Investigating the Non-disruptive Measurement of Immersive Player Experience

MYAT THURA AUNG

DOCTOR OF PHILOSOPHY

UNIVERSITY OF YORK

COMPUTER SCIENCE

September 2022

Abstract

There exists a phenomenon whereby individuals playing digital games enter state of intense engagement. One definition of this state is the theory of immersion, which defines immersion as a gradient process with barriers that players must pass towards achieving total immersion. The IEQ is a questionnaire that captures this experience of immersion, usually after playing sessions. This thesis aims to explore new methods that non-disruptively and granularly measure immersion.

The first study looked at whether pupil diameter, fixation rates, and fixation durations could be used to measure immersion over time. Replicating a previously published experiment, immersion was manipulated by informing them of either an advanced AI or a standard AI before play. No effect was found for the immersion manipulation. While there were significant effects on pupil diameter change and eye tracking, these were not conclusively indicative of immersive states. Issues with this study's design also revealed considerations incorporated in subsequent experiments.

The second study investigated a specific component of immersion in a rhythm game. Cognitive load was measured in a repeated measures experiment, where participants played difficult and easier levels. The NASA-TLX and electrocardiography were taken as measurements. Significant differences in heart rate variability, heart rate, and cognitive load were observed between different levels of difficulty. Results also demonstrated that repeated small questionnaires can also enable more granular measurements. Finally, four studies more were conducted to develop an IEQ short form. The first two studies used unidimensional and multidimensional item response theory factor analyses to construct the IEQ short form (IEQ-SF). The last two studies validated the IEQ-SF by replicating previously published IEQ results, and measured immersion in a pre-registered validation experiment.

This thesis provides novel insights on the non-disruptive measurement of immersion over time. It reveals considerations for research using psychophysiological measurements, and the development of short form questionnaires.

Declaration of Authorship

I, Myat Thura AUNG, declare that this thesis titled, "Investigating the Non-disruptive Measurement of Immersive Player Experience" is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

“Before I started my PhD, I could’ve finished a PhD. But then I started my PhD. . . ”

Peter York

Dedicated to my friends and all those who got knocked too hard off track.

Acknowledgements

I have been so fortunate to receive the support of so many people in my life, and I will happily and proudly exercise my right to thank as many of them as I can here.

I am forever grateful for the guidance of my supervisors Paul Cairns and Alex Wade. They both entertained my many barely-relevant questions on maths, statistics, computer science, and tangents into analyses that may have been better left alone. They both enabled me to undertake this journey, kept me from faltering catastrophically throughout.

In Burmese culture, we express a sacred and lifelong gratitude to all the teachers in our lives. Thus I will also never forget the help of my previous teachers, Rob Jenkins, Philip Quinlan, Judy Parke, Ashley Pearson, Penny Ventura, and Caroline Warnock, all of whom were instrumental to me reaching this point.

I would also like to thank my family. My parents Ko Ko Aung, Wah Wah Hlaing, and my aunt Yee Yee Aung for allowing me to persist so long in education. U Tun Aung, and U Thein Lwin, for instilling education in our values. Daw Si Si and Daw Khin Yee for their unwavering love in all my efforts.

Eternal thanks for Sagarika (Moni) Patra for tolerating me during this PhD, for teaching me so much maths and statistics, and for practising so much patience with me all these years. Other than my supervisors, I can't imagine anyone more instrumental in helping me finish this.

I am also grateful to Dariusz for teaching almost all I know of programming, to Rajitha for his ever enthusiastic friendship, Adrian for reminding me to stay grounded, Alvi for appearing at always the right time, Zakie, Alexis, & Solomon for sitting with me as I wrote this thesis, and Timur & Ashleigh for assuring me that a PhD is turbulent no matter where you are.

Thanks to my friends and fellow students at IGGI who found time for one another during their own respective trials. I would like to make special mention for Peter and Lisa, for all the game nights and food between our research, to Athanasios for being an invaluable mentor, to Valerio for working with me through the early days of the PhD, Charlie for his kind friendship, and to Ozan for the camaraderie in our statistics courses and teaching me the joys of music.

Special thanks also to Jo Maltby for making so many elements of my PhD immeasurably smoother.

I would also like to express my appreciation for Ruth, Umar, Owain, Edward, and Sidney for being such understanding and encouraging colleagues as I wrote up this thesis.

Finally, I would like to thank my examiners Craig Lindley for his time reading this thesis and providing many helpful revisions, and Helen Petrie for providing a keen, independent eye on me during my research.

Contents

List of Figures

List of Tables

List of Abbreviations

In the interest of easy reading, care has been taken to minimise the number of abbreviations used in this thesis. However, some terms are commonly employed for convenience, and are enumerated below:

AAI	**A**daptive **A**rtificial **I**ntelligence
EDA	**E**lectrodermal **A**ctivity
GEQ	**G**ame **E**ngagement **Q**uestionnaire
GSR	**G**alvanic **S**kin **R**esponse
HRV	**H**eart **R**ate **V**ariability
IEQ	**I**mmersive **E**xperience **Q**uestionnaire
IEQ-SF	**I**mmersive **E**xperience **Q**uestionnaire **S**hort **F**orm
mirt	**M**ulti-**D**imensional **I**tem **R**esponse **T**heory
NASA-TLX	NASA **T**ask **L**oad **I**ndex
PCA	**P**rinciple **C**omponents **A**nalysis
RMSSD	**R**oot **M**ean **S**quare of **S**uccessive **D**ifferences between heartbeats
RWD	**R**eal **W**orld **D**issociation

Chapter 1

Introduction

The proliferation of video games as a media of substantial and continually growing consumer interest (ESA, 2017) has seen with it a growing body of research on understanding the psychological mechanisms by which players engage with the games they play. Video games allow players to experience a broad variety of different mental states. The concept of one particular kind of state has, over the years, been assigned different names such as immersion (Cairns et al., 2014a; Jennett et al., 2008), presence (IJsselsteijn et al., 2000), engagement (Brockmyer et al., 2009) and flow (Csikszentmihalyi, 1990) among others. While the subject and description of this state will vary based on the word assigned, a general, broad description can be found commonly repeating across many such models: video games induce players into a specific state of attention- a form of participation such that their mental faculties are drawn overwhelmingly to the game, and away from other targets in a player's environment or mind. It is a state often described as enabling one to "forget the things around you, and you're focused on what you're doing in the game" (Brown & Cairns, 2004). For the benefit of easy reference, this state will henceforth be referred to as simply immersion.

Within the academic literature of games research, numerous models have formulated different definitions of immersion. Each of these attempts have usually focused on assumptions around a core research interest. Often, this is a set of cognitive components of game play, such as social elements (Konstan & Riedl, 2012), the sensory experience being physically present in a virtual environment (Slater, 2003), or a holistic combination of a game's virtual environment, story, and interactivity (Cairns et al., 2014a). Other times, a model is formulated driven by a research interest in the psychological processes rather than the elements of the media, like the cognitive state of engagement independent of any such elements from the game itself (Brockmyer et al., 2009). These various models have all arisen by what is likely to be a lack of definitional clarity within the field, which is outlined by work such as that of Michaillidis' review of immersion and flow (Michailidis et al., 2018).

At the same time, a different but related area of HCI research has been exploring the use of psychophysiological signals to measure the cognitive state of a person during different activities. This research often measures some physiological signal that can be captured from the human body, such as one's heart rate, pupil diameter, electrical conductance on the surface of the skin, or brain's blood oxygen levels, and then correlates these measurements to a corresponding index of activity such as calculating mental arithmetic (Beatty, 1982) or undertaking a driving simulation (Palinko et al., 2010).

These different models of immersion, and methods of measurement have all been developed in the context of an overarching problem, which is the latent nature of immersion. A general agreement among all the work detailed thus far is that immersion is defined as a latent variable which can not be directly observed or measured. Therefore, indirect means of measuring immersion have emerged.

It is therefore no coincidence that a growing area of research interest has been the intersection of the psychometric measurement of player experience and some form of corresponding physiological signal that might be capable of providing a more comprehensive picture of immersion during game play.

1.1 Motivation & Research Question

One such model of this cognitive state has been coined as immersion, based on the word's common usage among the general video game playing population (Brown & Cairns, 2004). This theory of immersion defines the experience as gradient, progressive degrees of engagement. Each stage is partially defined by barriers that players need to pass in order to achieve the next stage, whereby players graduate from initial engagement, to engrossment, to full and total immersion.

This definition of immersion is common measured with the psychometric scale of Jennett's Immersive Experience Questionnaire (IEQ) (Jennett et al., 2008). The IEQ aims to combine both the media driven elements of video games such as the storytelling and interactive elements with the player's cognitive states of engagement and immersion. The IEQ has been used to measure immersion across a range of different games and experiments since its conception, and has been adopted in a large volume of academic work with over 2100 citations at the time of writing this thesis. There are, of course, other alternative theories of engagement, along with corresponding questionnaires. A more detailed definition and discussion of the IEQ, as well as related and competitive theories of experience are provided in chapter 2, in section 2.1. Here, it is assumed that there are sufficiently many different scales of engagement or immersion that broadly agree on the existence of such an experience of engagement. Subsequently, it is initially stated here that immersion was the theory of choice for answering the research questions of this thesis.

Critically, by the very nature of being a psychometric scale, there is a limitation placed on the IEQ as a utility for scientific research: it requires participants to remove themselves from a state of play, and to spend the time to provide self reported estimations of their mental state of immersion by answering a 31 item questionnaire. By the very nature of doing so, their conscious thoughts are drawn to the IEQ, rather than any game that induces a state of immersion. That is to say, that the IEQ is fundamentally an experientally disruptive measurement tool that is not easily applied at the same time that video game play occurs. The result of this is that the IEQ is, like many other psychometric scales, usually restricted in its usage to only providing summative measurements of a player's experience after a play session, rather than during play itself.

Clearly, there are benefits to be reaped from being able to measure immersion at a finer granularity over time, but existing approaches do not meet this demand. This limitation of psychometric questionnaires has been one of the factors driving the promise of psychophysiological measurements as an alternative capable of providing real-time measurements of immersion. Research interest in such an approach to measuring immersion can be found in both the commercial (Ambinder, 2011) and academic domains (Drachen et al., 2010; Kivikangas et al., 2011; Nacke et al., 2014). A number of studies have successfully correlated psychophysiological signals with latent phenomenon such as flow (Bian et al., 2016; Nacke & Lindley, 2008), and some attempts have even been made to measure immersion specifically (Cutting, 2018).

Fundamentally, the IEQ and its problems described above can be resolved by developing a means to conduct non-disruptive measurements of immersion. Such an approach would be valuable in part to ensure the validity of the measurements, as well as to enable the effective use of immersion measurements towards other ends. It is worth exploring therefore, whether a temporally granular measurement of immersion can be achieved based on existing understanding of psychophysiological signals and their relationships to player cognition. Much work has been done in this space by games and HCI researchers, as described in the background chapter of this thesis. However, less work has been done in explicitly tying the psychometric measurement of immersion with that of a physiological signal.

Based on this motivation, the main research question of this thesis is thus:

Can immersion of real world games be measured non-disruptively at the same time as a video game is being played?

1.2 Methodology & Outline

The aim of this thesis was to find a means to measure immersion non-disruptively.

The specific measurement approaches taken in this thesis can be fundamentally categorised into two groups. The first and second studies form the lab based experiments attempting to outline the relationship between psychophysiological measurements and immersion as measured by the IEQ. The third is a four phase study involving the formulation of a short-form IEQ using unidimensional and multidimensional item response theory factor analyses the first two phases. Following this, a validation of a proposed short-form IEQ through a pre-registered analysis of previously collected experimental data, and finally a validation of the short-form IEQ through a new pre-registered online experiment.

The first study in this thesis sought to explore the initial relationship between immersion and the physiological modalities of pupillometry, eye tracking, and electrodermal activity. Pupillometry is the measurement of changes in pupil size, usually in response to stimulus such as playing a video game. Eye tracking also captures information from the eyes, in the gross movements of the gaze in fixations, saccades, and blinks. Electrodermal activity measures the changes in conductivity of the skin in response to stimuli such as a game. This was an exploratory study which sought to replicate the results from a series of previously published experiments on immersion. The approach of using a replication experiment was to constrain the element of detecting an effect to a case where an effect should be reasonably expected. By doing so, the exploratory aspects of this study could be focused on the use of psychophysiology to measure immersion. Analytically, psychophysiological signals were tested both relative to immersion as measured by the IEQ, as well as behavioural signals from the stimulus game itself. Results from this experiment showed that pupil diameter varied significantly based on the experimental stimulus, providing evidence suggesting that there may be scope to use eye tracking to measure player experience. Furthermore, limitations that were discovered in this experiment provided a basis to better inform and design subsequent experiments. Most were confounds on pupillometric data likely caused by the experimental design and complications of the stimulus game. These considerations on the choice of game and measurement modality were subsequently incorporated into the considerations for the design of the next study.

The second study involved the use of electrocardiography and heart rate variability as alternate modalities of psychophysiological measurement. Electrocardiography captures heartbeats, along with rhythmic patterns of activity such as the variability of heart beats. These usually vary as a response to stimuli, such as playing a game. This change in modality was required due to malfunctions with the eye trackers originally used in the first study, and a shift in target phenomenon was chosen in order to test two new hypotheses. First, would it be possible to produce interpretable and reliable psychophysiological data using more accessible consumer grade measurement apparatus? Second, would it be possible to apply a shorter psychometric questionnaire throughout an experimental session while still producing reliable measurements? Here, heart rate variability was measured in combination with cognitive work load as measured by the NASA-TLX. Armed with lessons from the previous study, and a better understanding of how to choose appropriate games for experiments, a rhythm game was chosen as stimulus based on its suitability to the analyses planned. Unfortunately, data collection was abruptly interrupted by the onset of the COVID-19 pandemic, during which period all lab based research could no longer be conducted. Nonetheless, adequate samples were collected for the planned analyses. The results from this study suggested that while the heart rate variability data was indeed more interpretable under a better experimental design, the short form questionnaire was the more statistically powerful measurement device and most importantly, the short form questionnaire was able to capture more temporally granular measurements of participants' experience of the game over the course of the experiment.

The third and final study in this thesis followed during the initial outbreak of the COVID-19 pandemic where a shift in research approach was required. During the lockdowns of the pandemic, lab based research could no longer be carried out, and so the new approach taken here was to combine the re-analysis of previously published data using the IEQ, as well as conducting new experiments online through survey research. The goal of this study was to maintain the original research objective of exploring a method to measure immersion more granularly over time, and ideally with assurances to the validity of such measurements. This resulted in a study that was driven by a combination of analysing previously collected experimental data, and collecting new data in online experiments. In this study, promising results from the use of the short form questionnaire in study 2 were taken as a basis to attempt the formation of a short form Immersive Experience Questionnaire. A four phase study was undertaken to this end. The first two sub-studies were conducted in order to systematically develop a short-form IEQ with the best standards available to robustly ensure the reliability of the scale in its miniaturised form. This was accomplished through the use both quantitative methods in the form of multidimensional item response theory factor analysis, and the qualitative understanding of the theory of immersion. In the third sub-study, the newly conceived short form IEQ was validated through the re-analysis of previously published experimental data, and in the fourth and final sub-study, the short form IEQ was validated by collecting new data through an online experiment without any use of the original full sized IEQ. Results from these studies provided evidence that a short form IEQ could be constructed as a miniaturisation of the original IEQ, with minimal impediments to construct validity and sensitivity.

1.3 Approach

The approach taken in this thesis was framed by the research question of whether a temporally granular measurement of immersion could be achieved using previously published work on psychophysiological and psychometric measurement of player experience and cognitive processes.

First, in order to hone in on an appropriate concept of player immersion, definitional differences between different models of player experience are explored in the literature review within the background chapter of this thesis (chapter 2. A rationale for the use of Jennett's model of Immersion is then provided, so that a concept of immersion is held constant across the different studies in this experiment.

Experiments within this thesis were conducted primarily with the use of commercially available video games, in a lab environment specifically designed to resemble a typical room that one would play games within their own homes. Care was taken in this area as a mean of ensuring the greatest level of ecological validity possible for these experiments. Games themselves were chosen on the basis of suitability to the goals of each respective experiment, and their source codes were modified to enable concurrent functionality and in-game collection of player telemetry data, in conjunction with measurement apparatus such as eye trackers and electrocardiograms.

With respects to the psychophysiological measurements collected, and subsequent analyses, restrictions on the scope of analyses were placed based on the principle of whether or not an analysis could be statistically reasoned to be meaningfully interpretable. The principle for this approach was based on the simple assumption that it is theoretically possible to keep conducting deeper and deeper analyses on the same dataset collected from a single experiment. This is especially true for exploratory studies in which there is an assumption that a further, yet to be defined series of analyses will be conducted following the testing of any primary hypotheses. In such cases, an analysis is considered based on whether or not statistical results could be meaningfully assessed, without risk of over-interpretation. The factors of interest here are diverse, and include reasons such as the prevention of over-testing, the prioritisation of simple analyses that are more interpretable rather than statistically sophisticated but more difficult to penetrate methods, and the balance of statistical power with the realities of working in an area where a-priori sensitivity analyses are conducted based on smallest expected effects.

Finally, an objective was also set to conduct research as generalisable and applicable as possible to video games in a broader context. To this end, the studies in this thesis were conducted with the use of real world video games as stimulus where possible. This is in contrast to a common approach in games research to adopt games specifically developed for experiments. Using such games provide many benefits, such as a better capability to manipulate elements of a game to achieve a good experimental manipulation. Generally, there are not always adequate assurances for researchers that games they develop for a lab experiment are necessarily representative of games that participants would lay of their own volition in the real world. With this in mind, a compromise was attempted for the work in this thesis whereby commercial games were used in conjunction with modifications in order to achieve a balance between ecological validity and experimental control.

1.4 Contributions

The primary contributions of this thesis can be found in outcomes relative to either the use of psychophysiological signals as a means of conducting player experience research, and the development of a short form Immersive Experience Questionnaire:

- The research in this thesis provides evidence to demonstrate that significant challenges continue to exist to inferring states of immersion solely from a physiological signal.
- This thesis provides evidence based guidelines on selecting commercial games specifically to be used as stimulus in psychophysiological experiments.

- The research in this thesis has produced a novel psychometric scale based on a previously published and well tested psychometric tool, in the form of the short form IEQ.

1.5 Ethics Statement

All research in this thesis was carried out in accordance with the University of York's Code of Practice on Research Integrity. All experiments in this thesis required screening by an ethics committee, at both the University of York's departments of Psychology, and Computer Science.

All recruited participants were aged 18 or older, and were made aware of their right to withdraw, as well as their broader rights, before taking part in an experiment. All participants were also required to provide their informed consent before involvement in any experiments.

Because of the nature of some of the hardware devices used in experiments in this thesis, both participants' mental and physical welfare were taken into consideration. Experiments were designed such that participants were not placed in any situations that may have risked their well being. The placement of measurement apparatus such as electrodes were chosen in a manner that did not cause, or at least minimised, participant discomfort. Games chosen as stimulus for experiments were selected on the basis that they would not be expected to cause distress by nature of being violent, graphic, or otherwise harmful.

All participant data was anonymised at point of collection, and stored in encrypted containers on a University of York provisioned Google Drive, and personal computing hardware.

1.6 COVID-19 Impact Statement

Like many other PhD researchers, my research and general PhD work was substantially affected by the COVID-19 pandemic. The inherent laboratory based nature of conducting experiments involving the collecting of physiological data meant that upon the arrival of the COVID-19 lockdown restrictions, all lab based research had to be halted. It is also noteworthy that at the time, I operated under the assumption that I required especially careful conduct for my health and safety due to prior incidences with respiratory diseases including a previous primary pneumothorax.

Acutely, this halting occurred during the middle of data collection for the second study involving the capture of electrocardiograph signals. The immediate impact at the time was the failure to fully collect the larger sample originally intended for that study. Furthermore, no subsequent work using any physiological methodology could be conducted during the remaining funded period of my PhD. Further omitted work related to this study included a critical second experiment originally planned to integrate the IEQ into joint use with heart rate variability measurements as well as the the NASA-TLX. Had the opportunity arisen, subsequent experiments that more robustly controlled for respiratory confounds through the use of a strain gauge would also have been prioritised as essential to any research examining HRV.

Following the COVID-19 lockdowns, research for this thesis shifted into a remote paradigm. Results from the second study highlighted an opportunity to develop a short form IEQ to be used in the same manner that the NASA-TLX was designed for, and thus the final third of my PhD work involved a four part study that took advantage of the opportunity to focus on non-lab based research.

It is my belief that the COVID-19 restrictions may have cut short my capacity to demonstrate the full breadth of my development of expertise in conducting psychophysiological research. After all, I believe that it was over time that I developed a better understanding of stimulus selection and development, the control of experimental confounds, and the robust analyses of psychophysiological data. However, I hope to have also demonstrated in this thesis that I responded to the disruption of the COVID-19 restrictions in the best manner I could account for, by taking the applicable skills and knowledge I had developed over the course of the first two thirds of my PhD and applying myself to the development and validation of the IEQ-SF in the final four-part study of this thesis.

Chapter 2

Background

The research question of this thesis is whether immersion of real world games can be measured non-disruptively at the same time that a video game is being played.

To this end, the literature review was framed and divided into two areas. First, a proper review of immersion and player experience as a whole is attempted, such that a well-founded motivation can be made for choosing to adopt the terminology and definition of *immersion*. This includes a foray into the literature for the underlying cognitive and psychological concepts tied to immersion. The second part of this chapter is a survey of research and methods to measure player experience.

2.1 Why Immersion?

Different games researchers have at various points attempted formal definitions on the observation that players of video games exhibit a form of selective attention that stretches across an assortment of sensory, cognitive, and emotional processes. For the sake of balancing brevity and completeness, a brief overview of research in player experience is established in order to justify the selection of a single model of experience to be studied in this thesis.

2.1.1 The Ability for Games to Stimulate

Early work such as that of Grodal laid the foundations outlining the multidimensional way in which players engage with the games that they play (Grodal, 2000). Grodal highlighted that the interactivity innately unique to games as a media were what defined its capability to demand from it users such a significant amount of attention and coordination. Consequently, a video game being capable of facilitating such stimulation was argued to demand from players a degree of engagement greater than that of other media such as books or film. This effort (though not specifically the exact formulation that Grodal described) then became the subject of research among the sciences including psychology, human computer interactions, and cognitive neuroscience. The resulting cumulative work after more than two decades of research is a field in which substantial definitional challenges still exist, and multiple formulations exist, from which we select one for the purposes of the experiments in this thesis.

2.1.2 Presence

Some of the earliest descriptions of the immersive experience were provided through the term "Presence". In particular, two bodies of work by Witmer et al.(Witmer & Singer, 1998) and Slater et al. (Slater & Wilbur, 1997) were the foundations upon which present day immersion research grew. Although both models of presence were primarily concerned with general virtual environments rather than digital games, they do identify aspects of user experience that are pertinent to the form of immersion that is the primary interest of this thesis.

Witmer and Singer defined presence as "the subjective experience of being in one place or environment, even when one is physically situated in another" (Witmer & Singer, 1998), and similarly Slater et al. defined presence as a sense of "being in the virtual environment", based on the extent to which a technology is capable of delivering an illusion of reality. This concept of immersion, in Slater's words, was the abstraction by which the components of presence were investigated. In contrast, Witmer and Singer defined immersion as a psychological state "characterised by perceiving oneself to be enveloped by, including in, and interacting with, an environment that provides a continuous stream of stimuli and experiences". These definitional differences can cause significant confusion, especially in the wider breadth of user research spanning the past twenty years.

Both models identified presence as a gradient experience, with differing degrees of presence. Both models described many commonalities in their respective components of presence. They both identified the critical requirement of input modality, responsiveness and feedback. Because both models were developed for research in virtual environments, proprioceptive feedback was also a significant topic of relevance. Both models describe presence as a process that isolates the user from their environment, specifically as a consequence of the vividness or reality of the virtual world; this idea is commonly expressed in the form of graphical fidelity, realism, and consistency.

Upon deeper inspection, when comparing the many factors influencing experience that are described by both models, differences become more nuanced and arguably less critical for the capture of experience measurement. Each model is more detailed in some areas, while less concerned with others; Slater specifically mentions the role of narrative or "storyline", while Witmer and Singer instead focus on immediacy and feedback. Certainly more obvious than these differences is the fact that both models appear to be domain constrained to the research environments in which they were developed. While they do mention digital games, qualities of games, and even include items in their questionnaires specific to games, they are not in themselves game play experience models or questionnaire. Instead, their primary intent was on the broader range of virtual environments.

Witmer and Singer's work in this area led to the development of two psychometric questionnaires (Witmer & Singer, 1998): the presence questionnaire (PQ) which aimed to measure the degree of presence experience in users, and the immersive tendencies questionnaire (ITQ) which aimed to measure the individual differences and personal susceptibility to experiencing presence. Items in the presence questionnaire were derived from non-empirical theoretical work that produced four factors of presence: control, sensory, distractions, and realism. Parallel to this, items in the ITQ were more focused with the activities carried out within virtual environments, such as participant perceptions of, and involvement in virtual activities. Both questionnaires were tested in several experiments utilising virtual environments and considerable samples of 151 student participants. Items in both questionnaires were correlated for internal consistency and discriminatory power, as well as construct dimensionality. Both questionnaires were statistically argued to validly measure their single respective objectives, and based on the factors each questionnaire was derived to contain, this argument is supported. Sub-scales were studied as opposed to factors due to several reasons, including an insufficient sample size. Relationships between performance and presence metrics based on the PQ were also observed.

In considering the applicability of these questionnaires to immersion in digital game play, there are compelling arguments for their usage. For example, the ITQ contains questions that exist specifically on a subscale concerned with digital games, such as "Do you ever become so involved in a digital game that it is as if you are inside the game rather than moving a joystick and watching the screen?", also with narrative involvement; "How frequently do you get emotionally involved in the new stories that you read or hear?". In fact, different items from the subscales identified in the ITQ have been found independently in other work that seeks to measure player experience. However, the ITQ is a test discriminating individual differences, and the PQ aiming to measure experience is more concerned with virtual environment interactions, rather than games explicitly. While the work by Witmer and Singer have proven to be useful points of comparison for other models that have since been published, it is evident that both of these models of presence are insufficient for the specific study of player experience when considering the context of digital game play. The models of presence critically lack properties such as specific focus on player experience in digital game play, clear descriptions of experience at different gradients, or the detail of optimal presence or immersion as described by flow.

2.1.3 Flow

Another recurrent theory adopted in both Psychology and HCI literature is that of Flow (Csikszentmihalyi, 1990) and the optimal experience. Flow theory was developed to model the optimal experience that a person reaches in performing any task such as common everyday work, or more recently, digital game play. Flow can be described a state of exhilaration, enjoyment, and intensive focus in which an individual challenges their body and/or mind to its limits to arrive at some prior expectations and exceed one's original abilities. In this state, it can be said that a person is so involved in their activity that not much else seems to matter. In describing the original model of Flow, it is important to note that it is not just a state that someone experiences, but also the cognitive and emotional state of an individual following an episode of flow. The parameters of flow are listed below in table 2.1 in the context of digital games. Flow theory was originally conceived with the objective of improving overall quality of life, but has since been adopted in HCI in common design philosophy to create optimal experiences in digital software, including digital games.

Much like how flow and enjoyment are closely tied to happiness, the relationship between game play and enjoyment has been suggested to contain the same relationships and properties (Chen, 2007). From the perspective of designers, it is desirable for a game to provide appropriate degrees and amounts of challenge to its players, and for games to be designed in such a way that the ’psychic entropies’ of anxiety from abundant difficulty and boredom from the absence of challenge are minimised. Using flow theory, the GameFlow model was developed comprised of design heuristics for producing optimal player experience in digital games (Sweetser et al., 2012; Sweetser & Wyeth, 2005). Critically, GameFlow is insightful in two ways. First, Sweetser et al. mapped elements of Flow to contemporary games literature at the time (shown in table 2.1). Second, they used this mapping of flow to specify criteria for games to meet in order to allow players to reach a state of enjoyment; such as stimulation for adequate concentration, challenging objectives meeting a player’s level of skill, support of skill mastery and ease of control. In GameFlow, immersion is defined as a state of reduced awareness of one’s surroundings and everyday life through an altered sense of time, and an emotional and visceral involvement in the game.

As a model, Gameflow is more focused on the enjoyable experiences of playing games, rather than the specific experience of getting engaged or immersed. GameFlow does incorporate immersion into its components, by describing immersion as a "deep but effortless involvement". Such a definition, while concise and apt, does not provide enough detail of the degree of engagement. There are, however, significant overlaps with the model of immersion to be described later in section 2.1.7, and these similarities will be discussed. However, because of the specified scope of interest, GameFlow is not a model of experience completely relevant to the interests of the research question in this thesis. The experience of engaging with a game extends beyond just enjoyment. So in spite of GameFlow’s commendable qualities in capturing the elements of playing games that contribute to enjoyment, it is a less appropriate choice than a model that more broadly details how players get engaged with a game, particularly the process by which they transition from not engaging with a game at all, to being fully immersed.

Further, when considering the nature of Flow as only describing the optimal experience, Flow theory cannot explain all gradients of digital game engagement on its own. For example, the state of flow could be argued as synonymous to the final immersive state of the immersion model (Brown & Cairns, 2004), but where flow differs is that it is singularly concerned with this state, while immersion is defined more broadly as a multi-stage experience, as will be discussed further below in section 2.1.7.

Games Literature	Flow
The Game	A task that can be completed
Concentration	Ability to concentrate on the task
Challenge Player Skills	Perceived skills should match challenges, and both must exceed a certain threshold
Control	Allowed to exercise a sense of control over actions
Clear goals	The task has clear goals
Feedback	The task provides immediate feedback
Immersion	Deep but effortless involvement, reduced concern for self and sense of time
Social interaction	n/a

TABLE 2.1: Mapping of Flow elements to core concepts in digital games research literature (Sweetser & Wyeth, 2005).

2.1.4 Emotion

As a high level model of experience, Flow and the experience of enjoyment involves the emotions of the individual. Csikszentmihalyi in fact states that control of the conscious requires the commitment of emotions (Csikszentmihalyi, 1990), and it is the subsequent information in consciousness that contributes towards pleasure and enjoyment. Therefore, emotions are worth discussing in the context of the research question in this thesis.

There are issues in attempting to discuss emotion, however, where it appears that there yet remains to be a full consensus among researchers on a definition of emotion (Plass & Kaplan, 2016). Within the domain of HCI specifically, Lang's theory of emotion across the two dimensional space of valence and arousal (Lang, 1995) has often been employed, such as in work by Ravaja looking at phasic emotional responses in a commercial video game (Ravaja & Saari, 2004). The valence-arousal model of emotion suggests that all emotions can be placed in a two dimensional space, where they vary based on degree of activation or magnitude (arousal), or they vary based on degree of positive or negative emotion (valence). For example the emotion of anger could be represented as a high arousal and negative valence state, whereas a mild sadness may be negative valence with a lower amount of arousal. An alternative formulation of this manner of representing emotions can also be found in the circumplex model of affect (Posner et al., 2005), which also represents emotions along two dimensions of arousal and valence. The framing of the circumplex model of emotion also describes emotions as the end product of a chain of complicated interactions between cognitive processes, and it is through these processes that physiological changes occur.

It is therefore this chain of processes that tie emotions to other cognitive processes underlying our experiences of the world. This of course includes experiences of playing video games. Fortunately, there is much greater consensus that emotions are in some way involved with the physiological state of a person. In a review of different appraisal theories of emotion, Moors et al. describe changes in appraisals as drivers for changes in physiological and behavioural responses (Moors et al., 2013). In addition to contentions as to whether emotions are the result of physiological changes or the other way around, Moors et al. also discuss the possibility that a bidirectional relationship likely exists between emotion and physiological state. This can be thought of as a feedback system, where by appraisal changes lead to physiological changes, which then lead to further appraisal changes.

Seminal work by Schachter has demonstrated this relationship between emotion and physiological state (Schachter & Singer, 1962), for example, that the manipulation of physiological state through the use of epinephrine injections resulted in participants experiencing emotions with greater arousal, such as a greater experience of amusement and fear than those in the placebo group. Here, epinephrine was used as a means to replicate the discharge of the sympathetic nervous systems, in order to observe subsequent effects on the experience of emotions in participants.

Interestingly, Schachter proposes the theory that emotions exist within a context of cognitive awareness (Schachter, 1964). This awareness is represented in the capability of a participant to recognise an emotion and appropriately label it. If the label for an emotion is accurate, then no further appraisal is needed. However, in the absence of a completely appropriate explanation, the label assigned may fall into a greater and more diverse range of emotional experiences when watching a movie, such as joy or fury. Schachter concludes that this results in a situation whereby humans can be manipulated into experiencing emotions. Such manipulations can occur, for example, when a lack of explanations or awareness is combined with a state of elevated sympathetic activity.

It then follows that if physiological arousal and emotion are tied together, then measuring physiological state may yield meaningful and useful information of a person's experience when playing a game. It is because of this relationship between physiology and emotion, that the emotion of playing games has been studied with physiological measurement methods in numerous studies (Carey et al., 2017; Ivarsson et al., 2013; Ravaja et al., 2006; Yannakakis et al., 2016). Here, a discussion is provided on work by Ravaja which explored the physiological responses and experience of emotion in the context of the valence-arousal model (Ravaja et al., 2006).

Ravaja et al. investigated this by exploring the phasic, or acute physiological responses of players to highlight events in a commercial video game (Ravaja et al., 2006). They found that physiological variations in skin conductance levels, and heart inter-beat intervals varied based on the valence-arousal scores of specific in-game events, such as falling off the level, nearly falling off, or attaining a goal. A conclusion that can be drawn here is that there is a relationship between physiology and emotion being measured. Furthermore, these emotions and physiological states are occurring in response to events within a video game.

Findings such as that reported here by Ravaja et al., are not always in support of the view that differences in physiological response to playing games are easily observable.

For example, Ivarsson et al., investigated heart rate responses to playing violent video games. Details of heart rate and heart rate variability can be found in the dedicated section 2.2.3, later in this chapter. In this study, they were controlling for both exposure to violent video games and the measurement confound of chest movement (Ivarsson et al., 2013). Here, they found differences in heart rate and heart rate variability during sleep among low-exposed gamers after playing a violent game, but at the same time also found no differences during actual play between violent and non-violent video games. Additional complications regarding the relationship between emotion and cognitive load are also discussed later in section 2.1.6 on cognitive load.

In general, it is at least clear that emotions are fundamentally involved in experience of any stimulus, including experiences of playing games. It is because of this that emotion regularly appears as an element of the various models of experience discussed in this chapter, such as in flow, in the previous section, as well as engagement and immersion in the sections to follow. However, emotions alone do not completely describe the experience of engagement and immersion. There are other elements, such as attention or cognitive absorption, that are additional requisites to engagement in addition to experiences of emotion. It is because of this that emotion is considered as a critical and fundamental component of immersive experiences, but not representative of them entirely.

2.1.5 Engagement

The Game Engagement Questionnaire (GEQ) was developed (Brockmyer et al., 2009) as a general psychometric tool to enable the measurement of a player's engagement with a video game. The underlying theoretical foundation of the GEQ is that the many different facets of player experience, including Flow, Presence, and Immersion, can all be captured in a single dimension. In other words, the GEQ assumes that these experiential concepts can all be pinned along the same scale (Norman, 2013). Brockmyer took an iterative approach to developing and validating the GEQ, including a Rasch rating scale model. The manner in which Brockmyer outlines each of these relevant models of experience, is detailed in a manner at least partially similar to what is described by Brown and Cairns (Brown & Cairns, 2004). For example, there is agreement between Brown & Cairns, and Brockmyer, in that flow is to be considered the highest gradient of engagement. They also similarly take presence as a core component of their respective models.

One key problem with the GEQ is that it was intended to study digital game violence, and this has influenced the construction of the scale itself. Questions such as "I feel scared", "I get wound up" or "I feel different" are expressly less concerned with an experience of engagement, and instead concern themselves with components of aggression, as explicated by the authors. Framing this as a problem of the GEQ is not to discredit the significant body of psychological work in digital game aggression. Rather, it is to understand that such specificity leads to an arguably considerable limitation in the generalisability and applicability to a wider breadth of games. For example, if this GEQ were to be applied to games where aggression is not of relevance to a researcher's interests, it is questionable whether it would validly capture the experiences that the researchers were interested in. At best, these questions may render relatively harmless consequences on the measurement validity of experience, but even in such a case they would present as additional, unnecessary items that reduce the efficiency of the scale.

This is in addition to the problem that the questions were not formulated and filtered through a traditional exploratory analysis, which would be the approach considered as good practice (Kline, 2000). While the Rasch models and validation were technically robust, they do not necessarily lend credibility to the theoretical ideas on which the questionnaire was designed. Indeed, the GEQ's purpose is stated as a tool to identify individuals whose propensity to engage with games is a risk factor for negative impacts. Therefore, setting aside any issues around the broader terminology of "engagement" being used here, the GEQ may not be a suitable scale for the measurement of engagement beyond the stated contexts of interest to the original authors.

These issues are further exacerbated by the way that new questions were constructed and added on after initial analyses revealed deficits in the questionnaire (Brockmyer et al., 2009). These items were added without any details as to how they were accepted to the questionnaire, or how they were individually validated in the context of the wider questionnaire. Instead, they were incorporated to the scale, and then the scale was immediately applied in an experiment validating the aggression hypothesis. By this point, the GEQ had simply claimed to achieve the measurement of engagement, with little theoretical inquiry into the nature of what was potentially being measured by the newest iterations of the survey.

A fourth issue is the experiment and example by which the GEQ was deployed. Here, Brockmyer carried out an experiment in which the GEQ was answered both prior to, and following a (violent) gaming session. If the interest of a researcher is solely to examine the experience of engagement, as was stated to be a primary goal here, then a strong correlation between a before and after administration of the test may suggest that the test is not necessarily modelling the experience in particular, but is instead discriminating the user and their individual susceptibility.

Some of the sub-scales of the GEQ (though they are not strictly treated as such, due to the unidimensional construction of the scale), also deserve questioning. At best, questions loosely related to presence only loosely address the idea of presence. Getting a Likert score of whether players lose track of where they are, or playing longer than they mean to, are not necessarily aligned with the broader theoretical formulations of presence that are discussed in the initial overview of the paper. These questions do not capture the essence of the experience of being within a game world, only the potential physical by-product of such an experience. In fact, items specifically pertaining to the perception of time have also since been challenged in work by Nordin, who has demonstrated a considerable lack of accuracy from participants in knowing how long time has actually passed (Nourbakhsh et al., 2017).

These issues collectively raise questions around the validity of the GEQ as a scale, and this is somewhat observable in practice if one considers the manner in which the GEQ has been examined as a scale. For instance, there are, to my knowledge, no controlled experiments validating the GEQ's discriminatory power in user experience between games and non-games.

A review by Norman has also made a point of the manner in which the GEQ both presents subscales of engagement in its figures and reporting, while at the same time forcing a definition of engagement that exists on a single continuum (Norman, 2013). This issue around dimensionality has also been a consideration for other scales such as the IEQ, and indeed there will be studies later in this thesis in chapter 5 that demonstrate that assumptions around dimensionality deserve a great deal of care which was not adequately addressed in the original development of the GEQ.

Based on these limitations, it appears that the GEQ is likely not an appropriate scale for the research of interest in this thesis. Both on the grounds of questionable theoretical assumption around the nature of engagement, and the components from which it is composed, and also on the grounds that the intended usage of the GEQ is a constrained context not aligned with the interest area of this thesis.

2.1.6 Cognitive Load

A particular field of games research aiming to utilise digital games as tools for education known as Serious Games has examined player experience through cognitive load (Greitzer et al., 2007). Cognitive Load Theory (Sweller, 2011) was originally developed in educational psychology as an approach to optimise the process of learning. Based on a combination of cognitive research and concepts from evolutionary biology, limited cognitive resources are comprised of the human capacity to process elements and the behaviours by which said elements interact with one another, known as element interactivity. In this model, information can be observed as combinations of elements, like for example, a mathematical equation. An equation contains several variables and manipulators, which can all be treated as elements of their own. Based on the complexity of the relationships between these elements, the knowledge an individual is attempting to learn (in this case, the equation) can be described as difficult or exceeding the cognitive resource limits of a person. The processing system for this information structure was described by five detailed principles, which primarily focused on the constraints by which human learning takes place.

In combination with commonly accepted understandings of attention and executive function, cognitive load theory in the context of digital game play begins to describe the processes of digital interaction more wholly. There has been research on the effects of massively multiplayer online games on the cognitive loads of players (Ang et al., 2007). In this qualitative study, cognitive overload was examined in players playing a 2D massively multiplayer online game– a game in which large numbers of players all exist and play on the same virtual world (Bartle, 2009). The results indicated some presence of cognitive overload and that perhaps desirable aspects of cognitive load in order to present players with a challenging experience. The implications of which indicate that there may be a profound overlap between cognitive load and certain aspects of immersion.

Cognitive load was also a subject of measurement in Joe Cuttings' PhD research (Cutting, 2018), and here it was specifically formulated as a part of the array of cognitive processes involved in immersion, rather than a surrogate of, or a competitor model for immersion. Indeed, this was treated as a means of constraining the investigation of the experience of immersion into one of its more well defined subcomponents. In this sense, cognitive load is a critical element that contributes to immersion, but is not representative of immersion itself.

When considering cognitive load theory in the context of different elements involved in playing games, the specific element of emotion comes to mind. An initial assumption might be, for example, that cognitive load theory does not completely form a theory or explanation that captures the greater experience of playing a game. After all, there are aspects of playing games that include the emotional or 'imaginative' (Brown & Cairns, 2004; Mäyrä & Ermi, 2005). However, it is worth mentioning that cognitive load has been tied to emotion. After all, emotional processing occurs in the brain like any other cognitive process. On this matter, Plass and Kalyuga have suggested multiple mechanisms by which cognitive load could be directly affected by emotion (Plass & Kalyuga, 2019). For example, medical students experiencing even a simulated patient death also measured significant increases in their cognitive load. In the context of games, simulated virtual worlds often draw the emotional investments of players– this is in fact a requirement for immersion, as discussed in section 2.1.7. It is therefore reasonable to suggest that cognitive load may be capturing more than simply the raw and sterile cognitive demands of simply playing a game as a task.

However, while cognitive load and emotion may be more closely related than one might mistakenly assume, this does not change the case that cognitive load theory on its own does not appropriately capture such closely tied experiences. Rather, the conclusion to draw here is that there should be careful considerations of cognitive load in the broader context of experience, because cognitive load is inevitably tied to emotions or other cognitive processes that are important to the experience of playing games.

2.1.7 Immersion

Finally, the subject of measurement in this thesis can be found in the model of Immersion. In a sentence, immersion can be described as "being absorbed in a new reality", usually one of a video game (Jennett et al., 2009). Here, the development and formulation of immersion as a model of experience is first discussed, following this, the measurement of immersion and the immersive experience questionnaire (IEQ) is discussed.

Theory of Immersion

Immersion as a concept and model of experience was initially developed by Brown & Cairns (Brown & Cairns, 2004) who conducted the preliminary grounded theory investigation into the concept of Immersion. This study involved the semi structured interviewing of video game players, and subsequent qualitative analysis to develop a theoretical framework for immersion. Through this, Brown & Cairns identified concepts, categories of concepts, and relationships between these categories. The resulting work was their theory of immersion.

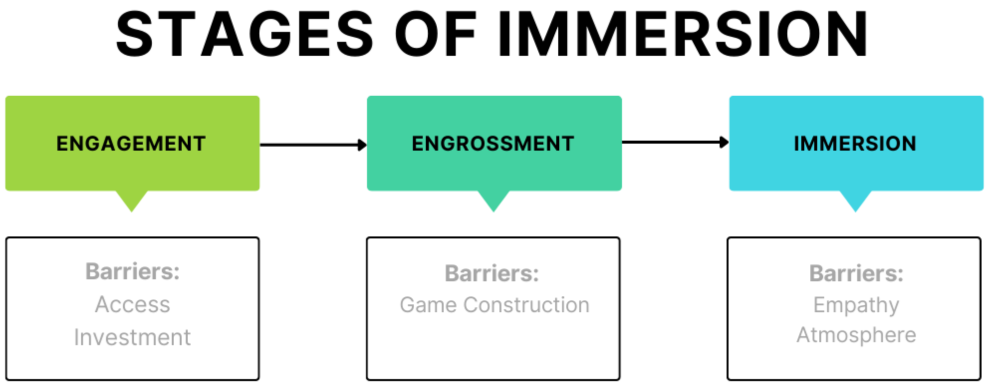

STAGES OF IMMERSION
ENGAGEMENT
ENGROSSMENT
IMMERSION
Barriers: Access Investment
Barriers: Game Construction
Barriers: Empathy Atmosphere

FIGURE 2.1: The progress stages of immersion and their barriers as described by Brown and Cairns, 2004.

Immersion is defined as a multi-stage process of engaging with a game. A flow diagram summarising these stages is presented in figure 2.1. There are three stages: engagement, engrossment, and total immersion. It is at this final transition that the experience of presence and immersion is obtained, in which players described presence as "a sense of being cut off from the world you actually inhabit", in a vein akin to the earlier work by Slater. Notably, each stage is viewed with the perspective of a barrier to be crossed in order to attain higher levels of engagement. The costs of breaking such barriers include time, effort, and attention on the part of the player.

The first stage, engagement, is the lowest level of involvement that one has to experience before reaching the subsequent stages of immersion. Under engagement, players invest their energy in the game they play, and begin to focus their attention on the game. Brown & Cairns identify two primary barriers to engagement in the form of access and investment demand. Access barriers can be found in examples such as the gamer's preference for certain types of games, or in the game controls such as the feedback from players required to play and progress through the game. Meanwhile, investment demands refer to the requirement for players to willingly commit their time and energy to a game such that they progress to its completion, or otherwise arrive at a terminus whereby they stop playing. Once these barriers to engagement are passed, players engage with a game until reaching such a point where they arrive at the next stage of (or rather, towards) immersion.

The second stage, engrossment, is described as a state in which players begin to lose awareness of their surroundings, and may even become less self aware. The latter of these properties refers to the nature by which players become increasingly emotionally involved with a game, such as with its characters or worlds. This investment of emotions occurs as a consequence of players having passed through the investment barrier in the previous state of engagement. Meanwhile, the barrier in engrossment is game construction. The game construction further facilitates the players' emotions to the point that they are directly influenced by the game itself. The construction element of this barrier was described by their participants as either visual elements, interesting tasks, or writing. The culminating result of engrossment is that players achieve a state in which they are almost autonomously engaging with a game and its virtual world.

The third and final stage is immersion. Interestingly, the original authors state this to be synonymous with presence. The experience itself is described as being cut off from reality, and detached to the extent that acutely, the game is the only thing that mattered. During the experience of immersion, the game is thought to be the only thing that is actively influencing a player's thoughts and feelings. The barriers to immersion are described as empathy, and atmosphere. Empathy is seen as the process by which players get attached to elements of the game, such as its characters. Meanwhile, atmosphere refers to the continued importance of elements from construction, a barrier from the previous stage of engrossment. Here, the emphasis on construction is placed in a manner that prioritises the demands of a player's attention. If the construction of the game continues to draw the player's attention, then they will continue to experience immersion while playing.

This theoretical formulation of immersion, particularly as a multi-stage process, provides a compelling framework with which to discuss the experience of engaging with a video game. Unlike the GEQ, this theory of immersion provides a clear theoretical basis for why it is a distinct experiential concept. Furthermore, similar to previous work by Slater, Wilbur, and Witmer (Slater & Wilbur, 1997; Witmer & Singer, 1998), there is knowledge drawn from existing games research on the design and technological factors of control inputs and graphics. Subsequently, work by Jennett et al. and the research in her PhD thesis lead to the development of the immersive experience questionnaire that continues to be used to capture and measure immersion (Jennett et al., 2008; Jennett, 2010).

Measurement of Immersion, and the IEQ

Drawing from the theory of immersion laid out by Brown & Cairns, Jennett et al., developed the Immersive Experience Questionnaire (IEQ), as a way of capturing the multidimensional experience described in the previous section.

The development of the IEQ itself involved the construction of pairs of positively and negatively phrased statements relating to immersion. These questions were derived from the contents of either Brown & Cairns' study (Brown & Cairns, 2004), or from findings of studies of other related concepts such as flow and presence. The questionnaire was then administrated in two experiments, before being further refined and validated.

The experiments involved the administration of the initial IEQ in either a task where participants played a game, or a control task in which participants completed a tangram exercise rather than playing a game. These were dubbed as either the immersive condition (playing a game), or non-immersive condition (tangram task). The difference between the two experiments was the inclusion of an eye tracking measurement for the second experiment, where gaze was captured to explore fixation differences between conditions. Because of the confounds involved in this eye tracking experiment, however, the focus here will be confined to the discussion of the IEQ specifically. The initial proposed IEQ did produce higher mean immersion scores in the immersive condition than the non-immersive condition for both experiments. Qualitative discussion of the experiments also included the observations that there may have been confounding factors in the tangram task that influenced the immersion scores of participants in this control condition. This could have potentially included the gamification of the tangram task by some participants, as well as the duration of the tangram task varying based on the challenge demands upon participants.

The subsequent refinement process of immersion, therefore, involved the abandonment of the positive/negative framing system for questions. Instead, a new approach using a multi-factor structure that also drew its item construction from both Brown & Cairns' theory of immersion, as well as related areas in flow, cognitive absorption, and presence. The factors themselves were also constructed based on these theoretical considerations, and six factors were formed: basic attention, temporal dissociation, transportation, challenge, emotional involvement, and enjoyment. This new formulation of the IEQ was then administered to a large sample of participants (263), and a factor analysis was conducted as a validation process.

This factor analysis confirmed that the majority of the questions in the IEQ measured the same underlying concept, based on a shared omnibus factor. Adjustments were also made such that rather than six factors, the IEQ consisted of five factors based on the scree plot produced by the principal components analysis. These factors were then renamed as Cognitive Involvement, Real World Dissociation, Challenge, Emotional Involvement, and Control.

It is this form of the IEQ that continues to be used at the time of writing this thesis, and this IEQ was validated in a final experiment by Jennett et al., where the same design as the first two experiments were used with the added adjustment of controlling for pace differences between the immersive and non-immersive tasks. Again, the IEQ was able to capture differences in immersion between the two conditions, and this result was used as a basis to conclude that the IEQ was likely to have been measuring immersion.

Conceptual and Theoretical Considerations of Immersion

The IEQ in this form is noteworthy for its inclusion of not just the concepts outlined in Brown & Cairns' theory of immersion, but also similar overlapped concepts from other areas such as flow and presence.

Interestingly, one manner in which the IEQ diverges from Brown & Cairns' theory of immersion is that the IEQ measures across five factors that, for all intents and purposes, exist in the same phase. Whereas Brown & Cairns described immersion as a gradient experience where players had to breach barriers to subsequently more engaging degrees of immersion, the IEQ instead captures the components discussed by Brown & Cairns in parallel. This is in part because a questionnaire is generally only administered as an omnibus series of questions to be answered in a single test, but it nonetheless follows that there might likely be either limitations with the IEQ as a measurement tool that does not appropriately capture important dimensions of the underlying theoretical concept, or that the original theory of immersion formulated by Brown & Cairns is not an accurate description of the manner in which players engage with games and get immersed. Perhaps what is being captured by the IEQ therefore, is only a fraction of the greater experience of immersion, limited in either the dimension of degrees, or in its coverage of the underlying latent components of immersion, or both.

Implications from the Analysis of Immersion

Further considerations and potential limitations of the IEQ arise when considering the manner in which the questionnaire was scored. In Jennett's study and subsequent studies during this time period, the IEQ was scored in two manners that now might be considered erroneous.

First, a single overall immersion score is taken, presumably based on the fact that the omnibus factor analysis presented evidence for most questions capturing the same underlying concept. This in itself is not too substantial an issue, as it has already been partly corrected by later research where a normalised, mean immersion score is taken instead.

More important is the subject of dimensionality. Since the IEQ is formulated as a multidimensional concept, there have been cases such as that of work by Denisova where it has been analysed dimension by dimension, in addition to the full IEQ score (Denisova, 2016). The question worth considering here is whether there is an optimal or preferred approach for scoring the IEQ. If the underlying theory of immersion is indeed multidimensional, then is it most optimal to be taking a single immersion score? In some cases at least, the answer would remain a firm yes, because analysing five factors individually would require error correction, thereby reducing the power of the IEQ. In this respect, the Rasch model proposed by Brockmyer makes more practical sense. However, the individual components are not without their merit or utility. Being able to disseminate the specific mechanisms by which immersion is taking place is helpful. Not all games, for instance, will present stimuli that emotionally engage their players. Likewise, not all games will pose a great degree of challenge, yet players may engage or immerse themselves in these games nonetheless. It is therefore at least a good idea for researchers to apply discretion in how they consider the IEQ, its scoring, and subsequent analysis. These decisions can potentially have profound implications on both the framing of the theory of immersion, as well as the statistical practices of its analysis.

Overlaps with Presence

However, in order to obtain Brown and Cairns' definition of presence, the final requirements of empathy and atmosphere are distinguishing factors from presence. These are described as a "growth of attachment" to the game and its constructions. The specific role of empathy for instance, explicitly requires players to become emotionally involved, enabling a transfer of consciousness to occur. This is a factor most pertinent to digital games, more so than simpler (for lack of a better term) virtual environments that may not concern themselves with (ludo)narrative or atmosphere.

Overlaps with Flow

Further overlaps were also identified with flow theory, which was actually described to be the most critical in achieving optimal immersion. A review by Michalidis et al. (Michailidis et al., 2018) even argued that immersion and flow were not uniquely different concepts, or at least provides an incomplete taxonomy of a player's experience. Drawing from literature in presence, flow, and immersion, the authors argued that based on a lack of understanding and adequate evidence for distinctly different measurements provided by flow and immersion, it would be hard to distinguish whether the two concepts were truly different from one another.

However, it is also important to note that this review arrived at such a conclusion primarily out of the same objective with which the first part of the review in this thesis is driven by — a need to navigate through the definitional challenges to arrive at the optimal tool for the addressing the research questions of interest. Therefore, there is no resolution to a definitional issue by simply taking their conclusion that immersion and flow are equivalent entities, as it would not simplify which of the two models to select from. Further, although the authors admirably consider the broader implication of experimental work conducted with alternative models in games research, much less effort is concerted to clarifying why flow might presumably be better than immersion, or vice versa.

An additional issue also arises when considering the arguments put forth by Michalidis et al. In reviewing the distinguishing features between presence and immersion, they conclude that presence is in fact more likely to be an independent entity. Further, they posit that based on experimental evidence, presence may even occur at a stage earlier than flow occurs in the model of immersion. However, if immersion includes presence (in the form of real world dissociation) in its framework of progression, then surely it is more likely that immersion is a separate entity from flow, given the latter's lack of accounting for any such staging or direct address of presence as an experience.

A final remark regarding the overlaps between immersion and flow can be made regarding GameFlow (Sweetser & Wyeth, 2005) specifically. In the GameFlow model of experience, enjoyment is described as a process which consists of eight core elements. Unsurprisingly, there are significant overlaps between these elements and that of immersion, namely in challenge, skills, control, and immersion. Firstly, it is interesting that GameFlow defines immersion in a manner similar to how it is defined by Brown & Cairns– that of an experience that dissociates the player from the real world. Similarly, the elements of challenge, control and skills are all significant parts of the earlier barriers to reaching engrossment, as defined by Brown & Cairns. A question therefore arises: do the underlying components of GameFlow really differ from those of Immersion?

In short, the elements of challenge and player skills are described similarly by both immersion and GameFlow such that the player must be challenged in an amount appropriate to their skill level to remain engaged with a game. Similarly, the control element is also mostly synonymous between the two scales, whereby players should not feel too much resistance from the manner in which they control their avatars within the games. Where the two models diverge most, are how immersion is defined in GameFlow, and how it is defined by Brown & Cairns.

At a first glance, the definition of immersion might appear to be the same. But the key distinction to be drawn is the manner in which immersion is framed as a requirement in GameFlow, whereas it is framed as the destination of maximal engagement by Brown & Cairns. GameFlow treats immersion as a necessity to achieve flow state: "players *should* become less aware of their surroundings", or "players *should* experience an altered sense of time". One could make the argument that even then, the definitions of immersion itself remain the same between the two models, it is simply that immersion is a sub-component of GameFlow rather than an entirely different concept.

One way to explore this thought is if one considers how the IEQ presents immersion as it is measured in real world play. There are instances where players might not attain particularly high scores in one of these dimensions of immersion that overlaps with GameFlow, such as challenge, yet they will achieve immersion nonetheless– does that mean that players are simply immersed under the model of GameFlow, but not truly achieving optimal enjoyment?

Seemingly, it is conceivable that immersion is a component of flow and/or GameFlow. However, immersion is also focused on the process of engaging with a game, which is a finer scoped concept than that of GameFlow. Subsequently, for the research question in this thesis, immersion still remains the concept of choice purely for the practical reasons of constraining the scope of measurement, as well as focusing the efforts of measurement into engagement rather than additional functions that contribute to a broader experience of enjoyment.

Emotions & Immersion

Earlier, emotions and theories of were discussed as an important and fundamental element of player experience. This is made more evident by the fact that emotions are defined as a critical barrier and element of immersion. Indeed, empathy is described as the final barrier of achieving an immersive state. The ability of a game to draw (potentially very strong) emotion responses from players is likely a significant reason that explains how immersive experiences take place at all.

Interestingly, the model of immersion provided by Brown & Cairns and the subsequent questionnaire do in fact dedicate an entire factor to this emotional aspect of playing games. In the third experiment validating the IEQ, Jennett et al found that the speed at which the tasks were experienced had a subsequent effect on the negative affect of participants (Jennett et al., 2008). The conclusion drawn based on this, was that as players get more swept up by the demands of a challenging task, their experiences of anxiety as measured by a separate scale also increased. Interestingly, increases in anxiety corresponded with increases in immersion, suggesting that emotionally charged contexts are a critical element of immersion. It is also based on this observation that Jennett et al. drew a distinction between immersion and flow: where flow is concerned with achieving a clearly positive state of enjoyment, immersion can be positively influenced by emotional states that might otherwise be seen as negative.

This manner of evaluating and quantifying the relationship between emotion and immersion is also not something that was done in the development and validation of the GEQ (Brockmyer et al., 2009). Not only does immersion incorporate from the broader array of theories of experience (which the GEQ also does), the formulation of the questionnaires were validated in the context of other affective states. This ultimately lends further support to the proposal that immersion is the preferable model of engagement to be studied in this thesis.

The Case for Immersion

If one takes Michalidis et al.'s presumed preference to default to flow given its richer literature and base of evidence, it would also be necessary to accept, based on the authors' own evaluation, that one no longer accounts for presence as a direct concept by only measuring for flow state. In fact, even if one were to agree with Michalidis' presumed choice to default to flow, there would still be a task of evaluating which specific variant of flow survey would be best adopted for study in this thesis. On this matter, Cairns argues that flow is just one constituent element of immersion, but also acknowledges a gap in research to solidify this theoretical model (Cairns, 2018). In general, the overlap between flow and immersion described by Michalidis et al. is largely a consequence of the authors' lack of address to any of the practical merits of choosing one model over another. For instance, no psychometric scale evaluation of immersion is provided, nor of any flow scale. Therefore, in the context of a literature review providing the background to the experiments to follow in this thesis, several evaluations are made to the defence of using immersion as a measurement concept.

First, the IEQ since its original publication, is now one of the most commonly adopted models for the measurement of player immersion, especially in user and player experience focused literature. When compared to other models of engagement discussed in this review, it is also evident that the immersion model is most suitable for specifically measuring the experience, rather than individual traits of engagement, or properties of a particular game.

Second, the means by which the model was conceived also lends it further credibility. This is based on the fact that the qualitative descriptions of immersive experiences were initially obtained from real players of digital games, and this descriptive information was the basis with which relevant theories of user psychology such as flow and cognitive absorption were further incorporated. So far, these are theoretically driven arguments rather than psychometric ones. In this domain, one alternative questionnaire in particular stands out as a considerable candidate, in the GEQ. To the credit of the GEQ, appreciable efforts were made to develop a robust scale with the use of both classical test theory and Rasch modelling of items during scale development. Additionally, the experimental validation of the IEQ was also conducted relative to a non-game stimulus, and it is on this basis that a clearer indicator of the scale's true conceptual validity was provided. Unfortunately, no such validation experiment was conducted for the GEQ. It therefore becomes less evident that the GEQ truly and validly measures engagement as a concept, which limits its capacity to address the research questions in this thesis. On the other hand, the IEQ has also tests of validity of its own in the years since its conception and continues to receive scrutiny by both its originators (Cairns et al., 2014a) and other researchers in this area (Michailidis et al., 2018).

We then conclude by drawing attention back to the definitional overlap of immersion and flow, and here, a simpler argument can be made. Ceteris paribus, on the grounds of attempting to capture a more holistic measurement of player experience, it would appear that immersion is the model either equally or even more likely to provide the more informative measurement. By merit of being designed to address games and the experience of playing games specifically, immersion is designed to at least conceptually tackle measurement in the specific domain of this thesis. Finally, by incorporating wider work around emotion and affect in its validation protocol, immersion has also made efforts to substantiate closer ties with fundamental theories of psychology that likely contribute towards experience.

2.1.8 Summary

In reviewing multiple theories of player experience, the goal was to justify the use of immersion as a basis of measurement, and the necessity of measurement in order to better understand and define player experience.

Through comparison, it has been made evident that theories of engagement and presence are both lacking in their explanatory ability to describe the wider aspects of player experience. From this complex assortment of concepts, it is clear that there is a necessity to disentangle the relationships between lower level function such as attention to the higher level player experiences that are of interest to games researchers and designers.

2.2 What Existing Research Has Revealed About Player Experience

With an understanding that immersion is indeed the most suitable player experience model of choice, the second half of this chapter is dedicated to exploring what previous experimental work has revealed about player experience. In this section, work is categorised by whether an experiment deployed psychometrics or psychophysiology as the primary approach to measuring immersion. Within both categories, the consideration of whether a study is included here or not is depended on whether the study reveals an insight of either player experience, or a broader cognitive facet such as cognitive load or attention. The goal of this section is to present a body of previous work that describes how immersive experiences have been observed, and in what contexts these experiences have been successfully manipulated experimentally.

2.2.1 Research using Questionnaires

Psychometric questionnaires have been the common method of player experience measurement. This is due to their relative ease of deployment in comparison to psychophysiological techniques, as well as a lower barrier to entry given the cost of physiological measurement equipment. What contributes towards a good, rigorous psychometric test is dependent upon an assortment of factors, such as reliability (both internal and external), validity, appropriate standardisation, and critically, the methods by which a test was originally conceived and constructed (Kline, 2000, 2014). In this subsection, experimental literature deploying psychometric questionnaires, especially research using the IEQ developed by Jennett et al. are examined.

Immersion influences non-play cognitive processing

In an experiment validating the grounded theory and immersion questionnaire, Jennett et al. (Jennett et al., 2009) tested the ability of participants to respond to auditory distractors that were specific to either the person, the game, or neither, during a play session of a 2D lab-developed game. This stimulus game was manipulated to be either high or low immersion based on the technological factors, similar to what was described by Slater (Slater & Wilbur, 1997). The manipulations here specifically included feedback in the form of sound effects and visual effects, variability in game mechanics such as size of objectives, and the overall graphical fidelity of the game session. Game and person relevant distractors were better recalled by players of the high immersion game, with irrelevant distractors on average being ignored by high immersion players. Inversely, irrelevant distractors were significantly detected more often by low immersion players. Higher immersion from the questionnaire was also reported by the high immersion players, confirming the applicability of this test, as well as the notion that immersive experiences can be manipulated experimentally. These results also indicated that a form of selective attention and memory processing occurred in participants who were better immersed in the game.

Social presence can contribute towards immersive experiences

Following this, Cairns et al. (Cairns et al., 2013) carried out a series of studies and experiments using the IEQ to explore different factors of immersion. Social presence was evaluated in conjunction with immersion using the IEQ and the Social Presence in Gaming Questionnaire in three experiments. First, participants were led to believe that they would be playing a pong game in three sessions. The first session would be against an AI, then another player online, and then another each other in the same room. In reality, they would play against each other in all three conditions, but the perceived belief that they would be playing against another person online increased reported immersion scores from against AI, and playing in a co-located environment increased immersion scores even further — though the latter difference was smaller and not statistically significant. The second experiment was conducted to dismiss confounds of demand characteristics, and here players played a racing game in either an online and mediated session, or against an AI. Again, a significant difference was found with increased immersion in social play compared to AI single-player play. Finally, to further investigate the original but insignificant difference between online and co-located play, a sample of significantly more players were asked to play the popular game *Mario Kart* on independent but proximally close screens. Here, no statistical difference between online and co-located play were found once again. Further data showing the effect of social play on experience have been presented by Hudson and Cairns (Hudson & Cairns, 2016), where players that win together also experienced higher cooperative social presence than if they lost. This effect, however, also varied based on the specific game played. These experiments reflected previous player reports preferring social play and supported the hypothesis that social interaction mediated changes in immersion within players, often with a positive effect.

Environmental lighting influences experience

The environment in which a player plays has also been found to be a significant factor in immersion. To test this, Nordin et al. (Nordin et al., 2014) asked participants to play a game on a mixed reality tablet version or a computer version, and a traditional console game in reduced lighting. In both experiments, reducing awareness of the environment through the use of a desktop and the reduction of environmental lighting, higher scores of immersion were produced in the IEQ. The second experiment in particular has demonstrated a critical need for immersion experiments to adequately control for environmental confounds. The results here have suggested that the nature of the room in which someone plays a game can have profound effects on their immersive experience.

Uncertainty is an important part of player experience

Using two psychometric questionnaires to examine uncertainty and immersion, Kumari et al. (Kumari et al., 2017) asked players to play a top-down shooter game under different conditions of certainty. Here, the manipulation was the visibility of the game environment and its objects. By reducing the light of the in-game environment, players received less information of the present game state and sub-factors of uncertainty consequently differed significantly between players. Through the PUG questionnaire, player uncertainty of disorientation was higher, and that this change in uncertainty was not reflected in immersion scores from the IEQ.

Uncertainty research in games is a more recently developed field, and this study acknowledged that the relationship between uncertainty and immersion was only exploratory. The work did, however, provide evidence that uncertainty as a concept is measurable by the PUG questionnaire and that uncertainty as a factor in games can be manipulated experimentally. This work by Kumari was followed by a study on uncertainty in games, with the work by Power et al. on developing the Player Uncertainty in Games Scale (PUGS) (Power et al., 2018).

Uncertainty is not something that is captured in the IEQ, nor is it something that is defined as a critical element of immersion in the original theoretical construction by Brown and Cairns. At the time of writing this, I am also not aware of uncertainty being incorporated as a factor or subscale in any newer variants of the IEQ, or similar questionnaires that broadly capture engagement. The findings by Kumari et al., suggest that uncertainty may be a dimension of player experience that is not necessarily involved with immersion. In considering this, it is possible therefore that immersion as a theory of player engagement is also not capturing other elements or dimensions that may be important to the total experience of playing games.

Controls play a significant role in immersion

The rise of popularity in smartphones birthed a new generation of mobile games that required a redesign of user input to accommodate modern touchscreens. These casual games were believed to be drastically different in game design from typical core digital games, and discussions of potential differences in experience arose (GDC, 2011). One of the first obvious differences between mobile and core games is a difference in screen size. To investigate this, Thompson et al. examined differences in immersion in players playing on different sized touch devices (Thompson et al., 2012). In the popular game *Fruit Ninja* where participants use slide touch inputs to slice flying fruits, participants reported higher IEQ scores when playing with a larger screen mobile device.

Controllers are also a primary barrier between the player and agency within a game, and taking elements of player control into account is critical to reaching optimal experience. Cairns et al. sought to investigate the influence of such input modality differences on the experience of immersion (Cairns et al., 2014b). When comparing accelerometer and gyroscope driven tilt controls in a mobile racing game, they found increased immersion measured by the IEQ for players playing with tilt controls in comparison to touch input. Interestingly, when comparing differences on the IEQ subscales, only control as a subscale was non-significant in its difference. The authors argued here, that one possible factor of this difference (or lack of) was due to prior mappings of concept to input, specifically that using tilt controls for steering a car in a racing game is naturally emulating how a steering wheel may function in real life. To explore this idea, a second experiment was run, comparing tilt to two different touch mechanisms (touching and sliding) to see if non-natural mappings would recreate similar effects. Here, a very significant increase in immersion was found, with particularly higher scores in the control subscale for sliding input. This would support the idea that input method is vital to user agency, and therefore by extension to immersion. The authors also noted that users in the sliding condition also enacted input behaviours not consequential to the game mechanics itself by sliding their fingers up and down in motion with the jumping character.

Research on the relationship between game controls and player experience, such as the examples described in this chapter, have highlighted the profound effects of player’s ability to control and execute their intent in a game, and the overall experience of immersion. One reflection is to consider how the findings presented here might fall into the theory of immersion defined by Cairns & Brown.

To this end, take the example statement "I found myself so involved that I was unaware I was using controls". This statement falls into the control factor of the original IEQ. However, this statement is in actuality a composite statement that can be further broken down. A semantic interpretation of this statement could be that the more important element is the term "involved", rather than the subject of "using controls". It is difficult to determine the causality of the relationship between these the experience of feeling involved, and the awareness of using controls. Recall that in the theory of immersion outlined by Brown & Cairns, controls are part of the access barrier to initial engagement. If this taken to be true, then the studies in this section are providing evidence in favour of the view that once the barrier of control is surpassed, it becomes less relevant than the game construction, empathy, or atmosphere barriers to experiencing immersion.

Finally, it is also worth considering the avenues for additional research that could be taken here. What, happens, for example, if the barrier of control or access is suddenly reintroduced at later stage of immersion? It would be interesting to see if upon reaching the state of immersion, players could fall out of immersion by experiencing new or unexpected barriers of control or access. Such work could contribute novel insights towards an improved theory of immersion and game design.

Individual differences in players can influence immersive experiences

Individual differences are also frequently confounds that need accounting for in human experiments, and games research conducted with player participants are no exception. Denisova, as part of her PhD, investigated the common individual difference of players' prior expectations and experience (Denisova, 2016). Specifically, to manipulate player expectations in game behaviour, Denisova et al. investigated player immersion when participants were led to believe that the artificial intelligence of the game was more sophisticated than that of a control condition when playing the commercially popular title *Don't Starve* (Denisova & Cairns, 2015b). Participants who played with a prior assumption that an advanced AI was present would consistently record significantly higher IEQ scores, despite an identical stimulus game in both conditions. This experiment displayed how considerable changes in player experience can occur purely based on the individual player's assumptions.

Another common individual difference in adaptive difficulty is the ability or skill level of the player. To investigate the effects of adaptive games on player experience, Denisova et al. dynamically adjusted difficulty of a game by manipulating time constraints to player performance (Denisova & Cairns, 2015a). Through this simple manipulation of challenge which went undetected by participants, Denisova et al. produced significantly higher scores in player immersion in players playing the game with dynamically adjusted timing compared to those playing with a standard timer. This result interestingly aligns with the findings by Jennett et al. in the original validation experiments of the IEQ (Jennett et al., 2008), where negative valence emotions could potentially be contributing positively towards immersive experiences.

Denisova also studied another common individual difference in the form of player preferences in UI and gameplay settings. Such preferences are significant enough to warrant most modern games to be published with options menus that allow players to customise game features to their preferences. To investigate player preferences in camera perspective, Denisova & Cairns compared player immersion in the acclaimed shooter game *Skyrim*. Players recorded significantly higher IEQ scores in first person play. By further use of the IEQ in experimental research, the body of work by Denisova has revealed many considerable confounds on player experience based on individual differences in the psychology of players.

Games with procedural content may introduce confounds that should be tracked

One of the games used by Denisova in her research was *Don't Starve*, a title which utilised procedural content generation. Procedural generation has become a popular game mechanic in the contemporary market, and in an experiment to investigate its effects on immersion, Connor et al. compared the use of procedural content generation against traditional human design (Connor et al., 2017).

Here, Connor et al. used the IEQ to compare immersion scores in an otherwise identical game that utilised procedural content generation or content created by a human designer for the levels of the game. No statistically significant result was found, though both a parametric ANOVA and a non-parametric Kruskal-Wallis test found results bordering on non-significance and significance, respectively.

Upon further inspection, some of the choices in this experiment resulted in limitations to the analyses. For instance, in an attempt to further explore the IEQ results at a finer grain, the authors opted to inspect differences between the groups for individual questions and hand select statistically significant items as potential justification for treating the borderline results as statistically significant. This might be an acceptable practice, if they had also inspected and found supporting evidence on the existence of clear subscales within the IEQ, but this was not done.

Furthermore, no management strategy for over-testing such as a correction for multiple comparisons was applied. In fact, at the design level of the experiment, the authors opted to compare a condition of procedural generation and human design. However, the information reported here lacks details on the steps taken, if any, to control for the confounds of this experimental manipulation. For example, the quality of implementation between procedural and human generation is a matter that is not only subjective, but interwoven with factors such as the expertise of the designer and the time invested in design. Importantly, this element was not adequately measured with, for example, a scale or even bare-bones Likert questions to capture a metric of procedural generation quality. Decisions regarding quality are multidimensional and complicated, requiring a bare minimum of quantifying why certain elements would have been chosen over others.

Overall, Connor et al. explored a very interesting and well motivated question. The usage of the IEQ's individual questions here draw attention to a need for careful conduct when considering the multidimensional nature of the questionnaire. For the work in this thesis, the lesson to be drawn here is to consider the statistical implications of testing many items in a single battery. The subsequent family-wise error rate would require appropriate management, and in certain cases it may be better instead to treat the IEQ as a unidimensional scale in order to minimise the negative impacts on statistical power. Finally, lessons should be taken from the potential issues here, around the subjective judgements of experience that surround immersion. In cases where it is possible, additional measures should be taken to try and capture other experiences peripheral to immersion that may potentially end up being important.

Making HUDs diegetic does not necessarily improve immersion for everyone

Another common and varying factor of game presentation is the user interface. To examine how differences in user interface could influence immersion, Iacovides et al. examined heads up displays (HUDs) as diegetic factors of play, where diegeses is defined as an element of the game either presented as a part of the world of the game and its narration, or outside it (Iacovides et al., 2015).

Iacovides et al. manipulated the HUD in the popular shooter game *Battlefield 3*, by either removing the display entirely or maintaining its default settings. Using the CEGEQ questionnaire designed to record the base necessary requirements for a game to produce positive playing experiences, they found that participants did not record statistically significant differences in CEGEQ scores after playing both versions of the game (e.g., differences in core experience). However, the presence of considerable effect sizes and a small sample of only 9 participants indicated that such a value was possibly the result of lacking experimental power.

Post-play interviews also revealed some players reported a preference for the non-HUD diegetic version of the game. To further evaluate this line of enquiry, 24 participants were asked to play the same versions of the game, though this time in a between groups design. Immersion was recorded through use of the IEQ and players recorded higher scores in the diegetic, non-HUD version of the game than the default non-diegetic HUD version. This difference, however, was only observed in players self-reported to be experts of first-person shooter games that played such games at least an hour per week. The implication here is that upon obtaining some expertise of a game, fewer non-diegetic elements of play are necessary for the same level of play and thus players are able to play a game in a presented environment where the HUD is less able to present a barrier for immersion.

This could be argued to be reflected by the statistically higher IEQ subscale scores for all factors except for real world dissociation. Interestingly, results from the IEQ also observed statistically significant differences in immersion as measured by the IEQ rather than engagement measured by the CEGEQ. It is rather awkward that real world dissociation was the sole factor that was not significantly different, as one might assume that a diagetic HUD would further facilitate the dissociation of the player from their real world.

There are some potential explanations for this, such as again, the lack of adequate experimental power. However, if the issue of power is temporarily set aside, a more interesting interpretation arises– perhaps the diegesis of HUD elements do not exist in a vacuum. By removing significant proportions of the HUD, Iacovides also removed information that may have been valuable to the player's ability to play the game. This would also explain why the overall IEQ scores were lower for novice participants playing with a diagetic HUD, compared to IEQ scores being higher for expert participants playing with a diagetic HUD.

The finding here further elaborate a picture of the influences of game construction on player experience. Furthermore, the IEQ specifically also presented results that could be interpreted as tying diagetic game elements to immersion, and that somehow, this is not related to real world dissociation. Importantly, there are also lessons to be drawn from the fact that manipulations of seemingly isolated elements of the game could have unexpected consequences on a player's experience, and these consequences could also occur contextually based on other factors such as the player's expertise with a game or genre.

Player behaviours can provide additional objective data

In addition to psychometric tests such as the IEQ, the viability of supplementary measurements such as objective behavioural outputs have been suggested. Prior to the PhD work of Nordin (Nordin, 2014), there was a belief that among potential behavioural measurements of immersion, the particular possibility that a player's perception of time could be a viable metric in describing their degree of engagement. Indeed, perception of time is explicitly mentioned in many models of experience, such as flow/GameFlow (Sweetser & Wyeth, 2005), engagement (Brockmyer et al., 2009), and immersion (Brown & Cairns, 2004).

Through the first five studies published in his thesis, Nordin demonstrated that players' ability to perceive and accurately report their perception of time was remarkably similar between different degrees and conditions of immersion. The reasons for this included the difficulty in correctly recording time perception under a retrospective paradigm, which is inherently dependent upon participants' memories as well as their willingness to engage with a psychometric test's questions on time.

Furthermore, Nordin et al. also state that there are also limitations in such experiments which do not reveal details of how participants are able to experience time when playing games due to the natural inability of the experimenters to inform participants of the true, deceptive nature of the experimental conditions.

Nordin et al. concluded that retrospective time perception tasks were unsuitable for the particular context of games research for these reasons, arguing that traditional time perception experiments in cognitive psychology were not directly translatable to a games research environment. Of additional interest are also the results Nordin et al. collected with regard to the relationship between time perception and immersion. He found that while each experiment directly manipulated immersion and produced significant results in player experience of immersion, measures in time perception consistently tested to be not statistically significant. This would indicate, according to Nordin et al., that there is a dissociation between immersion and time, although this may also simply be a measurement error in how time perception was recorded.

Based on Nordin et al.'s data, it also appears that perception of time seemingly diminishes as individuals play games under any condition. Grounded theory work done based on data collected from players also suggested that while players still perceived time, they opted not to record or pay attention to time in detail while playing games. From the perspective of the goals of this review, the multiple null results found by Nordin et al. are valuable. This is particularly the case due to the consistency under which they were produced, providing persistent evidence for his claims. More significantly, the work done by Nordin provides a context for the current goals of research stated in this review, showing that the value of recorded behavioural data through telemetry could likely be more reliable than self report data provided by participants.

Post-Play Reviews

In some of the studies discussed in this section, there is use of post play evaluations of participant experiences. This is another form of supplementary data that's used in combination with psychometric tests and measurement data in general. Gow et al. presented a pilot study using a paradigm in which participants were asked to play a shooter game followed by a questionnaire and a post-play commentary (Gow et al., 2010). Efforts were taken to guide participants without over prompting or over reliance on unreliably cognitive faculties such as the long term memory. During commentary, players were provided with a recording of their play and qualitative data was acquired in addition to the psychometric data. Gow et al. also coded these experiences into a category, emotional valence, and general set of original codes. The resulting piece of work is the presentation of a method that can be useful when researchers would benefit from a deeper inspection of their participants' experiences.

Similarly, Kivikangas et al. (Matias Kivikangas et al., 2011) have taken this technique into an psychophysiological experimental framework. Here, post-play commentaries are carried out with automatically generated recordings of particular events of interest, as well as time synchronised points in physiological data. The details of such physiological methods and the utility of this paradigm are discussed in the section below. The benefit of the approach used by Kivikangas et al. is that the use of physiological data provides a convenient avenue for detecting important points during play that can lead to relevant and useful post-play interview subjects.

The takeaway is that much like behavioural measurements, post-play reviews also provide a means to augment data captured by questionnaires. In many cases, they provide new information of participants' experiences, and effectively facilitate the formulation of new research questions for subsequent work, as was shown by Gow et al.

Ecological Validity & Commercial Games

The questions regarding how immersion can be manipulated or broken are vast, and in a way answering these questions is being proactively involved in design. McMahan et al. presented considerations for the use of games as research instruments following an experiment using the commercial game

Mario Kart Wii (McMahan et al., 2011). In particular, they discussed the trade-offs of ecological validity and development accessibility between commercial games to 'toy' lab created games for the use of experiments.

In the defence of lab games, the choices of manipulations made are often what would be considered design decisions by traditional game developers, yet they are considered inferior due to lower visual fidelity or missing narrative qualities. A poignant argument put forward by McMahan et al. was that ecological validity can be obtained by the presence of some population that plays the selected game. However, there are also some issues with this position.

For instance, does the lack of a present day community for a game indicate that a game is outdated and therefore no longer appropriate for research? If that is the case, then many published studies have already failed to achieve appropriate ecological validity based on the time of their data acquisition. Also, what about the applicability of previously published, older research to the present day? Are there qualities of player experience that are specific to a period of time?

The crux of the matter is that at present, there appears to be mostly qualitative assessments available in order to justify what may make a game ecologically valid or not. Ecological validity is important to measurement, as data made invalid by the conditions of measurement present a problem for any proposal of a generalisable measurement of immersion. Therefore, there is also a motivation for games research to present a logical framework with which a game can be justified as ecologically valid, and the minimal requirement to meet at present appears to be the use of a commercial game, or a very close replication of an existing commercial game such as that developed by Cutting in his work (Cutting, 2018).

2.2.2 Questionnaires continue to be a valuable way of studying player experience

Overall, psychometric questionnaires (with the IEQ in particular), have been deployed successfully as measurement tools in games research. Over the past decade, they have been paramount to developing a substantial and growing knowledge of player experience, as well as the factors that modulate immersion in video games.

2.2.3 Psychophysiological Research

Psychometrics alone cannot provide a comprehensive description of user experience. They rely on accuracy and adequate communication of self-reflection, which is difficult to validate. Furthermore, they are restricted to a single dimension and are unable to account for changes in cognition or behaviour over time, making them inappropriate for administration during live game play. One purported answer to this problem is the use of psychophysiological measurements as a correlate for immersion.

Reviews have been published presenting the ongoing viability and use of techniques in HCI (Dirican & Göktürk, 2011) as well as games research, specifically (Kivikangas et al., 2011). These techniques can be and often are applied in conjunction with psychometrics in order to obtain more robust measurements of user experience.

In this section, the different methods of measuring human psychophysiology are briefly explored, with definitions and understanding sourced from John Cacioppo et al.'s psychophysiology manual (Cacioppo et al., 2007).

Throughout this section, studies that have employed the use of various physiological modalities are mentioned, both in the respective subsections on each modality, as well as in a general section where many modalities were used in conjunction, as is common practice in this domain.

Pupillometry & Eye Tracking

Here, the subject of pupillometry and eye tracking will be discussed. Pupillometry and eye tracking are each distinct methods that both capture information from the same source, namely the eyes. Eye tracking is the method of recording eye motion and gaze locations across time and task (Carter & Luke, 2020). Pupillometry on the other hand, is the study of changes in the diameter of the pupil as a function of cognitive processing (Sirois & Brisson, 2014). In this section, both pupillometry and eye tracking research in games will be reviewed in order to explore the potential utility of the method for the following studies in this thesis.

First, a brief overview of the eye and its mechanisms are necessary. There are four primary eye movements (Purves et al., 2001): smooth pursuit where the eye smoothly tracks target stimuli as they move in the visual field, saccadic movements that rapidly shift the eye from one point of visual field to another, vergence that aligns the two eyes together, and vestibulo-ocular movements that maintain visual stability during movement of the head and body. These movements can be captured by modern eye tracking equipment through the recording of the pupils. If an individual needs to move their visual attention to a new target, a saccade would take place, and in turn the pupil would move towards the new subject of attention. Inversely, if one is attending to stimuli continuously, then a fixation takes place whereby the pupil remains still and fixed on the target. These enable the capture of measurements such as fixation duration, and fixation/saccade frequency, which can be manipulated in experiments to examine if they vary based on stimuli induced demands.

An additional eye movement that is of interest to cognitive research are blinks. There are three types of blinks (Stern et al., 1984). Reflex blinks happen as a protective response to potentially dangerous stimuli. Voluntary blinks similarly occur in response to some stimuli, but they are distinguished by the nature that they occur by choice of the individual. Finally, there are non-blink closures, which are eye closures that occur as in contexts such as sleep. Such non-blink closures are distinguishable from the other two types of blinks by the time taken to close the eyelids, and the duration of closure (Stern et al., 1984). Metrics of blink behaviour can also be found in the form of blink rate, blink duration, and blink frequency (Stern & Skelly, 1984). Because blinks have also been found to vary based on specific contextual demands, more involved blink metrics have also been tested. For example, the proportion of blink frequencies in relation to saccade durations have been found to be variable based on cognitive and perceptual demands during reading (Orchard & Stern, 1991). In general, blinks are another form of eye movement that have been tied to cognitive activity, and are therefore of interest to the research question in this thesis.

In addition to eye movements, the pupils themselves are also capable of physical adjustments in response to demands. The pupils have a dilation and constriction response to dark and light ambient conditions, respectively. This pupil dilation response has a reaction time of approximately 200ms (Sirois & Brisson, 2014). These changes are managed by pupillae– sphincter muscles that either constrict the pupil, or the opposite dilator muscle. These muscles are in turn neurologically connected to the brain, and it is through this connection that the dilation response is functionally tied to sympathetic activity and, in turn, cognition (Cacioppo et al., 2007). Based on this, functional changes in pupil diameter have been investigated in different cognitive contexts, such as attention, emotion, and stress. The key measurement in pupillometry therefore, is the pupils' diameter. Typically, this is taken as pupil diameter change, whereby differences in normalised pupil diameter between stimuli are calculated to test for stimuli induced pupil diameter responses (Sirois & Brisson, 2014).

Cognitive load in particular was an early subject of study for pupillary correlates, with a review by Beatty (Beatty, 1976) presenting 20th century cognitive research with early eye tracking. Cognitive overload experiments vary cognitive load by testing participant ability to process and recall varied length digit spans, and such experiments were introduced by Peavler (Peavler) with later experiments finding that pupils dilate during processable loads but constrict following overload (Granholm et al., 1996; Poock, 1973).

Similar digit processing experiments, including a digit sorting experiment inducing sustained cognitive load, have also found that blink rates converge around periods of change in cognitive load, with sustained dilation during sustained load (Siegle et al., 2008). Though these controlled experiments may appear less pertinent to general immersion research, they are among the earliest records presenting the use of pupillometry as a metric of cognitive function. There are important differences between such psychological experiments and games research, the most notable of which are the stark differences between task evoked stimuli and a typical game play session.

In HCI research, task evoked pupillary responses have been correlated with cognitive load during use of virtual driving simulators. Pupil diameter has been found to increase as a function of cognitive load induced by digit sequence errors during a driving simulator (Palinko & Kun, 2012). Mean diameter changes have also been found to increase as a function of cognitive load during simultaneous vehicle following and auditory stimulation, (Palinko et al., 2010). Using the pupillometric Index of Cognitive Activity (ICA), which measures cognitive load through dilation reflexes relative to short time intervals (seconds), Schwalm et al. found significant increases in ICA during lane changes, with even greater increases during lane changes and simultaneous visual search (Schwalm et al., 2008). Unfortunately, the original publication for the ICA appears to be a closed patent, making it difficult to judge the applicability of this signal processing technique for pupillometry. Most demonstrably, this pupillometry in HCI drew attention to its potential applicability for similar work in games research (Cox et al., 2006).

Returning then to eye tracking of eye movements, there has also been similar interest in measuring eye movements in order to investigate cognition. Saccades and fixations have become primary indices of interest, particularly in visual attention research. This is due to the acuity of the visual field, which deteriorates significantly as a function of distance away from fovea (the centre of the eye). As a result, humans must redirect the fovea to stimuli of focus in order to function with optimal visual performance. Saccadic eye movement volumes and patterns, and fixation volumes and durations have become well established indices of attention and cognition. Pupillometry is particularly advantageous in that it provides reliable measures even without conscious awareness on the part of the participant. A psychological overview of the relationship between saccadic movements and cognition can be found in Liversedge and Findlay's review (Liversedge & Findlay, 2000).

One of the earliest applications of eye tracking to commercial games I could find was a pilot study in 2009. Here, Renshaw et al. applied eye tracking to study player engagement in the game Tomb Raider between a difficult and easy level (Renshaw et al., 2009). This was done by measuring fixation durations throughout sessions of play. During play, they also intermittently asked participants how they were feeling in a manner that interrupted participants' play. The eye tracking in this study was restricted to fixation duration and location, which were graphed into cells divided across screen-space demarcations. Players fixated mostly in the central region of the screen. The research itself draws limited conclusions due to the nature of the small sample size, but did suggest that cognitive overload could be inferred from player inability to respond to the interrupting questions during play. More recently, the PhD research of Joe Cutting (Cutting, 2018) has involved numerous experiments with the use of eye tracking to study immersion. Cutting conducted a series of experiments with both pupillometry and eye tracking, and in particular was interested in the use of eye tracking to measure fixation patterns in order to develop and refine his distractor method of measuring attention during play. His thesis provided evidence to conclude that while promise exists in the use of such techniques, the design of a game may limit whether eye tracking is a suitable correlate for immersion. This was especially the case for self-paced games, which were the subject of Cutting's research.

Cutting's experiments specifically used self-paced games which permit the player to play under their own whims and speed. For this, a lab developed game mimicking the mobile title *Two Dots* was used, which is a simple game that required players to search for patterns of three identical items among an array of similar items. In a more complex version of the game, items above those selected by the player fall after actions, requiring players to adopt spatial planning strategies in order to find optimal scoring opportunities. When comparing between a hard and easy version of the puzzle manipulated by the using the aforementioned versions as conditions of play, a significant and large increase in pupil dilation was found in a 1-second window approximately 3 seconds before participant responses. Significantly fewer fixations, blinks, and saccades, and lower saccadic amplitudes were also found during the hard condition. As the use of *Two Dots* here only required a single input from participants, Cutting carried out a subsequent experiment. For the second experiment, three versions of the game were used: a replication of the game with players joining dots on the screen in order to reach an objective within a limited number of actions, a no-objectives version of the same game that permitted participants to play freely, and a control version of the game that contained one item identical across the array. Here, no significant difference was found for a 1-second window 1 second prior to response, but a significant difference was found immediately following response. The IEQ, which was used in conjunction with eye tracking, also recorded significantly higher immersion and challenge between conditions. The goal of Cutting's eye tracking experiments was to investigate whether experience could be measured with the technique, but some questions arise from the methodology.

For example, time series plots of pupil dilation for both experiments show patterns of variance that differ drastically on visual examination, yet the statistical analyses were primarily fixated on short periodic windows. While these analyses were well justified and well defined, they opted to ignore a dimension of the data that is meant to be of significant value to eye tracking as a methodology. Given previous literature, the question of whether eye tracking metrics (such as saccade volume, frequency, etc.) or pupil diameter metrics (such as change/difference) produced any significant results is left unanswered. Another question is that of the game used. Self-paced games are not necessarily atypical games, Cutting does cite popular commercial titles such as *Civilization* which are commonly considered to be core games. However, the particular game used and its demands of the participants mechanically are considerably limited in comparison. Cutting acknowledged this and argued that there was potentially a lower demand of cognitive effort because of the game, which critically highlights the difference between immersion or a general player experience, and cognitive load or processing. If this is the case, then eye tracking is most likely a technique incapable of measuring immersive experience on its own merits. This is not to discredit the nature of this work as initial exploration and investigation, these experiments do establish foundations which are informative for subsequent work using the same approaches.

Work by Strauch et al., has also applied pupillometry to explore changes in pupil dilation in response to difficulty (i.e. challenge) variations in a within-groups experiment where participants played pong (Strauch et al., 2020). They found that pupil diameter increased as difficulty increased, though the function was more quadratic (or specifically, inverted-u shaped), in line with theories of flow. Interestingly, they also found that for some participants, this inverted-u shaped relationship between pupil diameter and difficulty did not appear. This would suggest that the prior exposure and/or skill of the player has a significant influence over how they physiologically respond to a game. This interpretation would be in line with other studies discussed in this chapter that also explore the relationship between challenge and physiological responses, such as that of galvanic skin response in the following section.

Beyond academic publications, there is also a noted commercial interest in this research. Ambinder et al. carried out a series of experiments to test the viability of using physiological signals, including eye tracking, in commercial games as input and feedback devices (Ambinder, 2011). In particular, eye tracking was used to test input method viability for a first person puzzle game *Portal 2*, and proposed the general use of all psychophysiological signals as a means to derive metrics of arousal and enjoyment from games in order to feed game design and AI. From the time that this review was originally written to the subsequent submission and corrections, this commercial interest has only appeared to have grown.

Pupillometry and eye tracking have also become increasingly accessible, with improvements in access through new developments such as open source hardware and software (Kassner et al., 2014). There also exists a developing research community looking at the use of machine learning to capitalise on the rich, high volume data (Fridman et al., 2018). Summarily, there exists enough evidence to justify exploring its potential utility to measure immersion, and as has been discussed, some initial work doing so has already taken place.

Electrodermal Activity

Skin conductance (known in the past as galvanic skin response) is often used as an inferential measure of cognitive function. Now commonly referred to in research as electrodermal activity (EDA), usually scaled with microsiemens as the unit of measurement. The signal is sourced from the skin, modulated through the sweat ducts on the body that act as resistors. As sweat fills these ducts, conductance changes within the skin can be recorded using EDA measurement devices, usually in the form of electrodes placed on the surface of the skin (Cacioppo et al., 2007). As it is a measure of resistance, two electrodes are placed in proximity to one another in order to measure conductance. A constant current and voltage between electrodes allows for the measurement of resistance. EDA measurements are primarily comprised of a slow, tonic level that changes gradually, and a fast, phasic shift that occurs in response to stimulation. A summary of common EDA measures and their typical values can be found in table **??**.

Measure	Definition	Typical Values
Skin conductance levels (SCL)	Tonic level of electrical conductivity of skin	$2 - 20\mu S$
Change in SCL	Gradual changes in SCL measured at two or more points in time	1-3 μS
Frequency of NS-SCRs	Number of SCRs in absence of identifiable eliciting stimulus	1-3 per min
SCR Amplitude	Phasic increase in conductance shortly following stimulus onset	0.1-1.0 μS
SCR Latency	Temporal interval between stimulus onset and SCR initiation	1-3S
SCR rise time	Temporal interval between SCR initiation and SCR peak	1-3S
SCR half recovery time	Temporal interval between SCR peak and point of 50% recovery of SCR amplitude	2-10S
SCR habitation (trials to abituation)	Number of stimulus presentations before two or three trials with no response	2-8 stimulus presentations
SCR habituation (slope)	Rate of change of ER-SCR amplitude	0.01-0.5 μS per trial

TABLE 2.2: Common SCR (EDA) measures and typical values for each measure in healthy, young adults taken from Cacioppo et al., 2007. Key: SCL, skin conductance level; SCR, skin conductance response; NS-SCR, nonspecific skin conductance response

Like many physiological signals, EDA has many confounds. The properties of EDA signals will also vary based on the site of recording, with common electrode placements on the hand and fingers; specifically the fingertips (distal phalanges), mid-finger (medial phalanges), and opposing sides of the palms (thenar and hypothenar). However, these norms in psychology and neurophysiology do not translate well to HCI and games research, where both hands are often needed for use of devices. van Dooren et al. compared skin conductance across 16 locations on the body in order to determine which signals were most similar to that found on the hands, with highly correlative signals found on the feet, forehead and wrists (van Dooren et al., 2012). The use of preparation substances such as gels or alcoholic cleansers have also been found to change conductivity significantly, as does the temperature of the environment and hydration of the participant (Cacioppo et al., 2007).

The pathways to the eccrine sweat glands are complex and manifold, and the skin conductance response (SCR) is sensitive to an array of stimuli properties including novelty, surprise, and significance (Cacioppo et al., 2007). As a consequence, it is difficult to determine a single psychological process that the source of a SCR signal. Furthermore, as games experiment paradigms do not control psychological processes in the same manner as cognitive sciences, it is also difficult to follow suggestions of inferring through confound controls as is suggested by Cacioppo et al. It might, for example, be reasonable to hypothesise that SCRs could be observed based on in-game events of intensity or surprise, but deciding between a stimulus and non-stimulus SCR becomes a difficult analysis in such an experiment. Therefore, tonic SCL (skin conductance level) activity might be more appropriate for games experiments.

Particularly in the case of task activity and performance, SCL has been found to increase consistently when participants were under 'energy mobilization', which could be suggested to be a consequence of attentional load or stress (Cacioppo et al., 2007). Similarly, SCL activity has also increased under situations of emotional arousal, and has since obtained a label as a measurement of emotion; a review of EDA in games research by Westland explicates the metric as a measurement of emotion (Westland, 2011). This idea of emotional significance in the sweat response is largely due to the nature of the eccrine glands, in that they are entirely dictated by the sympathetic nervous system. As such, stress, fear, and anxiety can all be seen as causes of a thermoregulatory or hormonal response, based on the pathways of the hypothalamus, sympathetic ganglia in the spine, and the well known fight or flight response. Cognitive load has also been measured previously using EDA in an experiment using arithmetic and reading as tasks, and mean and accumulative EDA were found to be higher with increasing task difficulty (Nourbakhsh et al., 2012).

As a modality, EDA has been a common choice in games research of player experience. For example, Klarkowski et al. measured EDA from the palm, while changing the level of challenge experienced by participants within a game play session of a commercial game. Levels of challenge were defined by the relative demands of the game in relation to the skills of the player. In the boredom condition, the level of challenge was trivial to the player. In the balance condition, there was an appropriate amount of challenge relative to player skill. In the overload condition, the level of challenge greatly surpassed the comfortable thresholds of the player's skills. In a considerable sample size of 90 participants, they observed that EDA increased as challenge increased over all three conditions, showing that EDA was potentially suitable as an index of player's experiences of challenge.

Challenge has also been similarly studied by Nacke and Lindley, who measured electrodermal activity (though here it is referred to as GSR) of participants playing different levels in the commercial game *Half Life 2* (Nacke & Lindley, 2008). By defining subjective design criteria for how each level is constructed, Nacke & Lindley designed an experiment where participants experienced states of boredom, immersion, or flow, induced by different levels of the game. The results showed that logarithmically normalised galvanic skin response either increased when players played the level designed for flow experience, or decreased when players played the level designed to induce boredom.

In summary, electrodermal activity appears to have great utility as a summative measure of at least certain aspects of cognitive load, and in the context of the research question in this thesis, player experience with good existing evidence for a physiological relationship with challenge. Thus it is evident, that EDA is a viable measurement for cognition in modelling player experience for immersion research.

Cardiovascular Activity

The heart is intricately interconnected within the body, and of specific interest to psychologists is the relationship between the heart and the central autonomic network (CAN). This network is involved in the autonomic management of heart activity, connected to regions of the brain involved in emotion regulation such as the amygdala, and sensory input systems. As such, the CAN can be described as the system with which the heart is regulated based upon the information of intrinsic and extrinsic sensors (Appelhans & Luecken, 2006). Psychophysiologists draw behavioural inference from cardiac activity through electrocardiography (ECG) which measures the skin surface electrical activity of depolarisations in the heart. Measurements are typically observed across cardiac cycles, a single iteration of which starts at the end of diastole, the filling of the heart with blood, and ends with systole where oxygenated blood is pumped into circulation across the body. The recorded data are waveforms describing specific events of a cardiac cycle often labelled in the PQRST nomenclature that describes polarisation of the pacemaker node (P), activity in the myocardium or heart musculature (QRS) and the re polarisation of the ventricles (T) (Cacioppo et al., 2007). By analysing the waveforms of these events, researchers have been able to explore theories of the relationship between psychological function and heart activity 2.2.

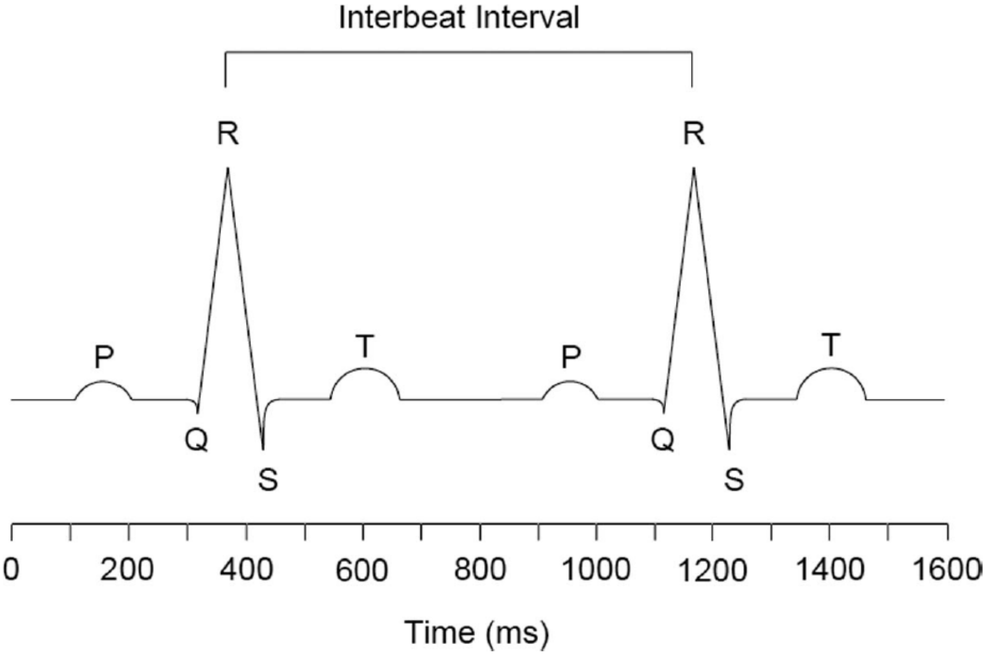


FIGURE 2.2: A hypothetical example of the cardiac cycle, here labelled as an interbeat interval. The interval in this instance, is defined by the R spikes which mark the depolarisation of the ventricles (Appelhans & Luecken, 2006).

In HCI research, heart rate and specifically heart rate variability is a common measurement inferring emotional activity (Dirican & Göktürk, 2011;

Kivikangas et al., 2011). Heart rate variability is simply, during normal sinus rhythm (Heathers). The underlying premise of using HRV is that it provides an index for the change in heart rate over a period of time, specifically in the form of variance of the distances between successive heartbeats. Where heart rate has been previously correlated with cognitive processes, heart rate variability purportedly indices the change in this process over time. This variability of the inter-beat interval has been related to both emotional (Appelhans & Luecken, 2006) and cognitive regulation enabled by the vagus nerve (Thayer & Lane, 2009).

Thayer et al. theorise that based on an ability to directly modulate cognitive performance, there exists an adaptive and flexible neural network involving the heart that is able to mediate cognitive efforts. There exists two theories modelling how this might take place, the polyvagal theory, and Thayer et al.'s neurovisceral integration model, which has accumulated empirical evidence over time. According to the model of neurovisceral integration instead, a regulatory system manages cognition, behaviour and physiology. Emotions are the supposed result of various combinations and interactions of these three sub-systems. Through the connection between the CAN and heart, cardiac activity is modelled as an extension of CAN activity, reflecting the ongoing management of cognitive and emotional states.

Heart rate variability has become more prominent in research (Heathers) in part due to the modern availability of heart rate monitoring equipment, often in the form of smartwatches or straps that contain an optical sensor detecting blood light absorption to infer pulse rate. Because such devices are more accessible than traditional electrocardiographs, they have become popular in HCI research (for example, by (Drachen et al., 2010)). However, the accuracy of such devices to ECGs are variable and care must be exercised in selecting the appropriate equipment (Gillinov et al., 2017). They do, however, provide consistent data for a general pattern of activity which is sufficient for HCI research uninterested in medically precise waveform analyses (El-Amrawy & Nounou, 2015).

Studies using multiple modalities

Experiments in HCI and games will often adopt multiple psychophysiological measurement techniques in conjunction. Such an approach seems only logical, as more data can only mean more opportunities for comprehensive coverage of a concept of interest. Select examples using this combined approach are discussed here in order to better inform using a similar multimodal design for experiments in this thesis.

Ravaja et al. examined the physiological indices of emotion during play of the game *Monkey Bowling 2* in an event related design (Ravaja et al., 2005). ECG, facial EMG and EDA signals were recorded to examine the physiological response to 4 different negative events. These data were strictly phasic and described increased emotional valence and arousal activity in response to negative events, particularly that a positive emotional response was elicited by some clearly negative game events, indicated by reduced corrugator and increased zygomatic and orbicularis oculi EMG activity. However, in other negative events marked by participant feedbacks, negative emotional signals were produced. The authors here note the potential insensitivity of facial EMG when considering the incongruence of emotion and event; however, they also suggest the potential involvement of game design instead as another potential factor. It is possible that a cheerful or comical presentation style present in the game could have produced from players an emotional response that mitigated the disappointment of a strict failure. This would be supported by the contrasting negative emotional signal to clear negative player feedback after a failure, indicating that some form of emotional management took place during play or that players genuinely found humour in their failures.

Drachen et al. carried out an experiment correlating heart rate and electrodermal activity to self reported player experience (Drachen et al., 2010). The psychometric test for experience used was the Ijsselsteijn GEQ, which correlated negatively with heart rate (the lower the heart rate, the higher the reported immersion), and higher heart rates with scores of negative affect from the GEQ such as tension. The authors proposed that a high heart rate is indicative of player frustration, but this is a questionable claim for two reasons. First, there is the argument that positive tension and negative affect scores are indicative of frustration; but if this was the case, then the dimension of challenge should also have been positively rated and players should not have reported the feelings of competence that they did. It is possible that some frustration existed that was not related to the challenge of the game, but such a case would be indicative of a problem with a game or system, thus calling into question the design of the experiment. Second, even if a high rate correlates with negative affect in first-person shooters as the authors suggest, it would be another step in inference to relate this to frustration; for example, there is a requirement for empirical data manipulating specifically tension as an independent variable in order to observe states of frustration. In other words, the term frustration here is used with little formality to defining concretely what it means. To the credit of the study, however, the fact that such inferences can be drawn from heart rate data alone suggests that there is a potential for this modality to provide meaningful data. Similarly, Drachen et al. also found a significantly positive relationship between EDA and negative affect and a negative relationship with challenge, immersion and competence. In this case, the authors considered challenge to be a multidimensional concept that required further investigation, but it is also possible that the GEQ does not appropriately capture the nature of challenge in a way that is consistent across different contexts. Perhaps more notable than either physiological investigations, was the involvement of constant interruption of participants during play in order to register the GEQ. This was done so that multiple points of recording were available from the GEQ to be correlated with physiological data. However, no details are provided on the stability of GEQ scores over the time course of the experiment, restricting any interpretations on whether psychometric scales might potentially be suitable for very coarse temporal measurements of immersion.

There have also been attempts made with a larger array of physiological measurement modalities. Specifically in games research, work by Nacke & Lindley used EEG, ECG, EMG, EDA, and eye tracking to measure flow in first-person shooters (Nacke & Lindley, 2008), though the EEG and eye tracking data were not included in the final analysis of this study. In this experiment, engagement and immersion were also measured using the GEQ, and the stimuli were a modified form of the commercial game *Half Life 2*, with experimenters manipulating the stimulus to induce a state of boredom or flow based on the level design. Unlike the experiment by Connor et al. (Connor et al., 2017), transparency is provided on details of design choices made to induce each mental state. Like many other studies, analyses were conducted using aggregates of measurements taken over the course of the experiment. The EMG data were used to infer valence within participants through orbicularis oculi activity, and a statistically significant difference between boredom and flow states in play. A similar difference was also observed for EDA between boredom and flow.

Chanel et al. used combined facial EMG, EDA, cardiac activity and respiration to investigate physiological compliance between cooperative and competitive play (Chanel et al., 2012). Here, compliance is defined as the relationship of physiological signals between multiple people inferring social experience in a multiplayer game. Compliance was computed between dyads separately for each signal, with EMG separated into zygomaticus, orbicularis oculi, and corrugator supercilii, and cardiac activity indexed by inter-beat intervals. Significant relationships between compliance and self-reported social presence was found for zygomaticus EMG, orbicularis oculi EMG, and inter-beat intervals. Further, compliance was higher for dyads playing competitively rather than cooperatively. Orbicularis oculi EMG in particular was found to be strongly predicted by the social GEQ subcomponents, which is explained by the involvement of the muscle primarily in smiling; thus if players smile together, they are more likely to report positive experiences in a questionnaire. Chanel et al., concluded that these results demonstrated a relationship between physiological compliance and the social context in which players interact with one another.

EMG and EDA joint experiments have also been conducted by Nacke et al. (Nacke et al., 2010), who demonstrated over three experiments that there exists a positive effect of sound on the engagement of players as measured by the GEQ. More pertinent is that they observed a lack of statistically significant effects of sound on the tonic physiological signals used in the experiments. These tonic measurements were effectively aggregates of the measurements taken over the course of the experiment, analysed with an ANOVA, and no effects in either tonic EMG or EDA were observed. The conclusion drawn by Nacke et al., was that an exploration of phasic signal properties may be more insightful in revealing any relationships with engagement (or immersion).

In a study utilising virtual reality to elicit arousal, McCall et al. observed electrodermal, heart rate, psychometric measurements, and post-hoc commentary self reports (McCall et al., 2015). The goal of the study was to investigate the role of memory in managing intense experiences and recall. To accomplish this, participants were asked to rate their arousal during a full visual-audio playback session following the VR experiment. Significant correlations between participant self-report and physiological signals were observed, especially during the most intense periods of the experiment (peak correlation = .64). This study provides two relevant observations: first, further evidence in support of Gow et al. suggesting the use of post-session commentaries; second, the use of events, or in this case, event windows to mark experiential shifts. Additionally, McCall et al. also showed that there was coherence between the physiological signals captured at the time of experience and their retrospective reports and appraisals of the experience. This suggests that participants, particularly those with an elevated awareness of their physiological state such as their heart bate, are capable of accurately remembering the emotional and physiological elements of virtual experiences.

Nourbakhsh used EDA and pupillometry to research cognitive load during arithmetic calculation (Nourbakhsh et al., 2013). Cognitive load was inferred from self reported difficulty, and several features for calculated from physiological signals: accumulative GSR, GSR frequency power spectrum, total blinks, and blink rate. The novelty of this work was the application of machine learning classifiers to cognitive load levels, using support vector machines and naive Bayes classifiers. Similarly, Nourbakhsh et al. also applied classification techniques to skin conductance data for a separate series of arithmetic experiments (Nourbakhsh et al., 2012). Based on the existing body of research indicating the use of physiology to measure cognitive load, it is understandable that these studies overlooked the validation of their data as that of cognitive load. However, basing classification of cognitive load entirely on self report with no data on participant performance is a severe limitation of both studies. Moreover, while there is arguable utility in the use of classification to categorise data into cognitive load levels, no new information of cognitive load was illuminated upon. However, these studies do provide value in that they present clear uses of classification algorithms to organise multidimensional time series signals, even in conjunction between different methodologies. Particularly that in the 2017 study, performance of classification based on just GSR features were similar between different experiments. The work by Nourbakhsh et al. has shown that both EDA/GSR and blinks are capable of indexing mental workload levels. Furthermore, the joint approach taken here has demonstrated that the combination of these two modalities together may potentially enable machine learning classification approaches, though the scope of success is practically limited to 2-class classifications.

More recent work by Jercic et al (Jerčič et al., 2017) used a novel technique of joint ECG and eye tracking measurements to investigate cognitive load in the auction game— a serious economic game. Jercic et al. designed an experiment within which heart rate and cognitive load were effectively coupled by the use of a biofeedback mechanism, whereby the heart rate determined the difficulty of the price estimation task. In doing so, a direct estimation of the expected cognitive load of the task was available for any given pupil diameter measurement. Using this assumption, the authors revealed a non-linear relationship between heart rate and pupil diameter, and therefore cognitive load and pupil diameter. They found that pupil diameter would increase with cognitive load until a threshold was reached where pupil diameter would decline, showing that pupil diameter would only be indicative of arousal within a certain range of cognitive load.

2.3 Conclusion

From the research reviewed in this chapter, there is sufficient evidence to suggest that player experience is tied to a set of psychological phenomena. Furthermore, these phenomena are potentially broader than any pre-existing psychological model of cognition like flow or selective attention. Immersion in particular also appears to be distinctly different from other experiential models, such as presence or engagement. Based on the body of experimental research using psychometrics, it has been demonstrated that this experiential process can be measured, at least approximately, by the use of psychometric questionnaires like the IEQ. Combined with a rich corpus of work on the successes that physiological signals have served other fields in cognitive research, the argument is put forth here that psychophysiological signals are potentially the most appropriate step forward for immersion research. This potential can be expressed in the ability to measure immersion granularly over time, and non-invasively, without a need to interrupt participants playing a game.

With this in mind, a most immediate priority would be to gather experimental data and research mapping physiological signals to the current psychometric model of immersion, building on the work of Cutting et al.

Chapter 3

Investigating the relationship between pupillometry, eye-tracking, and the IEQ

3.1 Introduction

Within academic games research, approaches to measuring player experience in video games have predominantly been with psychometric scales, as presented in the literature review. Among those studies that did adopt psychophysiological signals, it is not entirely clear what methodological challenges were encountered, or what the practical limitations of the data and analyses were from the perspective of building a reliable scale of immersion using a psychophysiological signal. Therefore, in order to explore the practical possibilities of measuring player immersion with physiological signals, an initial exploratory study was needed.

In order to conduct such a study, several critical decisions surrounding the design and logistics of an experiment required clarification. Namely, which signal modality to use as a measurement, what psychometric scale could best capture player immersion in the context of an experiment using psychophysiological signals, and how these two sets of measurements can be related to one another. In this experiment, the objective was to attempt to correlate a physiological signal in the form of pupillometry and electrodermal activity, with the psychometric measurement of player immersion.

3.2 The Psychometric Scale

The basis for much of the early work in developing engagement and immersion scales have predominantly been derived from the positive psychology theory of Flow (Csikszentmihalyi, 1990), which was adopted to measuring flow in games (Chen, 2007) resulting in works such as GameFlow (Sweetser et al.; Sweetser & Wyeth, 2005). Gameflow specifically has been published and since applied in design oriented research by its authors (Sweetser et al., 2012). However, as previously discussed, flow is likely to be too restrictive a definition to fully capture the experience of immersion that transpires when somebody plays a game. Therefore, more generalisable scales were considered to be used as an alternative.

The Game Engagement Questionnaire (GEQ) was developed with a focus on studying the psychological consequences of playing games (Brockmyer et al., 2009). On initial consideration, engagement is a more appropriately broad definition of player experience and therefore, the GEQ was a suitable candidate scale to be used in an exploratory study. However, the GEQ was developed with little emphasis on what conventional experiences of consumers of video games might have been. Instead, the scale was designed to be more attenuated with items intended to overlap with a psychometric assessment of aggression. Ideally, a psychometric scale that focused solely on capturing the essence of player experience would be preferable.

A few questionnaires with a greater focus on the solely games aspect of engagement can instead be found in work by Ryan, Rigby, & Przybylski (Przybylski et al., 2010; Ryan et al., 2006) which has explored the use of multi-factor questionnaires to probe at several aspects of game engagement, motivation, and immersion. Rather than focusing on the potential pathologies of game play, these works attempted to explicitly measure interactions with games, for example by capturing the mastery of controls that players felt they achieved. Here, self determination theory was used to construct a motivation driven model of engagement by incorporating The Player Experience of Need Satisfaction (PENS) scale, combined with scales for mood, self esteem, and future desires. The culmination of these make for a fairly comprehensive and compelling option for use in this experiment.

The last option considered and ultimately the questionnaire chosen for this experiment was the Immersion Experience Questionnaire (IEQ) (Jennett et al., 2008). A scale developed to measure player experience, the IEQ was designed to measure Brown & Cairns' definition of immersion (Brown & Cairns, 2004). As a measurement instrument, the IEQ assumes a multi-factor structure to be most representative of immersion, with the totality of these components forming the more holistic concept of immersion.

There are considerable overlaps with the scales developed by Ryan et al., including items to capture players' sense of control, presence, competence, and desire to play further. Therefore, the decision to use the IEQ was made entirely on pragmatic grounds- the scale developed by Ryan et al., was larger and therefore would involve longer experiments in additional to an already lengthy process of setting up physiological signal apparatus. In addition to this, choosing to use the IEQ also enabled an additional decision for this experiment- the experimental procedure and design itself. By using the IEQ, it would be possible to replicate previous work that used the IEQ to measure participants' immersion, meaning that there would exist a corpus of experiments in which immersion had been measured that this study could exploit by building a replication experiment.

3.3 Choosing Appropriate Signals

Having chosen to use the IEQ as the measure of player immersion, the next choice to be made was which psychophysiological signal would be used as a correlate for immersion. With such a broad set of options available in the literature, a set of priority criteria was first established in order to determine the most appropriate modality. First, the modality should have a previously reported relationship with cognition, or more ideally, player experience. Second, the modality should ideally involve minimal setup of apparatus for each participant. Third, the apparatus should cause minimal impedance to the participants' ability to engage with the game. Fourth and finally, modalities with richer data in higher dimensionality would be more desirable. Based on these parameters, eye tracking was chosen as the primary modality of interest, with electrodermal activity chosen as an additional modality to supplement eye tracking.

Eye tracking has a considerable presence in the history of both psychological and HCI literature, meaning that the first criteria for an established record was fulfilled. In addition to this, eye tracking has also been used to explore player immersion in the PhD work by Cutting (Cutting, 2018), meaning that eye tracking could be considered to have met the ideal standard of the first criterion. Furthermore, eye tracking hardware is jointly capable of capturing not just typical eye tracking metrics such as blinks, fixations, and saccadic movements, but also enables the recording and analysis of pupillometry and pupil diameter. All of these methods have prior literature that have indicated potential relationships with cognitive processes. Similarly, early and seminal work by Beatty, and subsequent work by Palinko in the HCI domain have both demonstrated a statistical relationship between pupil dilation and cognitive load (Beatty, 1976, 1982; Palinko & Kun, 2012; Palinko et al., 2010). Therefore, the fourth criterion for rich and high dimensional data was also met. Eye tracking apparatus is also less invasive than other modalities considered- for instance, EEG often requires the placement of multiple electrodes or worse, a gel conductive cap which would require participants to also wash their hair following an experimental session. On the other hand, eye tracking involves minimal setup beyond the calibration of participants' gaze location with that of the recording, fulfilling the second criterion. Finally, eye trackers do not interfere with the interfaces with which participants play a game, unlike something like an fMRI which would significantly restrict the movements of the participant in order to ensure data fidelity.

Electrodermal activity was additionally chosen as a modality due to the simplicity of analysing the signal- unlike eye tracking, EDA would record only one timeseries of participants' skin conductance, meaning the scope of analysis would not expand unreasonably by the addition of this modality. Furthermore, EDA had been previously correlated to player experience in the games research literature (Drachen et al., 2010; Kivikangas et al., 2011).

These two psychophysiological signals were considered to be capable of capturing enough data for there to be an exploration of potential correlations with immersion as measured by the IEQ, while being minimally invasive to participants' playing experiences, and easy enough to set up for multiple experiments.

3.4 Background

3.4.1 The Placebo Effect

The placebo effect specifically in the context of games experience occurs where a player's prior beliefs before playing a game can influence their experience of the game itself. A version of this effect was produced in experiments carried out by Denisova (Denisova & Cairns, 2015b), where participants were deceived to believe that the game they were about to play contained an advanced, adaptive artificial intelligence (AI). Participants in these experiments were informed that this AI allowed the game to adjust its difficulty to their skill level. This effect was demonstrated with the game 'Don't Starve', by describing the the game's world generation as being capable of adapting to the participant as they played. When immersion was measured through the use of the IEQ, statistically significant differences in immersion scores were found between the players who were deceived with the AI and players that were not. After the initial observation of this effect, a follow-up experiment (Denisova, 2016) once again produced this effect under a 2x3 factorial design to produce additional of the causal relationship between the immersion score differences and the deception of an artificial adaptive AI.

3.4.2 Experimental Manipulation

Given multiple previous instances of this placebo effect and the relative simplicity of the experimental manipulation of deception, Denisova's experiment using Don't Starve was selected as the basis for a replication experiment in this chapter.

Here, the simplest two condition version of Denisova's experiment is adopted, where one group is falsely informed of the presence of an adaptive AI in the game, while a control group plays the game as normal. The core manipulation therefore, is the presence or absence of a participant's belief in the presence of an adaptive AI.

The complexity of this study is instead introduced by the myriad of additional measurement systems in addition to the Immersive Experience Questionnaire, and the subsequent series of exploratory post-hoc analyses conducted in section 3.9.

3.4.3 Stimulus Game

Typically pupillometry studies use a task like doing numerical calculations to induce cognitive load such as that initially outlined by Beatty et al (Beatty, 1982). Pupil diameters are then correlated to the cognitive load, i.e. the difficulty, of the task. Similarly, in addition to following Beatty et al.'s original paradigm, Cutting has also investigated pupil dilation as the independent variable corresponding to the measurement of immersion. In the case of Cutting's experiment, the game chosen was "Two Dots", a self paced puzzle game. Like traditional psychology experimental designs, the format of Two

Dots allowed for the clear identification of a pre and post response window (Cutting, 2018).

In his experiment, the stimulus game used was the commercial game *Don't Starve*. This was a game with some more complexity in the data and forms of player interaction than that of *Two Dots*. Several reasons are discussed as to why such a complex game was chosen for this experiment, which are presented in greater detail in section 3.6.4. The key reasoning however, was that the more expansive nature of the game permitted for a broader set of opportunities for exploratory analyses. Also important was the fact that unlike *Two Dots*, participants playing *Don't Starve* were interacting with a game that was not self paced. Because the game forces players to abide by its own in-game schedule and cycle, and because there are real-time elements that demand the player's attention and decisions, the game produced a more natural environment for a event related paradigm.

As a result of choosing such a game, there were no clear response windows that could readily identified for a task-evoked paradigm, and event-response windows had to be carefully selected given the plethora of options for when an event is to be recorded. The definition of these windows and their subsequent analyses are provided in section 3.9.

In addition to the independent variable of adaptive AI deception as the experimental manipulation, the previously mentioned in-game clock of the game was improtant to the design of this experiment. The time of day within *Don't Starve* is a more temporally compressed version of time of day in our real world. The game presented players with different opportunities and requirements during its three in-game times of day: day, dusk, and night. This periodic cycle was to be treated as a within-samples condition. Every participant experienced the different times of days that the game had to offer, with a participant experiencing two full cycles on average (with variations caused by unexpected deaths).

In choosing such a game, there were consequences that affected the subsequent analyses of the data collected in this exploratory study. The decision to use such a complex game and specifically the use of the in-game time of day as an experimental manipulation resulted in considerable complications, the implications and limitations of which are explored in the discussion of this chapter.

3.5 Research Questions & Hypotheses

3.5.1 Research Question

The overarching research question of this exploratory study was "Is there a relationship between physiological measurements, in the form of eye tracking, pupillometry, and electrodermal activity, and immersion as measured by the IEQ?".

To answer this question, a collection of questions specific to each component of the experiment were explored.

Confirming The Placebo Effect

First, to confirm the previously recorded placebo effect, this study aimed to answer "Does the belief in the presence of an adaptive artificial intelligence produce a difference in immersion scores?". Here, the planned hypotheses were that:

1. There would be a significant difference in belief in the presence of an Adaptive AI (AAI) scores between the two groups.
2. There would be a significant difference in immersion scores between the two groups.

Exploring the relationship between pupillometry and immersion

Second, to explore the relationship between pupillometry and immersion, a simple hypothesis was stated:

3. There would be a significant correlation between pupil diameter and immersion as measured by the IEQ.

Tied to the presence of this relationship between immersion and pupil diameter, an exploratory hypothesis focusing on the difference between immersion states was also made:

4. There would be a significant difference in pupil diameter between the placebo (treatment) and the control group.

Finally, it is noted that the stimulus game had an inherent design element that produced different luminosity emissions between different in-game times of day. This day and night cycle was a core element of the game, driving players to play in different manners across the duration of the experiment. Here, hypotheses are stated based on this system of the in-game day and night cycle, and further details of the function of the times of day system in the stimulus game are described in section 3.6.4. Each time of day was treated as conditions in the game. This was done primarily in order to verify that different values of luminosity were indeed producing different pupil diameters:

5. There would be a significant difference in pupil diameter between different in-game time of day segments.

As a consequence of luminosity variability between different time of day segments, there was also an opportunity for further unplanned post-hoc analysis of within-segment differences between participant groups of interest, as later detailed in section 3.9.

Exploring the relationship between eye tracking measures and immersion

Similar to the previous subsection detailing pupillometric differences between experimental conditions, and different time of day segments:

6. There would be a significant difference in fixation rate and duration between different experimental conditions.

7. There would be a significant difference in fixation rate and duration between different in-game time of day segments.

The justification for additionally exploring fixations across time of day segments can be found in the fact that different in-game time of day segments demanded different forms and degrees of interactivity from participants. This therefore, had implications on the experience of participants and more acutely, different degrees of reactivity depending on the context of the time of day.

Exploring the relationship between electrodermal activity and immersion

Finally, in the same vein, as with pupillometry and eye tracking, hypotheses on the relationship between EDA and experimental groups were made:

8. There would be a significant difference in electrodermal activity between each group.

9. There would be a significant difference in EDA between different times of day.

Here, the plan was to compensate for the potential confound in luminosity variability between times of day on pupillometry and eye tracking, by using a modality not dependent upon physiological responses to visual elements.

3.6 Method

3.6.1 Experimental Design

The experiment was designed as an exploratory study with the aim to investigate the relationship between psychophysiological measures and immersion as measured by the IEQ. As a basis for the experiment, Denisova's Placebo Effect experiment was chosen as a replication source (Denisova & Cairns, 2015b).

In this experiment the independent variable was the same as Denisova's studies: the deception (or lack thereof) of the belief that there would be an adaptive AI present in the game. The condition in which participants were not informed with the deceptive information that an adaptive AI would exist, will henceforth be referred to as the Placebo condition. The treatment group therefore, were participants who played after having been falsely informed (and therefore presumably deceived) that the game would have an adaptive AI agent adjusting the difficulty of the session based on their skills. Participants were randomly assigned to one of each group at the start of the experiment.

3.6.2 Participants

An a-priori power analysis was difficult to conduct for this experiment for several reasons. First, the effect sizes found in Denisova's study were not entirely applicable to this experiment due to a different set of independent variables being captured in the form of the physiological signals. Second, if the conventional approach of conducting an a-priori sensitivity power analysis was taken, the resulting sample sizes required would be far larger than what would be financially or logistically possible for the scope of the experiment within the wider context of this PhD work. Therefore, a decision was made on pragmatic and practical grounds that data would be collected to match the time and budget allocated to this study of the PhD, which was approximately between 40 and 70 participants depending on the time required to recruit participants.

Ultimately, a total of 41 participants were recruited for this study, through a public recruitment portal available through the University of York website. 26 female and 15 male students participated in the study. The youngest participant was 18 years old, and the oldest 36 (Age mean = 20.92, std. 3.53). Participants were students at the university from an varied assortment of departments and comprised of both undergraduate and postgraduate students.

A demographics survey was also carried out in order to filter participants who had previous exposure to the game used in the experiment, in addition to general gaming behaviours. Regarding prior knowledge of the game, 12 participants had heard of Don't Starve before, and 29 had not. Of those who had, 6 had also previously played the game. Regarding prior knowledge of adaptive AI, 13 participants had also heard of adaptive AI prior to the experiment. Regarding familiarity with games, the majority of participants (39) had played games in general before, with 12 participants stating they played several times a week. 20 participants reported playing at least an hour on an average session, 15 participants reported playing between 15 minutes to an hour, and 4 participants played fewer than 15 minutes a session.

During data collection, 1 participant's data was partially removed from the sample due to hardware failure during acquisition, and 2 participant trials resulted in total data loss due to a software driver malfunction. For the 41 participants from whom psychometric data was collected, the groups were approximately equal (21 in Placebo-AI, 20 in Control). With respects to the physiological data, a total sample of 38 participants was collected, with some some caveats to this number provided in the analyses section below.

3.6.3 Materials & Equipment

Participants played the game on a desktop PC with sound playback through Audio-Technica ATH-M30X headphones at a volume which players were free to change to their comfort or accessibility requirements. Participants were also permitted to play the game with the peripheral set of their choice: either a mouse, or a mouse and keyboard. The variability of these hardware during the experiment are discussed below. The experiment itself was conducted in a lab designed to replicate the environment of a home dining room, with decorations and setting in place to minimise damage to ecological validity. The luminosity of the room was also controlled across every trial of the experiment by the eliminating all sources of light other than that emitted from the screen, which was kept standardised.

The Pupil-Labs "Core" open source eye tracking hardware was used for this experiment. The hardware comprised of a single front facing "world" camera which captures footage of the user perspective, and a pair of RGBD cameras to track the eyes. These sensors were mounted on a plastic frame, worn as if they were a pair of glasses sans the lenses. Pupil data was collected from both eyes using Pupil-Labs' 3d model, at 120Hz using the 3D model method. Front perspective footage was also captured in order to calibrate gaze data at 1280x720, 15Hz.

The data itself was captured under pupil-capture software version 1.8-26 was used for pupillometry data acquisition. Data was processed with pupil-player version 1.8-26, a software designed to replay captured data including the video recording of the participants' pupils, and front-facing camera. Use of the pupil-player software also included instances where post-hoc reprocessing of pupil tracking was required for certain participants for whom pupil tracking was recorded sub-optimally during the experiment.

EDA data was collected using a Mindmedia Nexus 10 MK-II encoder recording at 32Hz. However, when the data was exported using the provided software, the resolution of encoding was unavoidably down-sampled to 1Hz.

Pre-processing was conducted with the Pupil-Labs pre-processing software (Kassner et al., 2014), and *pandas* (version 1.3.5). Analyses were conducted using the Python programming language (version 3.7), with statistical tests conducted through the *pingouin* library (Vallat, 2018), and graphs plotted using *matplotlib* (version 3.4.3) and *seaborn* (version 0.11.2).

3.6.4 Questionnaires

Signals from psychophyisology were compared to psychometric data acquired using the IEQ (Jennett et al., 2008), which is comprised of 31 items on a 5 point likert scale. The items are divided across 5 subscales of immersion, and the survey is wholly designed to capture immersion in its different forms. The full IEQ questionnaire can be found in appendix A.4. The IEQ was used to measure the immersive state of the participants, which was a primary dependent variable in the experimental design.

In addition to the IEQ, 6 questions targeting the participants' perception of any adaptive AI (AAI) in the game were also added in order to determine the effect of the manipulation. The questions for this additional set were taken from prior work by Denisova (Denisova & Cairns, 2015b). It is noteworthy that these questions should not be treated as a robust psychometric scale, but rather an ad-hoc series of questions that were developed and initially tested in the original experiment run by Denisova. There are therefore, some differences to how these 6 questions are used in this experiment as compared to Denisova's, primarily in the nature of how the items are examined. Here, the interest in the AAI questions and subsequent score are solely to determine if the deception successfully took place. Denisova's original research was interested in the intricacies of the nature of belief in adaptive AI, and thus in the original experiment, each item was tested individually. However, in the exploratory context of this study, such an approach would be increasing the number of statistical tests for little gain in relevant information, and as such an aggregate approach of testing a mean score was used instead.

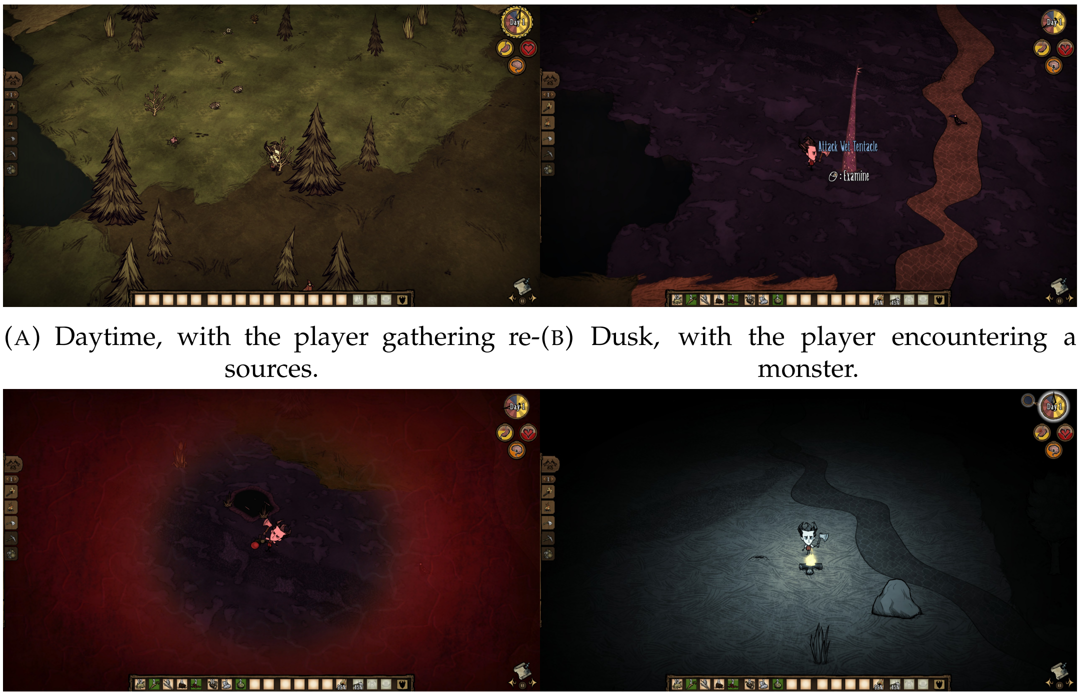
Additional questions were also included to record participant demographics data. These items were presented to participants digitally through the Qualtrics data collection system following the playing session of the experiment.

The full AAI survey can be found in appendix A.5. The perception of AAI questionnaire was used to measure the participants' individual perceptions of whether or not an AAI was present during their playing session. This perception score was to evaluate the successful deception of the independent variable in the experimental design.

Game

As this was a replication of (Denisova & Cairns, 2015b) game used for this experiment was Don't Starve. In order to ensure that hardware began recording at precise start and stop intervals for trials, a simple local server was developed to act as an intermediate communication point between the Pupil software and Don't Starve. The Nexus system, however, did not have any networking utilities or software APIs available to achieve similar functionality for the EDA data. Thus, EDA recordings were was initiated manually each trial by the investigator. This was considered to be an acceptable compromise to data quality under the justification that electrodermal data was often analysed at a lower signal frequency than eye tracking, and that the software provided would have restricted the dataset to a 1 second precision during data export anyway.

Event recordings for Don't Starve were chosen sparsely for the experiment, as the primary focus was to first inspect broader shifts in the play experience across the in-game times of day. In Don't Starve, players were expected to behave differently depending on the time of day within the game. Daytime and dusk were considered to be safe periods during which players could explore and gather materials, while night was dark and antagonistic which encouraged more defensive and passive game play. During night time, players were required to stay near an in-game light source at all times or perish, and a significant objective during dusk and daytime was to prepare resources in order to survive this dangerous period. Additionally, event triggers were also recorded when players were gathering through an item pickup event, as well as when players engaged in combat through a player attacked event. A game start event was also used in order to both demarcate the beginning of the trial session, as well as in order to determine a new run following a player death. These event triggers were programmed by modifying existing Don't Starve code exposed in a Lua script directory, and communicated to the physiological signal apparatus through the aforementioned server written for this experiment. Screenshots presenting examples from the game are shown in figure 3.1.



(A) Daytime, with the player gathering resources.

(B) Dusk, with the player encountering a monster.

(C) Combat example where the player is attacked.

(D) Night-time example, where the player is safe at a campfire.

FIGURE 3.1: Don't Starve Screenshots, demonstrative examples captured outside of participant data collection.

3.6.5 Procedure

Participants began the session with the introduction, information, and consent protocol (appendix item A.3 to detail the procedure and objective of the experiment. At this point, the information document (appendix item A.1) would explain to participants in the placebo-AI (i.e. treatment) condition that the game would involve an advanced adaptive AI that would alter the gameplay to their collective actions and experience. In addition to this, both samples were informed of the game and its details, as well as a vague objective for the experiment. During this part of the experiment, participants were also able to ask questions regarding the game or general experimental protocol.

Following this, participants would first put on the Pupil Labs eye trackers, followed by the headphones such that the headphones were worn over the plastic frame of the eye trackers. During this setup, the binocular pupil cameras were adjusted to align with participants' eyes, and calibration was carried out to establish pupil gaze. Participants were then equipped with the two Mindmedia GSR (EDA) electrodes to their index and ring fingers on their left hands. This was originally intended to be on participants' non-dominant/non-mouse hands, but all left handed participants opted to use the mouse with their right hands, resulting in left-handed participants with the electrodes on their left/keyboard hands, which was not considered to affect the results of the EDA recordings. EDA signals were tested with a simple respiratory response (inhale-exhale with a still body), and exaggerated finger motions were conducted to inspect noise induced by movement.

Once all apparatus were confirmed to be working, participants undertook a practice session during which recorded data was inspected in-vivo to ensure signal stability and general data reliability. Following the practice session, if any recording issues arose or noise was deemed too high, adjustments were made to the devices before participants continued on with the main trial. In the majority of cases for the eye-tracking modality, issues with data quality were ironed out at this stage.

The practice session consisted of a 8 minute session, with the duration chosen due to the length of a full day-night cycle in the game. During the practice session, participant were permitted to play as they wished in order to familiarise themselves with the game. While playing, participants were also provided with a standardised readout of instructions teaching them how to control the character and the game UI elements. During the whole practice session, participants were also able to have any further questions they had, answered.

For the main trial, participants played the same as they would have during practice with the exception that all communication was ceased with the experimental investigator. A main trial lasted 20 minutes and upon reset states such as player death, players had been instructed earlier to continue playing as normal rather than interact with the investigator. Upon completion of the trial, participants removed the recording apparatus and completed the questionnaire through a Qualtrics data collection page.

Participants were informed that the experimental investigator would be present to monitor hardware functionality as they played the game. During play, qualitative notes were also taken by the investigator of gameplay observations and signal behaviour. No structure or cross validation process was planned, and notes were taken spontaneously as experiments were conducted. In addition to these observational notes, several critical types of in-game events were recorded, which included instances when players were acquiring items, when players were attacked in combat, and when players died. These qualitative and event based data were recorded for the purposes of informing the later exploratory analyses conducted in this study.

At the end of this protocol, participants were then debrief and informed of the deceptive nature of the experiment. The debrief document can be found in appendix item A.2.

3.7 Data Pre-processing

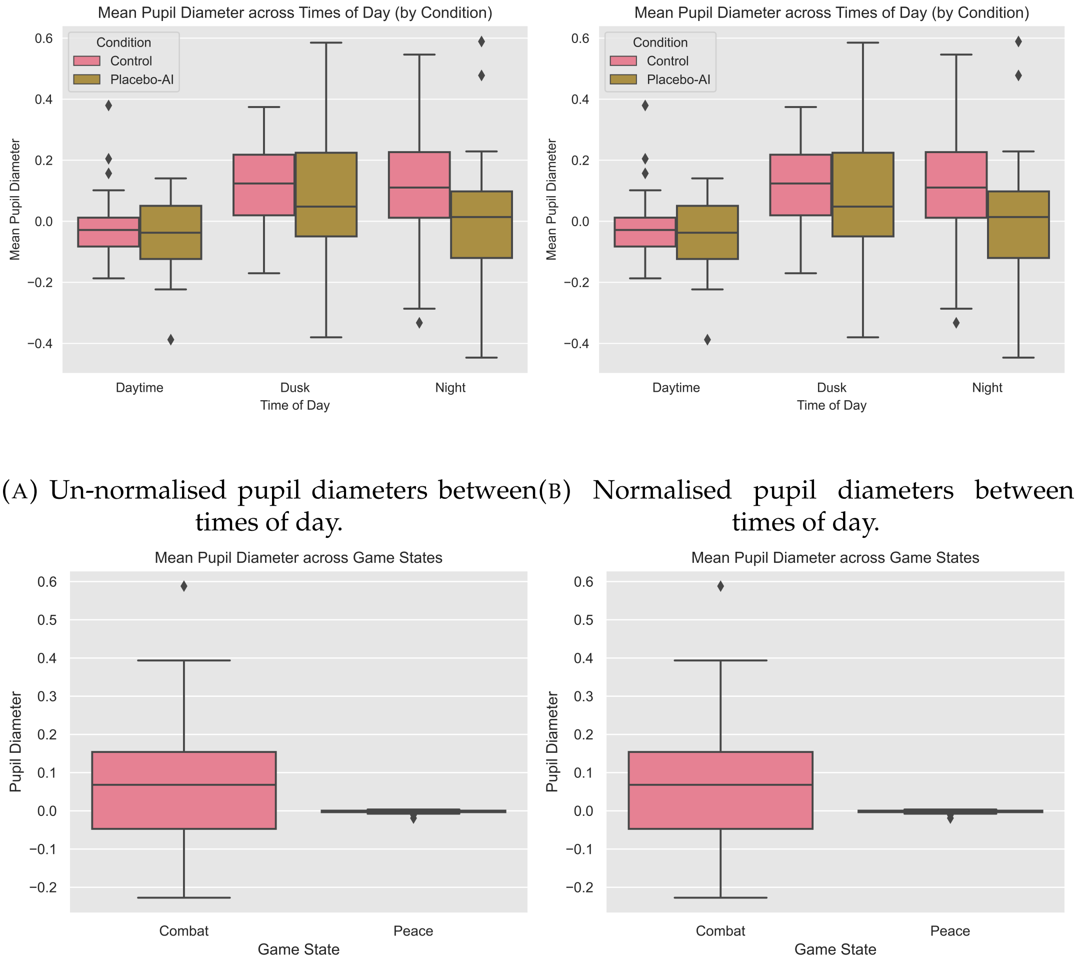
3.7.1 Pupillometry Processing

Pupillometry recordings undergo a processing pipeline in-vivo at the point of data capture, with methods as described in (Kassner et al., 2014). Blinks were also autonomously filtered during data acquisition using Pupil-Labs' detection algorithm (Kassner et al., 2014) during data processing.

During the data collection of this experiment, an issue arose in that an error with the blink tracking algorithm resulted in some discarded data, where the signal was not deemed reliable due to an unexplained and unacceptable amount of high variances in the data.

The first step of the pre-processing pipeline involved the calculation of a binocular value, rather than treating the data with each eye independently. The left and right pupil measurements were combined by averaging the diameter and confidence values for pairs of frames at each recording interval. In instances where only one pupil measurement existed, the pair was omitted. The resulting new binocular dataset was then used as the base for noise processing at the next stage.

Then, it was necessary to account for individual variations in pupil size among participants. In order to standardise pupil diameter measurements across all participants, a common approach is to transform pupil diameter into a relative and normalised measurement. Initially, an approach was taken where pupil diameter change was recorded relative to the previously measured pupil diameter (with a pupil diameter change of 0 applied to the very first measurement for each individual participant). However, the meaning of the measurement changes quite substantially with this approach, and the subsequent analyses would then be analyses of difference in pupil diameter relative to the previous frame, rather than pupil diameter directly. The knock-on effect of this would be that in cases where pupil diameter is relatively static, the pupil diameter change values would be very small. In essence, this approach to normalisation would not make sense. A common approach taken in research such as that taken by Jainta Baccino (Jainta & Baccino, 2010) is to subtract the pupil diameter during a neutral, zero-stimulus measurement of the pupil. However, in the present study, due to an oversight, no baseline zero-stimulus measurement was acquired. An attempt to normalise the full data by dividing the pupil diameter of a participant by the mean diameter from their full pupillometry timeseries was considered, but was not taken due to the realisation that luminosity varied substantially between the times of day within the game. This meant that the overall average taken from the experiment would actually be an average weighted with respects to luminosity. This is due to the fact that the daytime period's luminosity, which is the brightest, was much longer than either dusk or night time periods and therefore would have had a greater number of samples within the data. Most importantly of all, no reasonable assumption can be made that the average pupil diameter over the course of the full experiment would be equivalent to a zero-stimulus pupil diameter. In fact, as one would expect, normalising by taking the average pupil diameter over the full course of the experiment results in very little change in the analyses of this study other than changing the relative values of the measurements, as presented in figure 3.2. In the end, although less than ideal, it was decided no normalisation would be performed, given that all comparisons of pupil diameter were planned to be pairwise within-subjects anyway.



(A) Un-normalised pupil diameters between times of day.

(B) Normalised pupil diameters between times of day.

(C) Un-normalised pupil diameters between combat states.

(D) Normalised pupil diameters between combat states.

FIGURE 3.2: Comparison of the distributions of pupil diameter between normalised and non-normalised measurements. Observe that other than different relative values, the distributions are expectantly otherwise identical.

For noise processing, pupil data was filtered by first removing values for which confidence estimations were lower than .60. These confidence values specifically represented measurement confidence, i.e. how confident the pupil software was that the measurement captured from the eye would be correct. This confidence metric was calculated for every measurement taken in a recording, i.e. every frame captured by the the eye-tracker camera. A confidence cut-off threshold of .60 was chosen as per guidance from pupil labs documentation. Because data were removed during the confidence filtering process, there were now gaps in the timeseries. To correct this, a simple linear interpolation was applied to the data.

After filtering values below the confidence threshold, there was still considerable noise observed in the data through visual inspection. The source of this noise was determined to have been at least in part due to the tracking algorithm used, where as 3D eye modelling approach was considered to have caused greater signal noise than a simple 2D approach.

The next step in this processing chain was to apply a linear detrending correction to the data. The motivation for this was to account for gradual measurement confounds typically expected when using psychophysiological measurements. One critical example of such a confound can be found in the choice to use Pupil Labs' 2D model rather than the 3D model, whereby only the 3D modelling approach accounted for slippage. This choice resulted in the loss of in-vivo measurement re-calibrations accounting for slippage where the eye tracking glasses frame physically shifts from its original position relative to the participants' eyes as the experiment progresses. This meant that slippage had to be accounted for, resulting in a necessity to detrend during pre-processing. The specific detrending method applied was to simply subtract a linear fit of least squares of the timeseries from the timeseries itself. An visual example of the consequence of this detrending for one of the participants is provided in figure 3.3), and this would be principally representative for all participants. Implications for this detrending are discussed later in this chapter.

The next step of noise processing for the pupil diameter signal involved the application of a Butterworth bandpass filter. First, a spectral density graph was generated in order to determine the appropriate bandstop value (Figure 3.4). Based on the visual inspection of pupil diameter signals across participants, this value was determined to be 1Hz. An example of one such density graph is presented in figure 3.5), with examples of varying bandstop frequencies applied. The specific bandstop threshold applied here was largely inconsequential for the planned analyses, as a further aggregation would be taking place at the analysis stage. The primary purpose of this filter was therefore to remove any brief and sudden high-amplitude fluctuations in the data, as presented in the 4Hz example in figure 3.5.

3.7.2 EDA Pre-processing

The EDA data was captured by MindMedia software and hardware which preemptively applied its own pre-processing on the data. These pre-processing runs were not fully detailed in documentation, but it was estimated that a smoothing and denoising filter was applied during data capture. For this reason, as well as other issues with data reliability of the EDA signal, this modality of the study and all recorded data was considered to be potentially too unreliable to interpret due to a lack of clarity and confidence for any analytical results, null or otherwise.

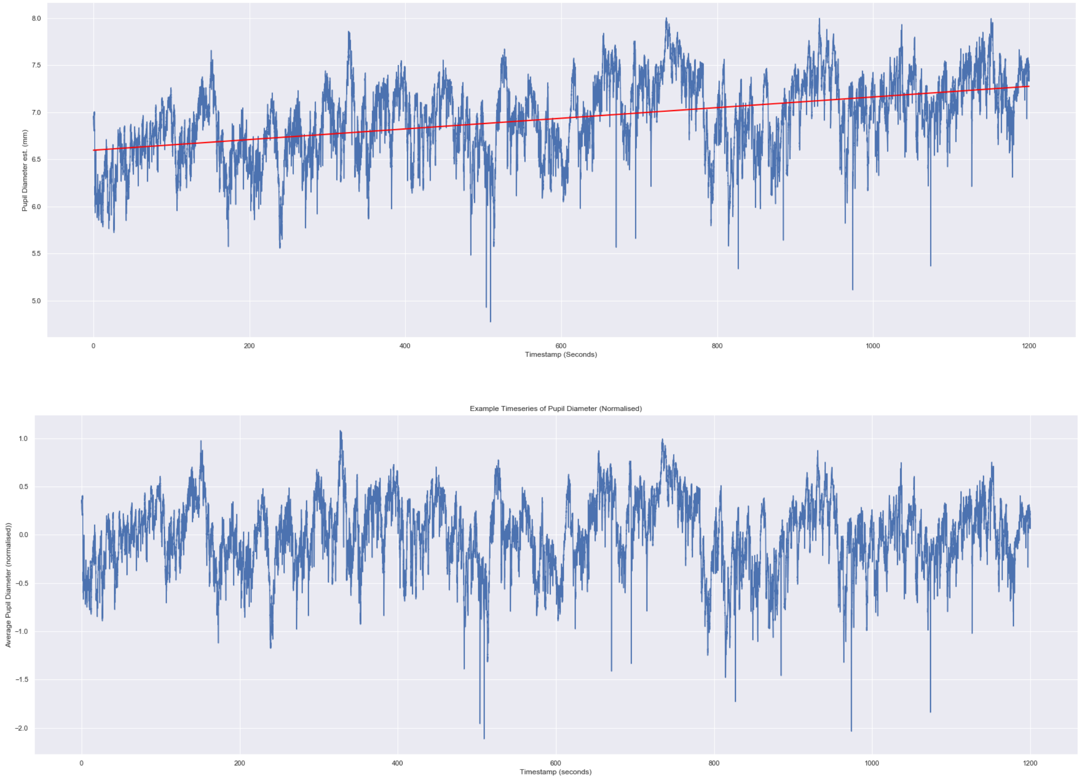


FIGURE 3.3: Timeseries of pupil diameter across the experiment for an example participant before (above) and after (below) linear detrending.

First, all dropped and *NA* samples were removed. Then, an additional set of erroneous measurements were detected, whereby unreasonably large EDA measurements were detected when, presumably, the electrode straps loosened around the fingers or some other sampling error took place within the apparatus. For context, the average mean EDA was 5.47, but 19006 observations were measured above 200. Although at a closer glance such frames only comprised less than 0.02% of the total number of observations across the full sample, the difficulty in tracing the root cause of these fluctuations as well as the sporadic nature in which they appeared within the data suggested that there may have been deeper issues with the data captured.

Given time constraints during the remainder of the allotted time for analysis, this part of the study was deprioritised and ultimately abandoned due to the reasons above.

3.7.3 IEQ and AAI Pre-processing

The pre-processing of the survey data was comparably straightforward, requiring only the calculation of the normalised scores for immersion (IEQ), and normlised scores for belief in adaptive AI (AAI).

IEQ scores for each participant were computed by taking the mean of responses to all 31 items of the IEQ questionnaire.

