the workshop motivation and goals.

Paper session 1 (15:15 – 16:30): Current challenges and barriers of measuring UX in MR

Coffee break (10 min)

Paper session 2 (16:40 – 17:55): Future directions of measuring methods for UX in MR We will have two paper sessions with max. 10 participants in each session. Each paper session is split up into two blocks of about five papers. We will show the pre-recorded video presentations in the blocks, including a short introduction of the presenter. The blocks end with an open discussion on the session’s topic. To engage the participants in the discussion, we will prepare ice-breaking questions.

Coffee break (15 min)

Breakout session (18:10 – 18:50): In the breakout session, all participants will discuss in small groups of 3–4 people for 20 min. The groups will be assigned in advance according to the paper topics. During the breakout session, the participants should brainstorm and aggregate their discussions in mind maps and charts. After that, the groups will present their outcomes to the workshop and open the discussion.

Closing and wrap up (18:50 – 19:00): Workshop results, including best practices and experiences from the field trip, will be documented. The remaining open questions will be wrapped up, and follow-up activities will be discussed.

We will end the workshop day with a virtual social event on Altspace, e.g., a basketball tournament. To facilitate the discussion in the breakout session and on the paper presentation videos, as well as to capture their outcomes, we will deploy collaborative online Miro boards, which allow for collaborative discussions remotely, and persist discussion results.

7 DISTANCE ENGAGEMENT
To incorporate both participants remotely, we will base the meeting on an online platform (Zoom or in line with the CHI 2021 centralized
Submissions will be selected by the workshop organizers based on the relevance to the workshop topic and their potential to engender insightful discussions at the workshop. At the workshop, accepted papers will have a 3-4 minutes video presentation. At least one author of the accepted paper must attend the virtual workshop. All participants must register for both the workshop and for at least one day of the conference.

Important Dates
- Submission Deadline: February 21st, 2021 at 12pm PT
- Notification: TBA
- Workshop Day: May 7th-8th/9th, 2021, virtual

For submission and further information, please visit: http://evaluating-mr ws.com/

REFERENCES
Evaluating User Experiences in Mixed Reality

CHI ’21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan