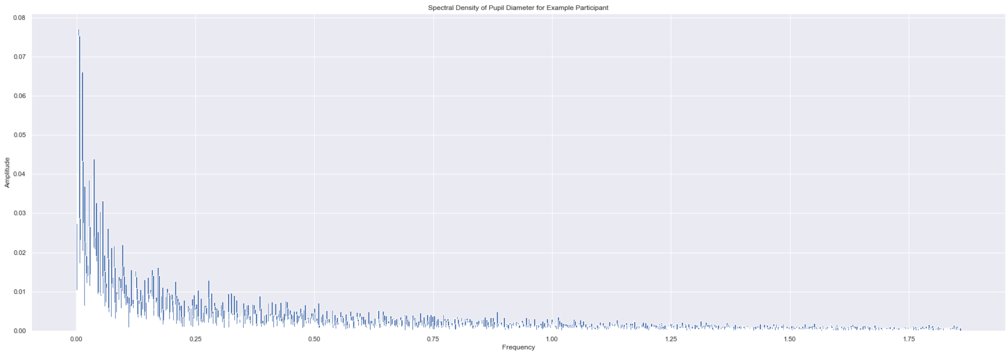


FIGURE 3.4: Example power spectral density graph of pupil diameter for a single participant. Based on a collection of PSD graphs including this sample, a bandstop value of 1Hz was used for processing noise from the data.

The AAI score was computed by taking the mean of all AAI questions taken from the study by Denisova. Note that this is a slightly different approach to how how belief in AI was explored in the original study by Denisova, where each individual item was explored. Here, the primary function is only to observe whether there were differences in respondents' overall belief (or lack of) in the presence of the adaptive AI.

3.8 Analyses & Results

3.8.1 Suitability of Data for Analyses

From this point forward, and for the remainder of the thesis, statistical tests reported have been chosen based on their suitability of the data for such tests. However, there are some implications to consider in terms of priority and correctness insofar as how suitability is determined.

First, when applying parametric tests, the assumption of normality is often considered to be a necessary requirement. However, this assumption comes with several caveats. First, is that commonly used normality tests such as the Shapiro-Wilks will test for the null hypothesis that samples were obtained from a normal distribution, which is not the same as testing against a null hypothesis that the sample does not come from a normal distribution, which can be argued to be the incorrectly assumed intent of a normality test, in the context of judging suitability for analysis. This can be demonstrated by observing that rejecting the null hypothesis is not equivalent to accepting the alternate, i.e. to reject the null of a Shapiro-Wilks test is to reject the null hypothesis that the data came from a normal distribution. This means that the absence of rejecting the null still does not concretely determine normality (Cairns, 2019), and that even in the known absence of normality (through such means such as simulation experiments), there are negligible results on the power of the test (Tsagris & Pandis, 2021), (Sawilowsky & Blair, 1992).

More pertinent to the objective of determining the suitability of data to a test is the presence of homogeneity of variance (or lackthereof) in the data.

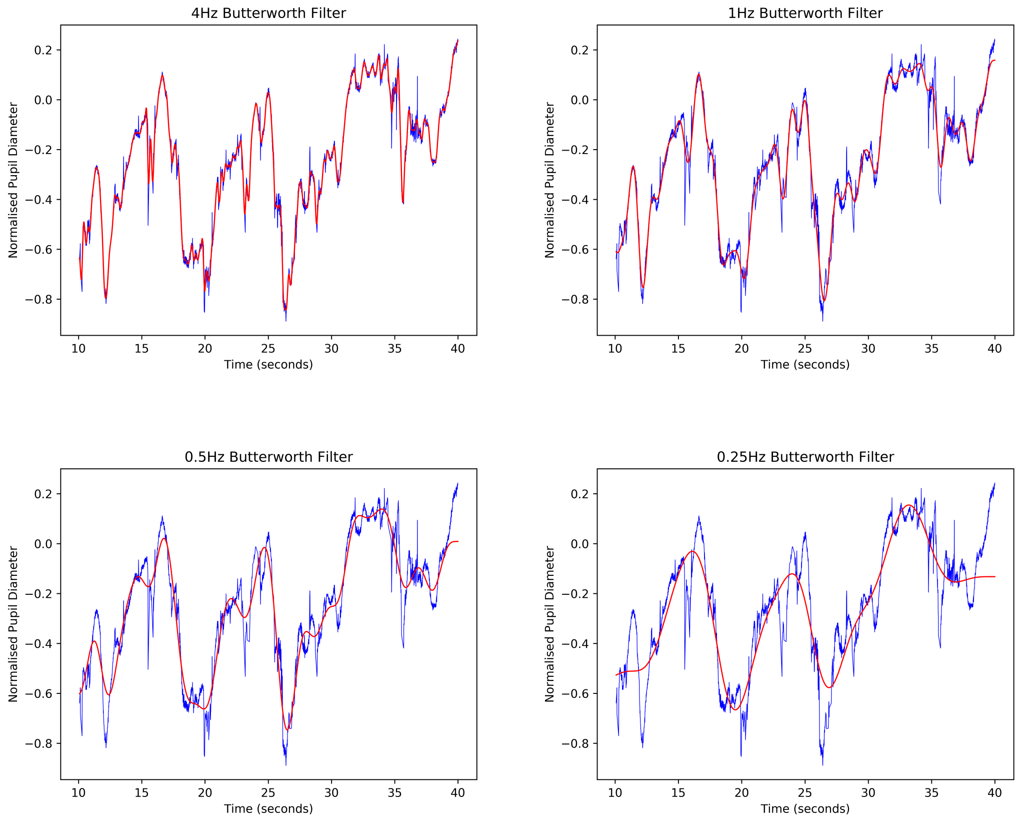


FIGURE 3.5: Illustration of various bandpass filters applied to an example 30 seconds of data. The objective here was to determine which filter would be most appropriate for the data. The 4Hz filter was not able to remove most of the high frequency spikes, while the 0.25Hz filter appears to smooth the timeseries too far. Between 0.5Hz and 1Hz, the differences are less discernible and a case could be made for either filter. In the end, 1Hz was qualitatively considered to likely be most similar to the 'true' data, while smoothing out the high frequency spikes.

Here, the Type I error rate can be substantially influenced by unequal variances (Ramsey, 1980), and although the suitability of a homoscedasticity test will vary based on the normality of the underlying distribution (Erceg-Hurn & Mirosevich, 2008), there is at least more robustness in tests of unequal variances than in tests of normality (Cairns, 2019).

Based on these considerations of the importance of, and tests for, normality and equal variances, the following actions have been taken for analyses in the remainder of this thesis: First, normality tests are not used to consider whether data is suitable for a parametric test or not. Second, the use of tests of homogeneity of variance is instead treated as a basis for determining the suitability for whether a testing protocol requires adjustment. Third, this adjustment is not necessarily towards non-parametric tests, as violations of homogeneity of variance also negatively influence the power of such tests (Zimmerman, 2000). Instead, more robust variants of a given statistical test is chosen, such as the Welch's t-test in place of a t-test (Zimmerman, 2004), or the Greenhouse-Geisser for repeated measures ANOVAs in the absence of unequal variances (Haverkamp & Beauducel, 2017).

For the purposes of testing for equality of variance, Levene's test for homoscedasticity was chosen over alternatives such as Bartlett's test. This decision was made based on the fact that Levene's test is more robust to non-normal data (Vorapongsathorn et al., 2004).

Psychometrics Data Suitability

For the test in section 3.8.2, the IEQ score data exhibited equal variances as determined by a Levene's test ($N = 38$, $W = 1.787$, $p = 0.19$).

For the test in section 3.8.2, the AAI score data exhibited equal variances as determined by a Levene's test ($N = 38$, $W = 1.44$, $p = 0.292$).

Pupil Diameter Data Suitability

For the analyses in section 3.8.3, the pupil diameter scores did not exhibit sphericity (i.e. equality of variances across all levels), following a Levene's test of sphericity ($N = 38$, $W = 0.842, p = 0.045$). Therefore a Greenhouse-Geisser correction was applied to the ANOVAs conducted, and Welch's t-tests were used to correct post-hoc analyses.

For the analyses in section 3.9.1, Levene's tests were again conducted which indicated unequal variances of mean pupil diameters between combat states ($W = 56.24$, $p < 0.001$). Therefore, t-tests in those analyses were conducted by using the Welch's t-test.

Fixation Rate Data Suitability

For the analyses of fixations, Levene's test of equal variances was conducted on the fixation rate between different times of day, and the fixation rate between different combat states. Equal variances in fixation rate were found for both times of day ($W = 0.191, p = 0.827$), and for combat state ($w = 0.351, p = 0.555$). Therefore, the analyses conducted in that section were made with no adjustments.

Electrodermal Activity Data Suitability

There were hypotheses tests planned to test differences in electrodermal activity across conditions and in particular, times of day, as discussed in section 3.5. However, as previously mentioned in section 3.7.2, the data for electrodermal activity was discarded due to several quality control issues. First, the sampling rate from the output of the recording apparatus was substantially lower than the measurement rate recorded on the device, with data down-sampled to 1Hz- a fidelity considered to be unacceptable for the purposes of this experiment, especially as the down-sampling technique used was not made transparent by the documentation of the device. Furthermore, this down-sampled data was initially explored and frequent instances of the signal dropping to 0 or missing measurements were found, leading to the belief that the data provided by this device was unreliable. For these reasons, the planned analyses for electrodermal activity unfortunately were not conducted.

3.8.2 Psychometrics

Placebo-AI vs Control

Based on (Denisova & Cairns, 2015b), the study aimed to replicate the placebo effect of expectation with an experiment using psychophysiological measurements. An independent two sample t-test revealed that there was no significant difference in the IEQ between the placebo-AI and control groups (T = 1.349, p = 0.187, $95\%CI = [-0.07, 0.36]$, $M_{control} = 3.749$, $M_{placebo-AI} = 3.604$, $SD_{control} = 0.269$, $SD_{placebo-AI} = 0.381$).

Effects of belief in advanced AI

Participants' belief in the presence of an advanced AI was statistically compared between those in the Placebo-AI group, and those in the control group, with an independent two sample t-test. No statistically significant difference was found in AAI scores between the two groups (T = -0.634, p = 0.531, $95\%CI = [-0.48, 0.25]$, $M_{control} = 3.167$, $M_{placebo-AI} = 3.281$, $SD_{control} = 0.441$, $SD_{placebo-AI} = 0.648$).

IEQ and AAI scores reveal a need for additional analysis

Based on these results, the possibility of an under-powered experiment was considered. Denisova has previously published finding moderate effect sizes under this experimental design for both the IEQ ($\eta^2_{partial} = 0.137$) and AAI scores ($\eta^2_{partial} = 0.154$), with a sample size of $n = 40$ participants. In comparison, the present study involved $n = 38$ participants for this analysis, which would make the samples relatively similar. Boxplots presenting the distribution of AAI and IEQ scores are presented in figure 3.6.

Therefore, if the possibility of an under-powered experiment was considered, some other underlying mechanisms would have been more likely to be the cause of reducing the statistical power, rather than the sample size or the underlying effect size of the experimental manipulation. The potential causes for this possibility are presented in the discussion of this chapter.

Further exploration of psychometric data

To contextualise the close (and non-significantly different) means observed in the AAI test scores, scatterplots of the AAI and Immersion scores were visualised for inspection (figure 3.7a). Additionally, a linear fit of the relationship between the AAI and Immersion scores was also overlaid on top of this visualisation in order to explore the relationship between the two dependent variables. From this visual inspection, it was observed that the gradient of

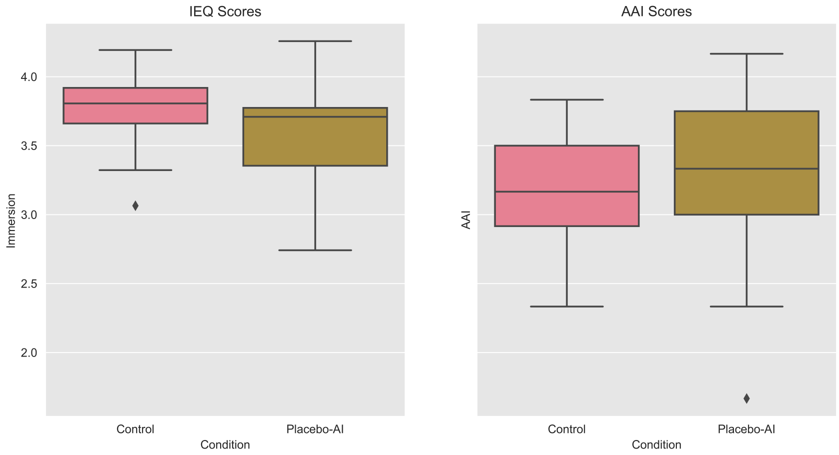
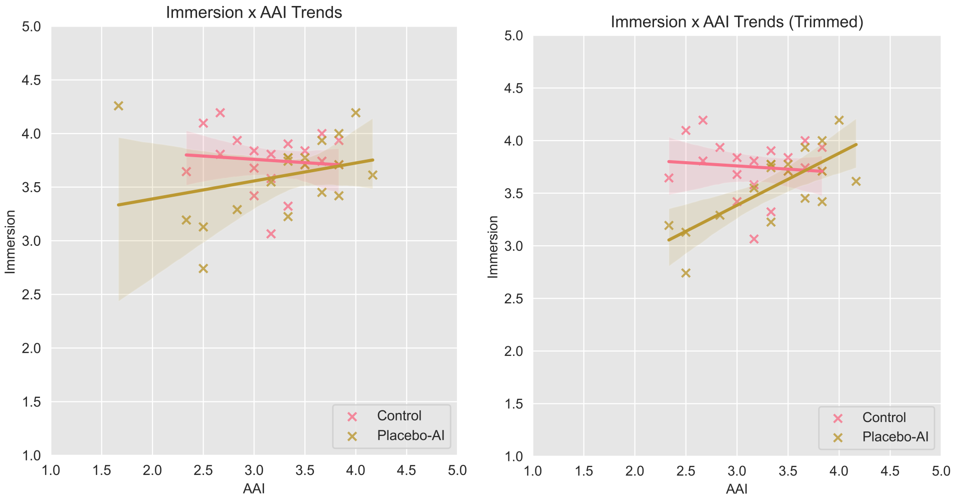


FIGURE 3.6: Boxplots presenting raw distribution (left), and trimmed distribution (right) of IEQ scores and scores of perception of advanced AI between the Placebo-AI and Control groups.

the Placebo-AI was being altered by one specific participant allocated to the placebo-AI group who scored very high on immersion, but very low on AAI. Regressions were computed to quantify this difference, showing that the outlier substantially displaced to the line of best fit ($p = 0.046, r = 0.318, b_0 = 2.934, b_1 = 0.22$) compared to the same sample with this outlier removed ($p = 0.367, r = 0.145, b_0 = 3.38, b_1 = 0.09$).

Qualitative remarks were noted during data acquisition for this specific participant. The participant in question was someone that reported not having played many games before, and described themselves as someone who was not a gamer. Despite this, the participant was observed to have performed better than the average participant in the game based on the investigator's subjective opinion of how they carried out their in-game actions, and survived the night. They were audibly reacting to events in the game such as combat incidences, where following a particularly difficult fight near the end of the session, they sighed a breath of relief. In another encounter, they responded acutely to the game with a sudden movement of the mouse arm. At the end of the study, they reported that the game was very fun, difficult, and quite stressful. The participant cited this stress as a significant reason to their not playing games much. Based on their IEQ score, the participant could have been considered as very immersed in the game (relative to most other participants) for the duration of the experiment. Despite this, their AAI score was the lowest among the whole sample, and by a considerable margin. This participant therefore either did not believe the game to have any adaptive capabilities or possibly failed to understand the nature of the concept, but just enjoyed the game anyway.

Given this potential outlier a second visualisation was also generated with this participant omitted, and both plots are presented in figure 3.12. The second plot with the omitted participant presented a result with a much



(A) IEQ x AAI Scatterplot

(B) IEQ x AAI without outlier.

FIGURE 3.7: Scatter graph of participants' AAI and IEQ scores. In the first plot with all data, observe the single participant at the top left, for whom a very high immersion score and the lowest AAI scores were recorded. When omitting this datum (who belonged to the placebo-AI group), the coefficient of fit for the placebo-AI group was much larger, as seen in the plot of the trimmed sample (right).

stronger positive linear relationship between AAI scores and IEQ scores in the Placebo-AI group.

In addition to being an interesting qualitative outlier, the participant was also a considerably skewing presence in the data. Linear fits were compared between the full sample including the outlier participant (presented in figure 3.7a, and the trimmed sample with this outlier removed (presented in figure 3.7b. A linear regression of the full sample ($r = 0.145, p = 0.367, SEM = 0.100$) resulted in a weaker and non-significant relationship between AAI and IEQ than a regression conducted on the trimmed sample which revealed a moderate and statistically significant relationship between IEQ and AAI scores ($r = 0.318, p = 0.046, SEM = 0.107$). A plot presenting this fit across the entire sample, irrespective of their experimental condition is presented in figure 3.8.

However, it is also clear that such a regression model actually disregards a critical mediator of the relationship between IEQ and AAI scores, which is the experimental condition in which participants were allotted. In both plots presented in figure 3.12, it is clear that there are two different effects between AAI and IEQ being presented. First, the control group presents a null to potentially slightly negative relationship between IEQ and AAI scores, whereas the Placebo-AI (i.e. the treatment group) presents a positive relationship between IEQ and AAI scores. Those who believe firmly in the AAI score remarkably higher IEQ values than those that do not, and the number of participants who did record higher AAI scores were more numerous than those that did not. Inversely, failure to believe in the presence of adaptive

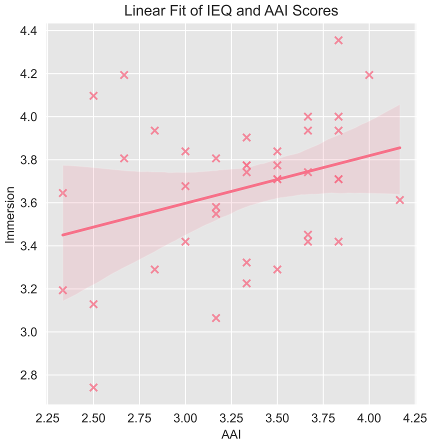


FIGURE 3.8: Linear fit to AAI and Immersion scores for all participants across both groups.

AI resulted in considerable detriments to immersion, as seen in the lower left quadrant of the plot in figure 3.7b. Interestingly, low AAI scores did not have such an effect on immersion in the control group, where participants were not ever exposed to any deception that may have failed to take hold.

The full picture of these results suggest that if there had been enough participants who had believed in the deception of an adaptive AI being present, the experimental manipulation may have replicated that of Denisova. Furthermore, given the number of participants in the treatment condition that did score higher AAI scores, it may have been possible that the study was under-powered enough to fall short of detecting a genuine effect of the deception. Had a larger sample been recruited, the number of participants who did believe in the deception may have been a greater enough proportion of the full sample that the distribution of the final IEQ and AAI scores would have closer resembled that of Denisova's original study.

3.8.3 Pupillometry

Pupil Diameter Results

In section 3.5, one of the stated hypotheses were that there would be statistically significant differences in pupil diameter between different times of day. This is based on the idea that the time of day within *Don't Starve* could be treated as a within-samples condition.

Every participant experienced the different times of days in Don't Starve, with most participants experiencing two full cycles. Different times of days emitted different luminosity and encouraged different in-game behaviours. Because of these differences, time of day was used as a variable to verify that the pupils were behaving as expected based on human anatomy.

Indeed, pupils were observed to be dilating differently based on the three times of days (Figure 3.9), which was expected given the difference in luminosity.

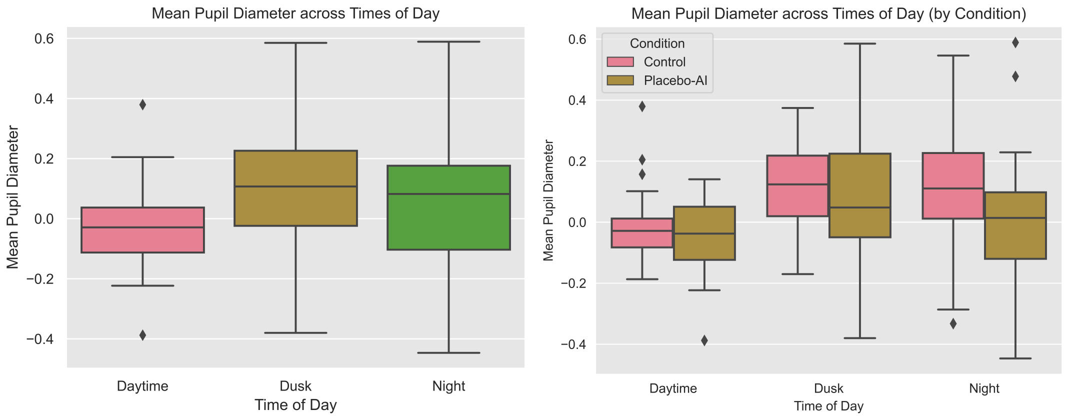


FIGURE 3.9: Boxplots of mean pupil diameter across the three different times of day. The left figure presents this across all participants, while the right also divides the samples based on their experimental condition. In both cases, differences are observable between different times of day. Note here that the average line represents the median, and conditions are labelled as: 0 = Control, 1 = Placebo/Treatment.

Visual inspection of the pupil diameter boxplots suggested there were potential differences between the conditions for each time of day. To confirm this, a mixed model ANOVA was carried out on the data. The between-samples variable were the experimental condition of either placebo-AI or control, and the within-samples variable were the times of days that all participants experienced. As predicted, there were indeed significant differences between the pupil diameter means of each time of day ($F(2,72) = 6.95, p = 0.002, \eta^2 = 0.162$). However, there were no significant effects between the two conditions ($F(1,36) = 0.24, p = 0.63, \eta^2 = 0.007$), and no significant interaction effects ($F(2,72) = 0.075, p = 0.928, \eta^2 = 0.002$). A post-hoc analysis of the time of day differences in pupil diameter was carried out and results Welch's t-tests indicated that there were statistically significant differences between daytime and dusk ($T(37) = -3.97, p < 0.01, g = -1.24$), but not between dusk and night ($T(37) = 1.89, p = 0.066, 0.35$) or between daytime and night ($T(37) = -1.83, p = 0.075, -0.55$).

Correlating Pupil Diameter with Immersion

Another of the initially stated hypotheses was that there would be a significant correlation between pupil diameter and immersion. Therefore, correlations were computed to test this hypothesis. The approach taken to this analysis was to conduct three independent correlations, one for each time of day. This method was considered to be most preferable when compared with either fitting a single model correlating IEQ scores and pupil diameter while treating the times of day as a third term, or alternatively computing three independent correlations with an adjustment made for multiple comparisons.

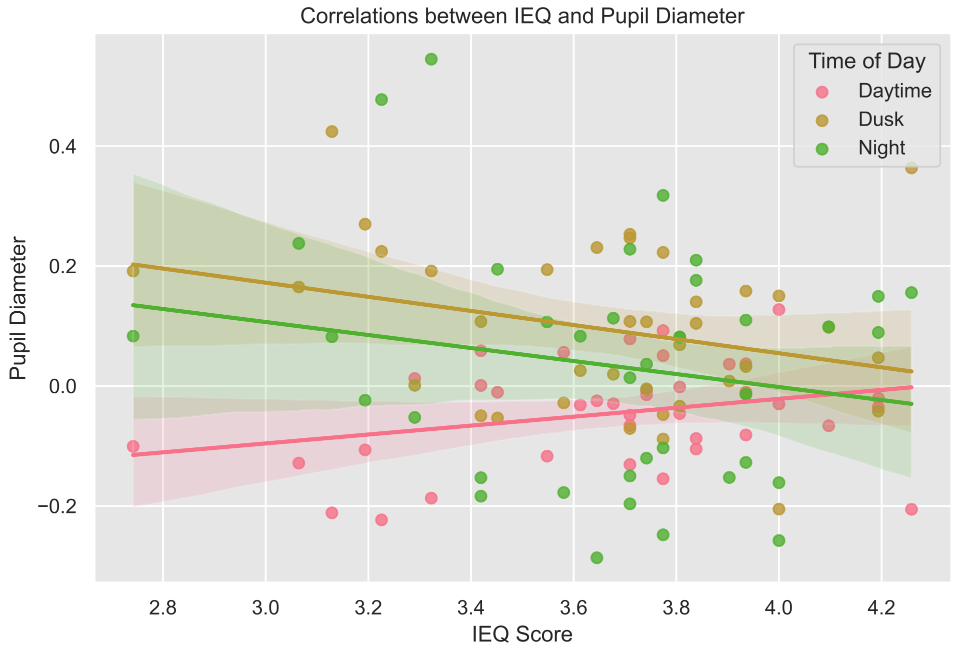


FIGURE 3.10: Correlations between pupil diameter and time of day, with a colour for each time of day.

With regards to the former option of fitting a single model, such an approach was not taken due to the imbalance of the sample whereby there was no guarantee that there would be an equivalent number of sample means for each time of day, nor was there a guarantee that the current sample size used to fit such a model would yield a reliably interpretable result. Additionally, corrections for multiple comparisons were not made given the already low power nature of the experiment. Further, the interpretations of the results of these correlations were made under an assumed context of potentially high noise, presence of confounds, and low reliability.

Figure 3.10 presents the scatterplot and corresponding correlation line plots for each of the correlations computed for this analysis. Statistical results from the correlations are provided in table 3.1. Correlations were computed using Pearson's correlation, with confidence intervals for r correlation coefficients computed using Fisher's transformation. No corrections for multiple comparisons were made. None of the correlations were statistically significant, though both Daytime and Dusk could have been considered to be borderline cases.

Interpreting these results comes with several caveats attached. First is the aforementioned acknowledgement that there are likely to be an inadequate number of samples to assume that the experiment is adequately powered. Second, it should be kept in mind that the times of days span different durations. Daytime lasts the longest by a considerable magnitude, which may have contributed to the trend of the correlation being positive as opposed to the negative correlations for the night and dusk pupil diameters. In general, looking beyond the lack of statistically significant correlations, it would be hard to conclude one way or another whether or not Pupil Diameter has any

TABLE 3.1: Correlations between Pupil Diameter and IEQ Scores, across different times of day.

Time of Day	r	$rCI95\%$	r^2	p
Daytime	0.289	[-0.03, 0.56]	0.03	.079
Dusk	-0.291	[-0.56, 0.03]	0.03	.077
Night	-0.186	[-0.48, 0.14]	-0.02	.263

statistical relationship to immersion scores. Implications and improvements for further attempts at correlating the two measures are considered in the discussion.

Fixation Rate

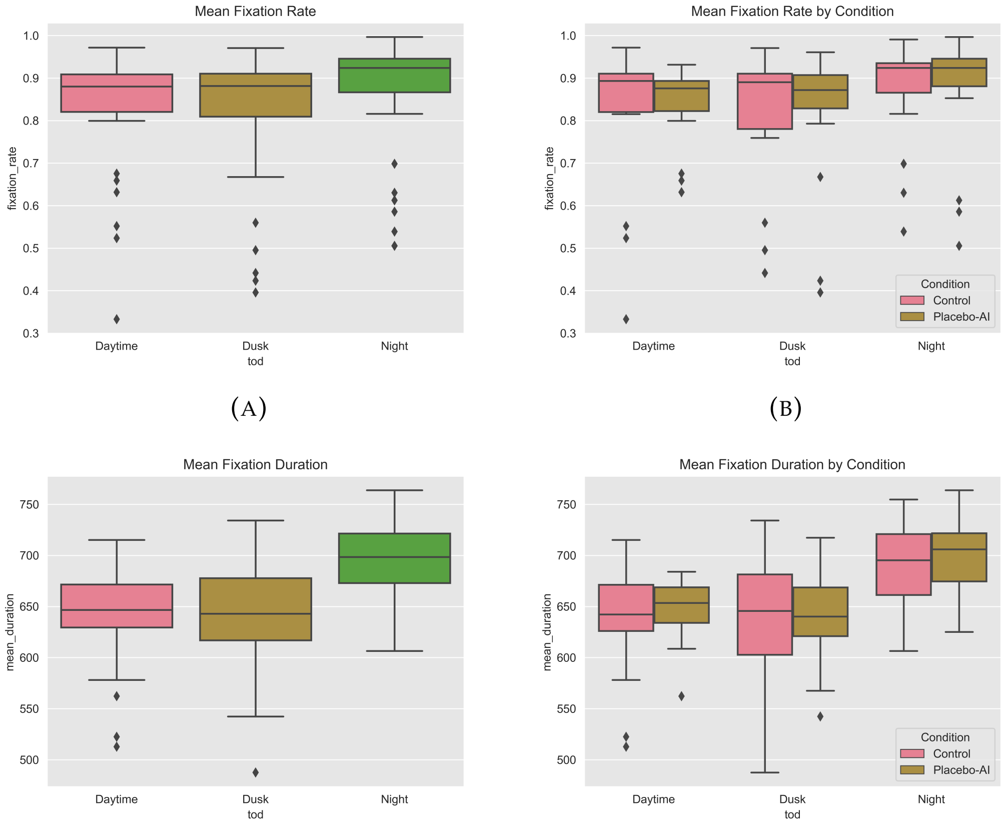
Another of the stated hypotheses in section 3.5 were that fixations and saccades would be different across experimental conditions and in-game time of day epochs.

Like the previous analyses for pupil diameter metrics, the initial point of exploration was across different times of day and between experimental conditions. For visualisation, there were two primary metrics computed. Average fixation duration was computed based by computing the mean of all fixations in each combination of time of day and condition. Fixation rates were also computed as a means of measuring the number of saccadic eye movements taking place in a given epoch. These rates are calculated by taking the number of fixations and dividing them by the duration of each epoch, i.e. each time of day. The distributions of these metrics are presented in figure 3.11.

As before, mixed model ANOVAs were conducted each for fixation rate and fixation duration.

First, for fixation rate, a mixed ANOVA was computed with times of days as the within-subjects factor, and the conditions as the between-subjects variable. Significant differences were observed between the times of day ($F(2,70) = 9.53, p < .001, \eta^2 = 0.214$), but no significant differences were observed between the conditions ($F(2,35) = 0.073, p = 0.789, \eta^2 = 0.002$), nor were any interactions found ($F(2,70) = 0.29, p = 0.751, \eta^2 = 0.008$). Pairwise comparisons were computed for a post-hoc analysis, and significant differences were observed between daytime and night ($T(36) = -3.363, p = 0.002, g = -0.33$), and dusk and night ($T(36) = -3.945, p < 0.001, -0.35$), but not between daytime and dusk ($T(36) = 0.598, p = 0.553$).

Another mixed model ANOVA was computed for fixation duration, and again significant differences were found between times of day ($F(2,70) = 51.3, p < 0.001, \eta^2 = 0.594$). Post-hoc pairwise t-tests found statistically significant differences in fixation durations between daytime and night ($T(36) = -9.55, p =< 0.001, g = -1.38$), and dusk and night ($T(36) = -7.67, p =< 0.001, g = -1.19$), but not daytime and dusk ($T(36) = -0.114, p = 0.910, g = -0.013$).



(A)

(B)

(C)

(D)

FIGURE 3.11: Boxplots of each fixation metric. Subfigures 3.11a and 3.11b present fixation rate distributions across times of day, and separated by experimental conditions. Subfigures 3.11c and 3.11d display mean fixation durations across times of day and grouped by conditions. The fixation rate boxplots reveal numerous outliers, which may be a consequence of measurement error.

3.8.4 Pupillometry results indicate a need for further analysis

Statistically significant differences in pupil diameter, fixation rate, and fixation duration were observed between the different times of day in the game. In addition to these differences being significant, the effect sizes were predominantly quite strong, often with an absolute $Hedge'sg$ greater than 1. However, it would be mistaken to take these results at face value.

First, it is difficult to extract definitive meaning from the pupil diameter between the day and night times of day given the luminosity differences between the two game states, meaning it is impossible to disentangle effects on pupil diameter from luminosity changes, from effects on pupil diameter due to gameplay elements.

A case therefore, could then be made that the focus should then be on the differences found in fixation and saccadic measurements between the times of days. However, in this case the issue with interpretation is that the source of the significant differences are largely uninteresting. This is due to the behaviour of the players during night time being considerably constrained, to the extent that most players have almost no actions to take during night time. This results in many players having few elements to visually search through, and many players simply fixating on the player at the centre of the screen or at the in-game clock awaiting the daytime to arrive. The result is the same- the significant differences in fixations between times of day yields disappointing value for exploring the measurement of immersion using pupillometry.

Finally, the distributions accompanying these results, presented in figures **??** and 3.11 indicate that though the distance between the means could be considered to be strongly separated, the 95% confidence intervals suggest that there is no statistically significant difference here.

The results from the planned exploratory tests therefore, did not reveal any reliable capability to capture immersion through simple pupillometry measurements over epochs defined within the gameplay. Therefore, a new exploratory analysis was conducted in order to try and control for the luminosity and gameplay confounds that plagued these initial set of results.

3.9 Further Exploratory Analyses

3.9.1 Gameplay Events

One of the purposes of collecting in-game events in as part of the data was to try and observe the consequences of specific situations on the players' psychophysiological signals. In particular, states of combat, death, and non-interaction were explored.

Passive Play vs. Combat

In *Don't Starve*, numerous high-risk situations can unfold where the players' investment of time is acutely at stake. Because of the possibility of a permanent loss of progress in any given session, players were expected to- and qualitatively did report experiencing tension in these situations during play whenever their character was threatened. A common case of such a threatening incident which was recorded during the experiment was combat, whereby players are attacked by monsters in the game world. Often, these combat engagements are not initiated by the players, and can even be unexpected when the aggressor is initially mistaken to be a non-hostile entity by the player. For example, one player believed that the game's frogs were not likely to be a threat until they were attacked upon approach.

Based on the stark experiential difference between these moments of combat and passive play such as during item gathering and exploration, it could be considered that there may have been differences in emotionality or cognitive load from the differences. These differences may have, therefore, resulted in observable differences in measurements taken during the experiment, and thus the pupil diameter and fixation measurements were once again explored to see if such differences could be found.

Defining Combat States

In Don't Starve, combat encounters are numerous and continue until a player flees, or a combatant is killed. In many cases, the player may even be fighting several combatants simultaneously. To define combat as recorded in the in-game event data, timeseries windows were generated around the point of an instance where the player is attacked. If the player is attacked by multiple enemies, then there can be multiple markers for an attacked event.

Therefore, to define a simple and generalisable instance of combat, combat sessions were defined by the 2.5 seconds before and after the first attack event in a sequence. A sequence was simply defined as a series of in-game frames in which no state transition occurred (i.e. the player did not change from peaceful gameplay to combat gameplay). Therefore, if a player had been attacked in the 2.5 seconds before or after a given frame of pupil measurement, that frame would be labelled as being in "combat state".

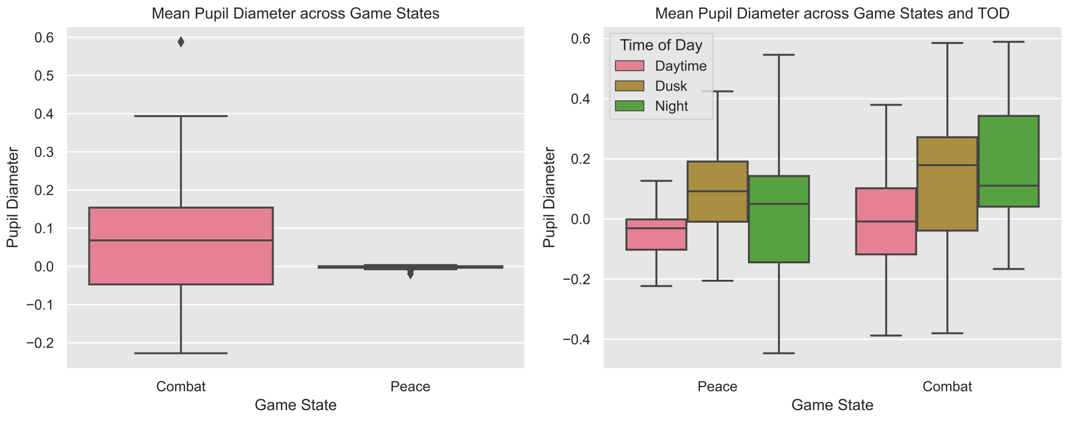
In addition to combat, an intermediate state of alertness between peace and combat was also defined as *picking*, where players would have opted to move towards a gathering node such as a tree to cut down or flora to gather from.

All frames captured in this experiment were then labelled as combat or peacetime, independently for each player. The resulting dataset could then be separated between each game state in order to compare any pupillometry measurement.

I. Mean Pupil Diameter across Peace vs Combat

Mean pupil diameter was computed within game states for each participant. The distribution of mean pupil diameter between game states as well as between game states across each times of day are presented in figure ??. Variance of the mean pupil diameters of participants within peace was near zero ($M_{combat} = 0.069, SD_{combat} = 0.187; M_{peace} = -0.0018, SD_{peace} = 0.004$), with very little distance between the minimum and maximum mean pupil diameters ($min = -0.019, max = 0.004$).

This drastic difference in variances was considered to be most likely due to the simple simply fact that combat comprised only a minuscule portion of the game, and thus the two contrasting data are disproportionate. Across all participants, there were on average 1888 pupil diameter frames recorded during combat, and 140297 pupil diameter frames recorded during passive/peacetime play. Given the much lower number of combat frames, the average diameter was therefore more likely to vary due to relative under-sampling.



(A)

(B)

FIGURE 3.12: Boxplots presenting mean pupil diameter of participants within each game state of either combat or peace. Please note that the boxplot presenting pupil diameters between just the game states is not erroneous and that the mean pupil diameter among participants at peace did not, in fact vary much at all.

With these considerations in mind, a repeated measures ANOVA was computed to test for any significant differences between mean pupil diameters at different times of day, across combat states (visualised in figure 3.12b). Greenhouse-Geisser corrections were applied due to an expected violation of the assumption of sphericity. A statistically significant difference was observed in mean pupil diameters between each time of day ($F(2,74) = 7.911, p = 0.0013, \eta^2 = 0.176, \epsilon = 0.895$), and combat states ($F(1,37) = 5.75, p = 0.02, \eta^2 = 0.135, \epsilon = 1$), but no significant interaction effects were found ($F(2,74) = -5.1, p = 1, \eta^2 = -0.16, \epsilon = 0.904$).

Finally, a different approach to investigating the effects of combat on pupil diameter in a more robust manner was considered.

II. Mean Pupil Diameter around Peace to Combat State Transitions

An improved approach to analysing the differences in pupillometry based on game state was devised by replicating the shorter epoch windows originally adopted by Beatty et al. (Beatty, 1982), whereby the mean pupil diameters were computed in short windows around the events of interest. In the original Beatty study, these were task-evoked epochs of interest, whereas here they were event based epochs defined by taking frames around the centre of a combat transition.

Defining Epochs: Here, combat transitions can be determined by first defining an *Epoch* as a collection of frames with a homogeneous state label. For example, a collection of consecutive frames labelled as peace will comprise a single epoch, which will be labelled as "Peace". Were this collection of frames

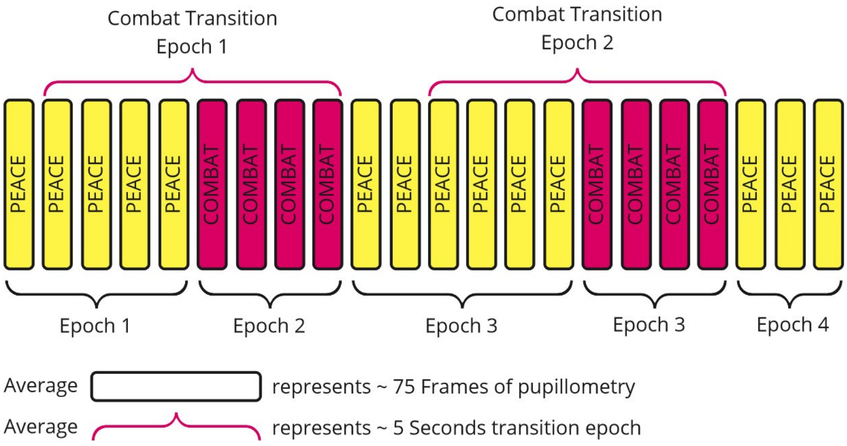
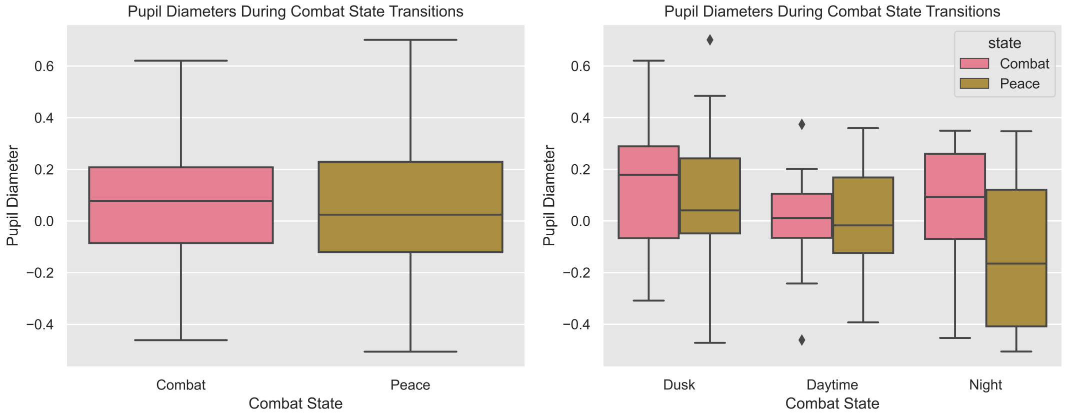


FIGURE 3.13: Illustrative diagram of the extraction of combat transition epochs. Here, transition epochs were defined as 5 second windows surrounding the change of a state from peace to combat. Transitions were then labelled as individual epochs within each participant.

not consecutive, and there were to be a smaller subset of consecutive "Combat" frames within the middle of this collection, then there would instead be 3 epochs: "Peace", then "Combat", and then "Peace". These Epochs are the initial basis on which combat transitions can then be defined, which is done by simply finding all frames where an epoch labelled "Peace" ends, followed by an epoch labelled "Combat". Such instances demarcate the player having recently had a period of peaceful play to then be attacked. These transition frames were then treated as centroids from which a *Combat Transition Epoch* could be sampled. *Combat Transition Epochs* are comprised of 2.5 seconds of frames, equally sampled from either side of a combat transition, meaning that 2.5 seconds of pupillometry from peaceful play, and 2.5 seconds of pupillometry from combat play are collected within a time window where one immediately precedes the other. An illustration accompanying this description of how epochs were defined and sampled can be found in figure 3.13.

By constraining the comparison of mean pupil diameters to these smaller windows, two problems with the prior analysis approach had been eliminated. First, taking an equivalent number of frames around the primary event of interest allows for comparison of means in equally sized samples, thereby eliminating sample size based differences in variance. Second, the examination of pupil diameter in a smaller window allows for the detection of more acute, phasic behaviour in the signal of interest. Consider that it could theoretically be possible that the power of pupil diameter in reflecting affect is greatest in short and phasic responses to stimuli- in such a case, the investigation of pupil diameter over larger samples spanning multiple minutes may have been an incorrect approach. A total of 65 combat transitions were observed in the data, across 31 participants, with an average of approximately 2 (precisely 1.88) combat transitions per participant, and a maximum of 5 combat transitions observed in 1 participant.



(A) (B)

FIGURE 3.14: Distribution of mean pupil diameters in the 2.5 seconds preceding and following a state change from *peace* to *combat* (left), and from *peace* to *combat* across different times of day (right). Sample means between *peace* and *combat* were very close ($M_{combat} = 0.06, M_{peace} = 0.04$)
.

A two-sided paired samples t-test was conducted between peace and combat mean pupil diameters, indicating no statistically significant difference ($T(2,30) = 1.12, p = 0.273, M_{combat} = 0.06, M_{peace} = 0.04$). The lack any detected significant difference indicated that mean pupil diameter in this experiment did not capture or reflect meaningfully upon game states and corresponding affective states that were likely to have occurred during the experiment. A boxplot displaying the distributions of mean pupil diameters around combat transition epochs is presented in figure 3.14a, clearly showing very close medians and largely overlapping IQRs. Here, a two way repeated measures ANOVA was not conducted due to a considerable imbalance in the number of combat occurrences between times of day, resulting in an imbalanced sample. There is, however, evidence to suggest that the combat instances during night induced a greater pupil diameter ($M_{combat} = 0.05, M_{peace} = -0.12$, and the implications of this are discussed below.

II. Considering (and rejecting) Further Investigation of Pupillometry Data

The next stage which was considered was to investigate whether fixations, saccades, and blinks may have differed around these combat transition epochs. However, this idea was ultimately not pursued based on several reasons. Initially, pupil diameter was the most discerning metric of interest as the mechanics of pupil dilation were not directly influenced by the overt behaviours of participants. To expand on this, a participant engaging in combat does not directly manipulate their pupil diameter as a function of *choosing* to fight a monster in the same way that they would influence their gaze movements. Fighting a target would necessarily implicate the movement of their gaze to the target of interest, thereby directly altering the metrics of fixations, saccades, and potentially even blinks to capture more than simply the affective state of the participant. A valid counterpoint would of course be to point out that pupil diameter is also always capturing luminosity responses, but again, the luminosity of the screen is not within the direct and comprehensive control of the player. Consequently, if the noise of pupil diameter was considered to already be a substantially high impediment on its ability to infer a participant's affective state, then fixations, saccades, and blinks were likely to be even worse. Furthermore, given that the statistical tests so far were intentionally applied without corrections for multiple comparisons due to a limitation to the power of this sample, choosing to conduct a battery of tests on three further metrics at this stage in the study could be considered to veer the statistical investigations of this chapter from simply exploratory into the territory of significance fishing. As a result, the exploratory analyses were halted at this point, and the overarching set of results were considered in the context of improving for the next study going forward.

3.10 Discussion

The overarching research question in this study was "Is there a relationship between physiological measurements, in the form of eye tracking, pupillometry, and electrodermal activity, and immersion as measured by the IEQ?". In order to answer this question, an experimental design by Denisova (Denisova, 2016) that has been previously replicated was used. The results of this study do not support the alternative hypotheses of either the IEQ scores showing a difference in immersion due to the experimental manipulation, or the efficacy of pupil diameter to distinguish affective states. No statistically significance differences of IEQ scores were observed between the experimental or control conditions, and while there were instances where pupil diameter was statistically significantly different in comparisons of interest, no meaningful interpretation could be made due to the effects of experimental confounds.

The findings here reveal critical considerations that must be taken when conducting a psychophysiological experiment, especially in the context of using commercial games as experimental stimuli. Furthermore, they also reveal important steps forward that can and should be taken for subsequent studies seeking to infer affective states or changes using physiological measurements. Finally, a number of specific considerations can also be made from the broader lessons of this study.

3.10.1 Failure to Replicate Experimental Manipulation

Originally, this experiment was designed with the use of an established manipulation of immersion as a platform to explore the initial outcome of psychophysiological measurements of immersion. This study therefore involved the replication of previous works that were assumed to be functionally suitable for a simple exploratory psychophysiological study, and that such measurements would yield interpretable (though not necessarily clear) results.

However, as it turned out, it is possible that even the core manipulation of the experimental design did not function as intended, despite adequately managing to replicate Denisova's previous work (sample size, participant documentation, etc).

Therefore, it is important to initially define what a failure to replicate means, in the context of this experiment. There were two manners in which a failure could take place. The first form of failure, is a failure to correctly replicate the previous experiment by Denisova, in the sense that the version of Denisova's experiment run in this study was not equivalent to its previous implementations. The second form of failure, is a failure to reproduce the previous findings of Denisova in that belief in an adaptive AI would induce differences in immersion. Here, we will discuss each of these forms of failure in detail, as reasons for why either of these failures could have occurred are potentially interesting considerations for future work.

One possibility is that the addition of physiological measurement devices have altered this study from one that is, strictly speaking, a replication of Denisova's experiment. This argument can be made on grounds of the negative ecological effects of such devices. For instance, it may have been the case that the use of multiple physiological signal apparatus may have resulted in a reduction in the overall immersion of participants that undertook the experiment. In particular, it is noteworthy that the use of electrodermal activity electrodes, placed on the fingers, may have been a great impediment on the ability for participants to feel immersed, as the hands are proactively involved in playing the game. Although this requires a further experiment to confirm the existence of such deleterious effects on immersion, it is not unreasonable to assume that whatever modality is chosen for future experiment should take greater considerations any factors of discomfort caused by the signal capture equipment.

Then, on the possibility that a failure to reproduce an effect occurred, there were some considerations made. Setting aside the default possibility that the Placebo effect of belief in adaptive AI on immersion is not actually real, there were other possible explanations for why the effect was not reproduced. For example, the sample size in this experiment was cut short of the target sample size. However, it was unlikely that the failure to replicate was due to a difference in experimental power caused by sample size differences, given that the sample sizes here were approximately the same as those in Denisova's original experiments.

In general, one of the lessons learned here is that the inclusion of physiological measurements not only involves the challenging management of apparatus, but also the subsequent effects that such equipment may have on participants' player experience as well.

3.10.2 Evaluation of Phasic Responses to Game Activity

One analysis in particular in this study aimed to conduct a robust examination of psychophysiological indicators for immersion by focusing down on smaller windows of play. This analysis involved the comparison of pupil diameter in the moments immediately prior to, and following an intense in-game moment of combat. This approach was also useful for reasons beyond simply being better controlled than the rest of the analyses in this study. Most importantly, it permits for the investigation of acute and phasic physiological responses to in-game events, which is something that psychometric based measurement of affect are unable to accomplish accurately.

It is possible that some of the results in this experiment returning a lack of statistical significance were simply due to the lack of control for multiple other confounds discovered over the course of the study. This may have been the case, rather than the lack of an effect outright.

Particular focus should also be drawn to the null results in pupil diameter between combat states. Here, no comparison of interactions between time of day and combat states was conducted due to an imbalanced and underpowered sample. However, there is a large enough difference in medians of pupil diameter between combat states at night to indicate that a larger, balanced, sample may have detected this difference. There is a theoretical basis on which this can be hypothesised: at night, combat state is an all or nothing affair. Players are either in a state of dire danger that is certain to end in their deaths, or there is no combat at all. It stands to reason therefore, that the acute cognitive load and activity of players would be greater in combat at night than in peace. Unfortunately, limitations to the sample that came as a result of the event based paradigm of the experimental design restricts a conclusive interpretation on this difference.

In the future it would be beneficial to continue this line of investigation as it is not only one of the more rigorous means of comparing two sets of physiological data, it is also a comparison that could only take place given the high temporal dimensionality of using physiological signals in the first place. The comparison of two means of a signal over the course of an experiment is hardly substantially different from that of two different questionnaire scores, whereas the comparison of short samples of signals in moments of interest, over multiple moments of interest (65 instances of combat, in the case of the present data), is something entirely unachievable with traditional psychometrics.

3.10.3 Consequences of absent baseline measurements

One of the critical shortcomings of this study with the pupil diameter data was the lack of a normalised/comparison baseline, especially with respects to luminosity. Ordinarily in vision and eye tracking research (e.g. more recently in immersion research by Cutting, 2018), a luminance-stable, zero stimulus measurement is usually included in the experiment in order to compare and subtract effects of luminance on pupil diameter. There are numerous ways this can be done, but that it is done at all is most critical. In Cutting's case, a baseline was computed by taking pupil dilation towards a fixation point prior to each trial, and in the case of Jainta Baccino (Jainta & Baccino, 2010), a baseline measurement epoch was included at the very start of the experiment.

For the present study, the omission of a baseline measurement may have limited the ability during analysis to normalise pupil diameters anatomically across participants, however, two important factors may have limited the damage to validity that could have transpired. First, the use of a within subjects design meant that ANOVAs and post-hoc comparisons should have adequately controlled for inter-participant variations in baseline or average pupil diameters. Second, while pupil sizes do vary between individuals, the variance is not so high that it would invalidate statistical analyses outright. Li & Huang have for example, used optical coherence tomography to sample a standard deviation of 0.85mm to 1.02mm in pupil diameter depending on whether participants were Asian or Caucasian (Li & Huang, 2009). The pupil dilation response can be several magnitudes this number, depending on the starting and ending pupil states, and there is no basis on which one might expect equivariant dilation between participants in this study. As a note, it is possible that a similar amount of variance could have been caused by simple adjustments of the eye tracker glasses from participants' heads as they undertook the study, which was accounted for in pre-processing by detrending the data.

Nonetheless, a baseline measurement would have been useful for additional reasons than simply being able to calculate a sample normalised pupil diameter, and that would have been to provide a non-immersive stimulus for comparison with the main experimental stimulus of playing an immersive game. At the very least, a lesson was learned that future studies should incorporate a baseline measurement for at least this reason alone, even if the modality being used is not pupillometry.

3.10.4 Considering possible confounds

The use of a baseline measurement is also relevant when considering an experimental task such as the one in this study, where there luminosity varies over the course of the experiment. For a 20 minute session involving a game with greatly variable stimulation and luminosity however, it is less clear how to take an appropriate baseline for participants as they play.

In particular, the time of day within the game would emit different luminances, and a single baseline prior to playing would likely be inadequate for comparison to any number of points of measurement during play. A participant could have potentially cycled through day, dusk, and night, two times during 20 minutes. This is also without mentioning the other mechanics which could have adjusted screen luminance such as the existence of a torch or another light source, or the colour of the world at that location. Therefore, at least in the context of the present study, the lack of a baseline measurement (or the inclusion of one thereof) alone would not have been the only critical confound that limited a more rigorous analysis.

Rather than controlling for the luminosity of any game, it may therefore be more preferable to design an experiment around a stimulus that is most suitable for the constraints a psychophysiological experiment places on the design of a stimulus game.

3.10.5 Reflecting on the choice of stimulus

Beyond the consideration of a baseline measurement, luminosity variances are generally not ideal for any experiment seeking to compare pupil diameter. The particular use of games as experimental stimulus make the management of luminosity very challenging, as they often produce different levels of luminosity based on in-game activity. One solution to this is to create a lab-designed game as a stimuli, as done by Cutting (Cutting, 2018). However, such an approach potentially constrains the ecological validity of the task, at least relative to what might be considered an "immersive game". Alternatively, there may instead be games or types of games that allow for the systemic control of luminance levels across the duration of an experiment. Such a game would have several critical, desirable properties. First, it must not have stochastic elements that vary what appears on the screen, and therefore no procedural generation or even emergent gameplay that varies the on-screen elements can exist within the game. Second, the game must allow for the identical presentation of on-screen and in-game elements between different participants while remaining "natural". Here, natural is assumed to mean that the design of the game itself would always have permitted for the same on-screen events to take place, irrespective of who is playing the game and what they are doing.

These criteria may appear to be impossible to meet while also meeting requirements for ecological validity, but there are, in fact, games that have both critical properties. For example, Tetris can be used as an experimental paradigm in such a way that all participants are presented with identical in-game objects in a luminance normalised paradigm, while remaining the same game in design and presentation as originally intended. Another example would include rhythm games, whereby the a level played twice would involve playing the same song, in a manner that could easily be luminance-normalised without impeding upon the design or presentation of the game.

3.10.6 Challenges with physiological apparatus

Throughout this study, parts of which were either mentioned or inferred in the write-up of this chapter, there were apparatus failures and data corruptions that prevented the collection and analysis of a full and complete sample of data as originally planned. The use of psychophysiology as a measurement paradigm is challenging and non-trivial, and in future work, greater care must be taken to ensure that there are stable data streams and contingency plans in place to deal with corruption of data or hardware failures. One such strategy may be to have resources planned to re-sample those participants whose data failed to be recorded for whatever reason. In general however, a greater lesson has to be conveyed that in any experiment aiming to use psychophysiological signals as a measurement of interest, considerations have to be taken a-priori in order to best stabilise the data acquired during collection.

Another lesson is the greater utility of psychophysiological modalities that have fewer points of failure. For instance, electrode based modalities that are susceptible to movement interrupting data integrity are less preferable to paradigms for which such a potential source of noise and interruption is irrelevant. In general, however, these signals often have a confounding source (or several) that have to be appropriately controlled for in a given experiment. In the specific case of this experiment, it is clear that modalities which require electrodes that must necessarily be placed on parts of the body which move during play are entirely inappropriate. Beyond this, when considering future work in general, for any given modality, careful judgement must be made about whether the apparatus required can impede upon the experience in any manner.

3.10.7 Conclusion

In this chapter, a first experiment was conducted to explore the use of pupillometry as a means of inferring cognitive states based on immersion in a game. The goals of this study were not met due to both a failure of the experimental manipulation to induce the expected outcome differences between groups, as well as a series of unexpected failures to appropriately manage confounds specific to the psychophysiological modalities chosen for this study. Critical experimental factors for future work were learned over the course of this study that can be incorporated to improve the rigour of future work, including the subsequent investigations in this thesis, irrespective of experimental modality chosen.

Chapter 4

More Granular Measurements of Mental Load with Heart Rate Variability in a Rhythm Game

4.1 Introduction

The exploratory study with *Don't Starve* revealed the challenges in getting a fine granularity psychophysiological signal that is clear enough to interpret or gain information from. Management of variability in data is one of the key constraining factors in conducting an experiment using such signals, with many confounding sources that can range from the game, the signal itself, the within as well as between individual biological differences, and perhaps most importantly, the cognitive process intended to be studied.

The purpose of this study, therefore, was to design an experiment where there is a deliberate manipulation of a specific part of a game, to observe if finer granularity from a physiological signal could be obtained with respect to a single cognitive facet. This defined goal then resulted in three clear requirements: a signal with appropriately managed extraneous confounding variables, an appropriate game that allows for a well controlled experimental design, and an appropriate cognitive facet intended for measurement.

The research question at the centre of this study is whether a physiological signal previously correlated or associated with cognitive activity in the literature, can consistently capture a more granular measurement of a player's experience during game play. For the purposes of this study, the modality that was chosen was heart rate variability (HRV), the game chosen for the experiment was a popular, commercial rhythm game titled *Osu!Taiko*, and the cognitive facet measured was mental work load, through the NASA-TLX. The reasons for these selections follow for the remainder of this introduction.

4.1.1 Choosing the appropriate measurement modalities

As discussed previously in the initial literature review as well as the first study, there are numerous viable choices that can be made when determining which physiological signal would be suitable for a player experience study.

The first study in this thesis incorporated pupillometry and electrodermal activity as correlate measures of immersion. One of the limitations mentioned was that both of these modalities had significant unexpected issues which arose during the experiment, and this was evaluated at length in the discussion of the previous chapter. It was concluded that such issues would have also occurred with other modalities, such as unexpected events complicating interpretation of a signal. More importantly than this however, irrespective of the particular idiosyncrasies of a given modality, video games usually have properties that are antagonistic to producing good experimental environments in which clean measurements can be taken, due to their very nature of being novelty driven devices. For a given oversight or unexpected confounding variable in the *Don't Starve* experiment, the confounding effect of the game would have also influenced an experiment using another physiological measure. The stochastic nature of *Don't Starve* would have added difficulties to interpreting impulse responses with heart rate variability, EEG, or even facial myography. Therefore, the premise with the experiment in the current chapter was that the specific choice of signal actually matters as much as the choice of game to be used as an experimental stimulus.

Another key learning was that the complexity of the modality itself did not inherently result in more information that could be acquired from the experiment. For example, in the previous study, the signal modality was a combination of pupillometry and electrodermal activity. Not only did the increased number of measurement devices possibly impede on the experience of players, the added dimensionality would have increased statistical complexity had the electrodermal activity equipment not failed. In other words, at least within the context of this particular lab, the management of signal-noise ratio became unmanageable with even a few additional sources of complexity. Furthermore, a single physiological modality known to correlate with affective states such as cognitive load could be more than adequate to begin modelling experience with more granularity than a psychometric test. For instance, an eye tracker could measure pupil diameter, gaze locality, gaze duration, blink frequency and number of fixations, all from a single physical source, each sampled at over a hundred times per second, resulting in a high-dimensional time series richer than one psychometric test. A single added dimension such as taking multiple measurements over time alone would have introduced more useful information than a single psychometric measurement covering the entire experiment. Meanwhile, all the additional dimensionality introduced by the eye tracker combined with five electrodermal electrodes did not compensate for the experiential disruption upon the player caused by the equipment. This experiential disruption was observed during the previous study through unprompted remarks from participants. Due to the lack of existing data on discomfort and a formal investigation into this particular aspect of psychophysiological player experience research being out of the scope of this thesis, observations and unprompted remarks from participants in the *Don't Starve* study were considered as a basis for not choosing eye tracking or electrodermal activity for this experiment. Whatever modality is ultimately chosen, it is therefore clear that minimising discomfort should be a critical point in the decision process.

There was an argument to take advantage of both the pre-existing research environment that was already prepared for pupillometry, as well as the fact that pupillometry alone would reduce discomfort compared to the previous experiment, it was still the case that the particular equipment used for eye tracking was not designed with comfort as a central priority. Based on our previous observations of the possible importance of comfort, alternate modalities were explored to see if a measurement device that prioritised user comfort was available. An option that was more accessible with a less computationally expensive output was heart rate variability. Logistically, heart rate variability mismeasurements or lost data caused by hardware failures were less likely as there were fewer sensors which were also more physically robust. The data output of an electrocardiograph was also smaller in size and lower in dimension, meaning that the pipeline from data acquisition to analysis could be completed more easily. Finally, some participants in the previous experiment had noted discomfort with wearing the eye trackers on the head. Modern commercial electrocardiographs are designed to be worn during strenuous physical activity with minimal disruptions to mobility or comfort. Based on this comparison, and numerous apparatus failures during and following the previous study, heart rate variability was chosen as the modality most appropriate for this study.

4.1.2 Game Choice

One of the observations from using *Don't Starve* as an experimental stimulus was the importance of tracking the causality between an event in the game and the signal being measured. A simple way to deal with this was to ensure that the game was unable to behave in unexpected ways, which can be achieved by simply picking a game with deterministic and linear sequences of events. Linearity ensures that there are no unexpected variations in the gameplay that can not be accounted for when processing or analysing the data. *Don't Starve* was a game which was procedurally generated and open world, meaning that numerous stochastic elements were central to the design of the game. On the other hand, the game chosen for this study, *Osu!Taiko* had several qualities which made it suitable as a stimulus for a psychophysiological experiment.

The first and most critical requirement was to constrain the decision space of the game, forcing players into a limited set of actions that they can take at any given time. By confining the playing space of a game into a so called linear design, one can be aware of not only the finite set of actions a player can take at a given time, but also what subsequent actions can occur later on in a play session. This addresses one of the critical issues with *Don't Starve* in that an open world design such as that of *Don't Starve* resulted in players acting in such a large and variable playing space that more data would have been required in order to fully sample a reasonable part of this space among our participants. Such rare occurrences then led to unknown cascading effects within the data that became intractable due to their infrequency. A more appropriate game to be used as a stimulus would therefore incorporate the aforementioned limited playing space, ideally one with few options that are still interesting enough to engage players in the right context.

Second, the game must be deterministic (or as close as possible) in the way it presents sensory elements during play such as visual graphics or auditory elements. Stochastically generated content is undesirable in a psychophysiological experiment due to the effects of large variability, as demonstrated in the *Don't Starve* study. This is mostly pertinent when the physiological signal is directly related to a sensory modality critically involved in playing, resulting in considerations such as the need to minimise luminosity variance in a stimulus used for an eye tracking study. Conveniently, this also meant that measuring a physiological signal such as Heart Rate Variability was less likely to be confounded by the elements of the game as a player does not interact with a video game using their heart rate. However, this does not mean that it is then possible to completely ignore the necessity for deterministic game elements, as any unforeseen events can have unexpected consequences on data, even indirectly.

Third, the game must be easy to explain and learn in the brief time frame of an experiment. This means that games with long periods of initial exploration, or long gameplay loops are not ideal for a setting where experiments can only last an hour or two at most, and in most cases even less (for instance, the *Don't Starve* study only had participants play for a total of 25 minutes). In the same vein, the game should not have too many complex mechanics for the participant to learn before they can achieve basic control of the game such that the learning period does not interfere with the intended experience.

Another requirement was a game that allowed for appropriate handling of failure states, or, ideally, the total removal of a failure state. Failure states were very common in *Don't Starve*, especially among participants of the study who were playing the game for the first time, or games at all, in some cases. Such a failure state meant that the total duration of game play was variable, due to a subset of the sample now having to wait an additional period of time to load a new instance of the game. In many games, including *Don't Starve*, dying also meant that participants had to start again from a far earlier point in the game, resulting in other feelings that may have impeded engagement, due to feelings of lost progress or frustration. Finally, a failure state would require the measurement apparatus to appropriately synchronise with an additional state of the game, resulting in more overhead setting up such an experiment. One approach to managing failure states would be the complete removal of a failure state, punishing participants for their mistakes in other ways such as detracted points from their final score. A game that allows for this alternative to a total reset of game state would be preferable, as it would then be possible to consistently test all participants for the same duration of time.

For the sake of dedicating more resources to the analysis of the data, rather than the capture, an important requirement was to choose a game for which the data was readily accessible. This would help to ensure both data quality and completeness, as the game developers would have exposed relevant elements of the game for an experimental investigator to then extract in-game information. In addition, this would ideally aid in synchronising the participants physiological data with the in-game event data.

Based on these above requirements, it was determined that a rhythm game would be an ideal type of game to use for the experiment in this study, due to a multitude of reasons. First, rhythm games are more understandable to a general audience, given the familiarity of many casual video game players to titles such as *Guitar Hero*. Rhythm games also generally involve very simple inputs (four buttons in the case of this study), and rule sets that are generally relatable to pre-existing understandings that people may have of playing music. Rhythm games are also structured around music, which is inherently linear. Deterministic gameplay, graphics, and audio can also therefore be achieved by ensuring that the game is relatively simple and player inputs are designed to revolve around the immutable properties of the songs they are playing to, such as the arrangements of the notes or the tempo of the track. An alternative type of game that meets most of the criteria listed and was also considered for this study but was ultimately not chosen were bullet hell games, such as *Just Shapes & Beats* or *Beat Hazard*.In the end, these games were not chosen due to lower control over the difficulty elements of a level, the requirement to reset the game after reaching a failure state, and less convenient access to the game's API and code base.

The specific game chosen for the experiment was *Osu!Taiko*. This is a rhythm game with a fixed, linear sequence of pre-determined and easily manageable events, meaning that players could take no divergent paths for players to take. Each event would also be presented to participants in the exact same order and at the exact same timestamp within an experimental epoch, meeting requirements for determinism. *Osu!Taiko* also has a rich ecosystem of user-generated content that could be conveniently re purposed for an experiment, including the possibility for an experimenter to change the number of notes within a level, or to change the tempo of a level so that the difficulty could be adjusted for experimental design purposes. *Osu!* and its various game modes, including *Osu!Taiko* also continues to enjoy, at the time of writing this, considerable popularity with a broad set of video game players, and its dedicated Reddit community holding an active 1000 users on its forums at a given time of day. All of these features made *Osu!Taiko* a suitable choice to be used as the experimental stimulus for this study.

4.1.3 Choosing the cognitive process of interest

The third requirement for a better designed psychophysiological experiment was to constrain the psychological phenomenon of interest to a more defined theoretical construct than immersion. This approach was preferred over attempting to continue measuring immersion, as a more focused and well defined target for measurement would be most likely to yield clear and interpretable results. Meanwhile, even though the ongoing research on immersion continues to grow, there are also ongoing discussions around its conceptually broad definition (Cairns et al., 2014a). Therefore, a clearer focus would most likely be ideal to create the circumstances for the greatest chances of interpretable and robust measurements.

One consideration on how to hone down immersion was to consider the fact that in its current definition, immersion is considered to be a multifaceted model of experience. Therefore focusing on a single one of these factors could constrain the scope of what to measure into a more manageable domain. However, this approach was considered to be lesser to that of an alternative option to instead focus on a psychological area that has close ties to immersion.

Attention and cognitive load are areas of psychology that have already been explored in the context of immersion. Cutting in his PhD thesis attempted the measurement of attention during game play with pupillometry (Cutting, 2018) with an arsenal of studies that developed an experimental design of using distractor recall tasks. However, this approach yielded mixed results. Although Cutting found that participants' performance in the distractor recall design did produce informative results of the engagement, he also found that pupil diameter and cognitive load did not vary significantly based on differences in cognitive load. Given that the modality chosen here was not pupillometry or eye tracking, but instead electrocardiography and heart rate (variability), cognitive load was considered a good potential target for measurement.

In the context of immersion, cognitive load could be considered as a form of task-induced mental activity where the task is considered to be the act of playing a video game. The clearer definition of cognitive load relative to immersion, allows it to be used as a target for the measurement of cognition using physiological signals. As the demands of a game on the player increases, so would cognitive load. The greatest advantage of cognitive load was the fact that there was already an established form of measuring cognitive load in a generalised task context.

4.1.4 Choosing the method of measuring cognitive load

For the psychometric tool itself, there was an additional desirable property other than validity and reliability that was identified from the results of the experiment in Chapter 3. This was the ability for a test to be quickly completed, such that it could be applied multiple times at intermissions within an experiment. Using such a test would meet one of the requirements to enable the design of experiments where physiological signals can be analysed in conjunction with multiple measurements of a cognitive process over time. This would be possible as the test measuring cognitive load could be completed more than a single time during the experiment.

To meet this need, the NASA-TLX was chosen as an instrument (Hart & Staveland, 1988) in order to quantify cognitive load by collected repeated responses over time. This psychometric test has seen widespread application in numerous industries and fields of research (Hart, 2006), and its short length makes the NASA-TLX an optimal choice for repeated applications on the same participant within a single experiment. It is important to note that the NASA-TLX is an instrument measuring mental task load rather than cognitive load directly. Nonetheless, the NASA-TLX was argued to be applicable to the experiment by taking into consideration the previously stated argument that playing a game is a task that induces a degree of cognitive load. Within this context, the mental work load measured by the NASA-TLX can be considered to be approximately analogous to cognitive load.

4.1.5 Research Questions & Goals

For this study, he overarching research question to be answered was "Is there a way to measure cognitive workload while participants play a rhythm game?". This research question was formulated to aim to achieve several goals for the study in this chapter.

The first goal of this experiment was to determine if a change in game-induced mental load would produce a corresponding change in physiological activity among participants.

The second goal was to explore if there were other possible causes of variation in HRV. After all, the NASA-TLX even in its small form factor, cannot capture information continuously despite repeated applications. Therefore, heart rate variability could be observed for any differences across known task load demands.

The third goal was a general goal to explore the feasibility of administering a psychometric test multiple times during a video game session. Although the NASA-TLX is not a measure of immersion, task load sensitivities to interruption would still be worth investigating to see if changes in measurements appear due to repeated measurements.

4.1.6 Hypotheses

With these research questions in mind, an experiment exploring the consequences of challenge in a video game was designed.

Specifically, the study aimed to determine whether playing levels of *Osu!Taiko* at different levels of difficulty would induce a corresponding, and matching change in both the task load and HRV measurements.

Before we proceed to stating hypotheses, it should be clarified that the terms challenge and difficulty are frequently used in this chapter. For all intents and purposes, from the perspective of conceptually viewing player experience, these two terms are treated interchangeably such that challenge and difficulty both refer to elements of the experience that increase the player's cognitive load. For the benefit of readability, referring to difficulty in the context of the stimulus game appeared to make the most sense, and so that terminology is used here. The nuanced difference between the term is partly based in the terminology of games as stimuli, and the terminology of the IEQ. Challenge is a sub-scale factor of the IEQ, whereas difficulty usually refers to elements or mechanics of a game that increase the challenge of playing the game skilfully. For instance, increasing the difficulty of a level in a game would also increase the challenge subscale score in the IEQ (were it applied).

On the subject of defining hypotheses to answer the research questions, four hypotheses were formed for this experiment, both based on analysing per-condition results in the NASA-TLX and Heart Rate Variability. Additionally, some supplementary hypotheses were formulated dependent upon the results following testing of the two central hypotheses.

The first hypothesis was that there would be a performance difference between the different difficulty epochs (discussed below in section 4.2.1) of the experiment:

1. There would be a significant difference in performance as measured by accuracy, between the easy difficulty and hard difficulty conditions.

Here, only the easy and hard conditions are chosen for comparison for two primary reasons. The first is to reduce the number of tests in order to minimise the familywise error rate. The second reason is that participants were instructed to treat the practice session as a space to learn the game and ask questions, which understandably could lead to unexpected variances. This was required as participants could not necessarily be expected to attempt to score optimally, which would cause any performance measurement to become less valid.

The second hypothesis was that there would be a difference in measured task-load between epochs:

2. There would be a statistically significant difference observed in the NASA-TLX scores between each level of challenge in the game.

Since the NASA-TLX measured cognitive load, it would follow that a minimally interactive task such as a resting state recording would not commit as much load for a participant as completing a trial block of the game. This hypotheses aimed to test this, in addition to testing differences in task load between two different levels of demand from playing the same game.

The third hypothesis centred on heart rate variability differences:

3. There would a statistically significant difference in HRV would be observed between different levels of challenge in the game.

Since HRV was used in this experiment as a physiological index of cognitive load, it would make sense that just as with the NASA-TLX varying significantly between different difficulties, there would also be a statistically significant difference in HRV between different conditions of the experiment.

The fourth hypothesis was formulated to tie together the NASA-TLX task load index scores with the HRV metrics:

4. There would a statistically significant positive correlation between HRV and NASA-TLX scores.

This would mean that any increase in HRV would have a corresponding increase in task load as measured by the TLX, thereby inferring that increases in task load could be measured with HRV.

4.2 Methodology

4.2.1 Design

The objective of this study was to attempt a more granular measurement of cognitive activity using heart rate variability. To this end, the aim of the experimental design was to measure the effect of varying levels of difficulty on either (or both) cognitive workload and heart rate variability. The experiment was therefore composed of two dependent variables (NASA-TLX scores, and Heart Rate Variability), and one independent variable (stimulus difficulty). The start of this section will define each measurement, with further details surrounding the technical specifications of each measurement found in section 4.3.1

Difficulty was defined as four separate states of engagement with the stimulus: resting, observation, easy, and hard. Resting was defined as an initial five minute period in which participants were asked to stay seated with eyes closed, as well as refraining from interacting with the investigator and the computer. Observation involved watching an AI complete a perfect playthrough of one level of the game (that participants would not get to play). Easy and hard stimulus were defined through objective metrics pertaining to the statistics comprising a level of the game. These four conditions could therefore also be seen as a gradient from minimally interactive (resting state and observation) moving towards maximally engaging (hard difficulty).

Then, following every trial for each difficulty, cognitive workload was measured through the full NASA-TLX psychometric test, including the second half of the TLX which recorded participants' weightings of each factor through the forced choice task battery. This second half has been treated as optional in many studies, the consequences of which on the scale reliability and sensitivity have been mixed (Hart, 2006). Heart rate variability was defined as the root mean square of successive differences between heartbeats, where the point of each heartbeat is defined at the R moment of the QRS complex. The collective set of differences between every R-R interval within an ECG recording was used to compute HRV, and each participant's ECG recording lasted the full duration of the main experiment including an initial resting state period. Consequently, the experiment itself was designed with a within-subjects paradigm in mind, and the remainder of this section details reasoning and justifications behind each of these decisions.

The experiment was designed around the fact that participants would have to complete successive levels of the game, which involved changing between different songs to play through as the game was musical. Therefore, the presentation of stimulus could be planned around natural breaks between experimental trial blocks, where one block comprised of a small collection of songs that correspond to that trial's difficulty. Breaks in gameplay between these blocks of levels could therefore be seen as natural by participants, and that period could then be used as a trial for collecting NASA-TLX data by having participants complete the questionnaire.

A second critical reason for taking a within-subjects design for this experiment was the ability to counteract interpersonal variations in heart physiology. Given that every individual would have varying levels of fitness, cardiac age, and engagement with the stimulus, it was paramount to control for as much variation as possible, and the best manner in which this could be achieved was to take a within-subjects design (Heathers).

Additionally, taking a within-subjects design for the experiment would reduce the total number of participants required for an adequately powered study. Although there are considerations to be made for the additional time needed within each participant to complete the experiment, if one considers the 5 minute total duration of a condition, and compare that to the total duration of ECG setup and debrief which exceeds this amount, it becomes clear that there is a more efficient use of time when taking a within-subjects design.

There were a number of additional considerations to account for by designing the structure of the experiment in this way. First, songs are not exact in their duration, and because a variety of different songs were chosen for ecological validity, the subsequent variation in duration of each song meant that the exact duration of each experimental block varied by a few seconds. However, trial blocks were designed to stick as closely to 5 minute intervals as possible. Furthermore, because of the task difficulty driven nature of the experimental stimulus, the order in which participants experienced the different difficulties of the game was randomised between participants in order to counterbalance any order effects.

4.2.2 Power Analysis

A power analysis was conducted in order to calculate the appropriate number of participants required for this study. However, there has been no previous work applying either the NASA-TLX or heart rate variability to an player experience experiment where the independent variable was the difficulty or challenge of a rhythm game. Consequently, there were therefore no previously measured effect sizes with which a power estimation could draw from.

This was further complicated by the fact that the effect of the experimental manipulation on the measure of the NASA-TLX scores would possibly be different from that of the HRV measurements. Therefore the ideal approach would have been to determine the smaller of the two effect sizes, and recruit a sample based on this lower bound estimate.

There was also a third constraint to the sample and recruitment for this experiment, which was the inherently time intensive nature of a psychophysiological study. Even with experimental stimuli designed to last no more than 15 minutes, an additional allocation of time of up to 30 minutes would be required for each participant due to the logistics of preparing for and debriefing a psychophysiological study. These complications include the placement of the ECG on the participant, ensuring that the device was not disturbing the comfort of the participant and that the signal measured was stable, and then removing the device following the main experimental procedure. This also therefore placed a constraint on the total number of participants that could be recruited due to these limitations of the laboratory in which the experiment was conducted.

Based on all of these factors, a smaller effect size of Cohen's $f = 0.2$ was assumed to be expected, which would be somewhat smaller than effect sizes recorded in previous studies utilising the IEQ. Such a conservative estimate was chosen in the absence of any effect sizes previously recorded with either of the measurement systems used in this study. With an $\alpha = 0.05$ error probability and a desired $1 - \beta = 0.8$ power for an experiment with four conditions (within-subjects groups) and two measured variables, an a priori power analysis estimated that an approximate 52 participants would be required in order to achieve adequate power.

4.2.3 Participant Recruitment

Participants were recruited through the University of York's Department of Psychology participant recruitment platform, and all participants were financially remunerated, with no rewards given out for course credits or study requirements among psychology undergraduate students. This was done in order to attempt to recruit more motivated participants with a wider variety of participant backgrounds as possible, though they would still predominantly remain students.

In actuality, because of national lockdown caused by the Coronavirus pandemic, only a partial total sample of 41 participants were recruited for the study completed the full experiment and recorded answers for the NASA-TLX. Of these 41 participants, a smaller sub-selection of 31 participants had complete physiological data that passed all acceptance criteria (detailed in section 4.2.4). Therefore, the resulting experimental results could not be expected to meet the originally specified criteria established in the power analysis. Nonetheless, analyses were conducted with this caveat in mind, assuming a modest expected effect size of $d = 0.35$.

The participants were of mean age $M = 22.55$, with reported genders of 30 female, 10 male, and 1 other. 37 participants were right handed and 5 were left handed. 15 participants had heard of the game and game mode used for the experiment, of which 9 had played the game. 29 participants had experience playing video games, with an average of 13.86 years of experience gaming. 10 participants reported being able to play musical instruments, with an average of 8.25 years of experience. Finally, it should be noted that the Demographics reported here pertain to the larger sample of 41 participants that completed the experiment and completed all questionnaires.

4.2.4 Acceptance Criteria & Data Quality Measures

Acceptance criteria for participants were established such that only participants who completed the full experiment and then completed the NASA-TLX and demographics questionnaires were retained in the final sample. On the psychophysiological aspect of data collection, lessons drawn from the *Don't Starve* study led to the decision that acceptance criteria were made stricter to ensure data quality. These included requirements for participants to have complete measurements from the very start of the experiment until the end of the play session, complete epoch labels for each segment of the experiment that annotated which conditions they were in at a given point of the study, and accurately logged keyboard input during the experiment. Participants were also requested to ensure that they had eaten at some point prior to the experiment, and the experiment itself was always scheduled to take place after noon. Additional instructions to not consume caffeine was also given to participants. These additional constraints on sampling respondents were taken in order to combat the interpersonal variability in cardiophysiology that were modulated by their circadian rhythms and food intake at the time of undertaking the experiment. Of note, the ingestion of caffeine was checked at the initial introductory briefing of the experiment, where all but one participant reported following the instruction to avoid caffeinated food or beverages preceding the study. The culminating sample of physiological data specifically was comprised of 31 participants, as described in the previous section (4.2.3.

4.3 Materials & Equipment

Participants completed the experiment on a personal computer in the University of York Computer Science Home Labs, which utilised rooms designed to emulate a home environment. The room in which this experiment was conducted was designed to simulate a typical study or computer room. The personal computer on which participants completed the study was equipped with a GTX 960 graphics card, and participants played on a mechanical keyboard which is a peripheral designed for playing games that is often preferred by gaming populations that play on a PC.

The full NASA-TLX was used for this study, including the additional dual forced choice section used to calculate weighting coefficients for individual task load factors. This questionnaire was delivered to participants in a digital form using local instances of the *Qualtrics* questionnaire platform.

Within this questionnaire, an additional question was included which consisted of two simple likert scale items scoring participants' agreement on the statement "After a certain point, I gave up trying to play well.", and "I enjoyed the songs." These items were included as indicators of potential confounds from either subjective impression of the music chosen for the experiment, or difficulty that participants considered so insurmountable that they disengaged from the task.

Heart rate variability measurements were obtained using the *Polar Pro* chest strap consumer grade electrocardiogram, with an unofficial recording app written by an engineer of the *Polar* team. They were able to provide additional information on data formats and pre-processing utilised by the device, however these procedures were non-transparent and also immutable for all purposes of the experiment. Therefore, for any replicability purposes, the same hardware and software stack would be required. To this end an apk package of the software used was backed up for posterity, and can be requested for replicability purposes from the author of this thesis.

The ECG device was worn by participants using an elastic chest strap with a rubberised surface in the area of the electrodes to optimise the detected signal. This strap was cleaned with ethanol based medical cleaning wipes between each participant's session due to ongoing concerns with the Coronavirus pandemic at the time of data collection.

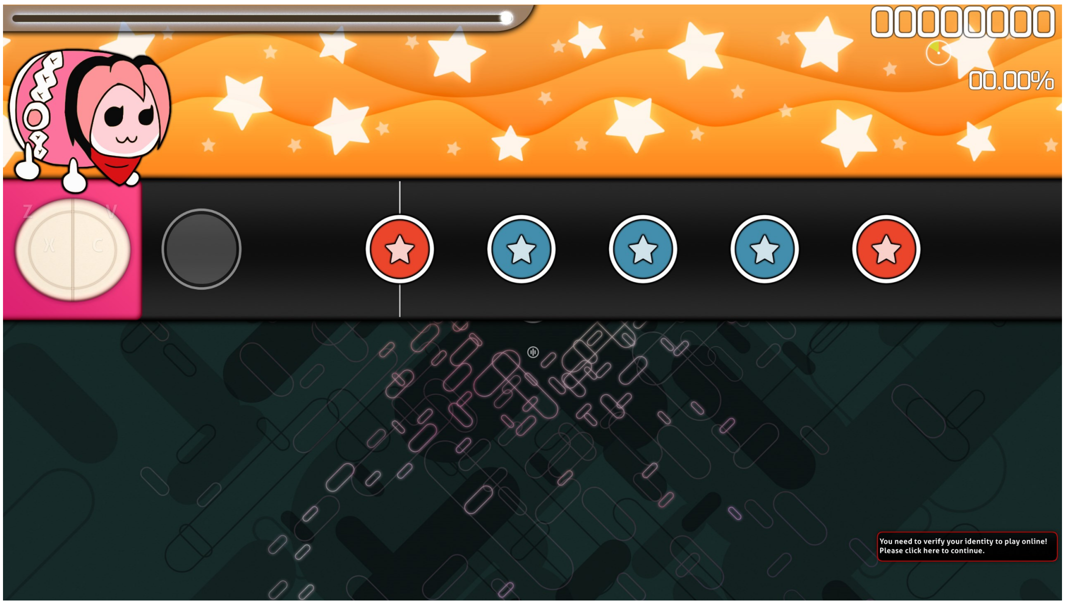
Data processing and statistical analyses were conducted using the Python programming language. ECG data was processed with the *pyhrv* (Gomes et al., 2019) and *biosppy* (Carreiras et al., 2015) packages, and statistical analyses were conducted using *pingouin* (Vallat, 2018).

The python programming language was also used to develop a keyboard input logging program for this experiment, as well as a logging and timer system that was used to synchronise the players' game session with the data fed from the electrocardiogram.

4.3.1 Measurements

As stated previously, the full NASA-TLX was used, including score weightings recorded from the second half of the questionnaire. TLX scores were computed for analyses using both their raw values, as well as weighted scores.

After consulting with the developer of the Polar recording app, it was clarified that the device was computing heart rate by detecting R-R intervals which were measured and fed into their HR algorithms. The detection algorithm for finding R-R intervals was proprietary and could not be further disclosed by the hardware and software providers. It was confirmed however, that processing of this data included moving average and median filters, a low pass filter, and a time cut-off filter for anomalous R-R interval lengths. Further details of these processing parameters were not provided due to non-disclosure constraints from *Polar*. ECG samples were acquired at approximately 130Hz, and all recorded data were exported to flat tabulated formats for analysis. The sampling rate was approximate due to small recording fluctuations in the hardware given the consumer grade nature of the device.



You need to verify your identity to play online!
Please click here to continue.

FIGURE 4.1: Screenshot of a *osu!taiko* game, displaying an influx of red and blue notes that the player has to hit upon overlap with the drum on the left hand side of the screen. The score and completed duration of the level are presented on the top right hand side of the screen. A drum avatar visualising the player's performance is animated on the top left corner of the screen.

4.3.2 Game

The experimental stimulus was the popular free, open source video game *osu!*, which is a simple rhythm game chosen for its deterministic and identical play-by-play procedure that lends itself for experimental research. More specifically, the *osu!Taiko* mode of the game is based on a popular arcade rhythm game titled *Taiko no Tatsujin*, and was chosen for the experiment due to its greater simplicity as compared to the main game mode. In *osu!Taiko*, players only interact with the game through two input buttons, which strike either the centre or outer rim of a drum, with the objective of the game being to score as many points as possible by hitting the correct notes with the right timing. A screenshot of the game is presented in figure 4.1.

In this game, participants were required to strike drum notes, presented as circles coloured either red or blue, at the right time upon the overlap of a note with the drum displayed on the left hand side of their screen. Specific keyboard buttons were dedicated to striking either a blue or red note on the drum, and this input layout is presented in figure 4.2. This input layout was intentionally designed for this experiment to be diagetically aligned with the in-game action of striking either the centre of the drum (red notes), or outer rim of the drum (blue notes). Players were also given freedom to start levels at their own pacing so that they could position and prepare themselves before a trial block in the experiment.

The game also included additional mechanics that demanded further contextual interactions from participants. The collection of these mechanics are

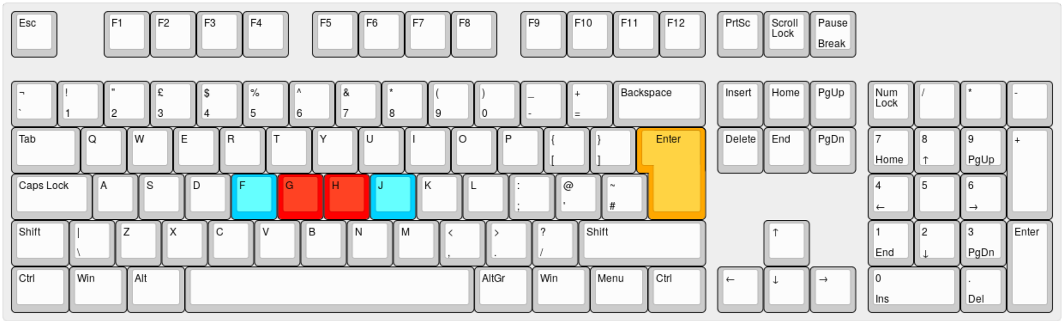
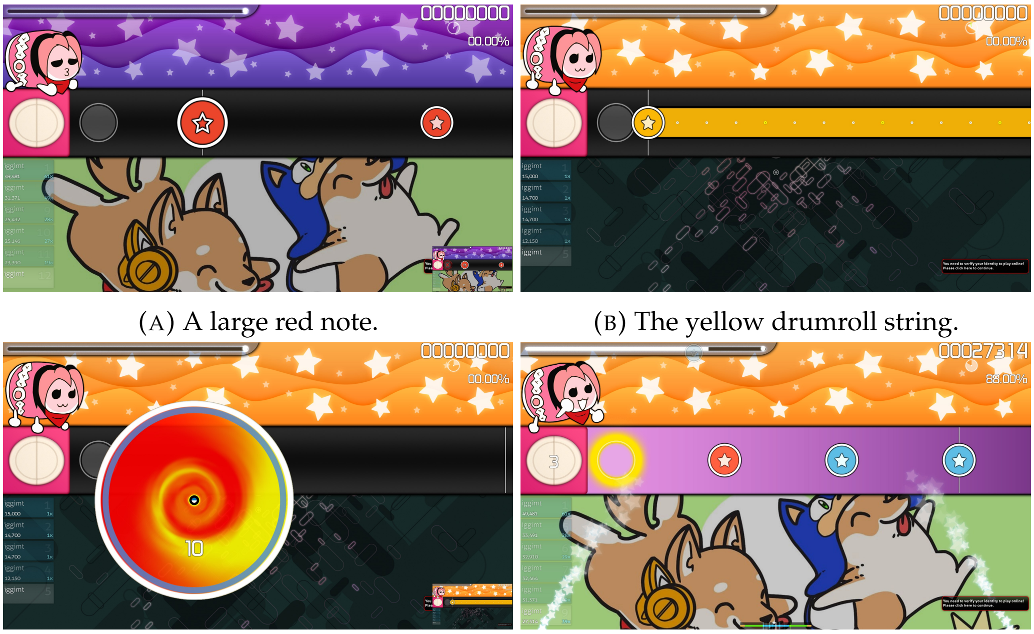

Esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12 PrtSc Scroll Lock Pause Break
¬ ` ! 1 " 2 £ 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 _ - + = Backspace Insert Home PgUp Num Lock / * -
Tab Q W E R T Y U I O P { [}] Enter Delete End PgDn 7 Home 8 ↑ 9 PgUp +
Caps Lock A S D F G H J K L : ; @ ' ~ # 4 ← 5 6 →
Shift | \ Z X C V B N M < , > . ? / Shift ↑ 1 End 2 ↓ 3 PgDn Enter
Ctrl Win Alt AltGr Win Menu Ctrl ← ↓ → 0 Ins . Del

FIGURE 4.2: The keyboard layout for *osu!taiko* used in this experiment. Colouring of inputs map to either the inside of the drum (red), outer rim of the drum (blue), or starting a level (yellow).

presented in figure 4.3. Occasionally, a larger note would appear instead of the regular sized coloured notes (4.3a). These large notes could be struck with either one or two buttons corresponding to that note's colour, with bonus points awarded to players if they opted to correctly time and strike both buttons of that note colour. Sometimes, a yellow stream would appear in place of the usual coloured notes (4.3b). These streams represented drumrolls, and players were able to strike any coloured note continuously to score additional points during the duration of the drumroll. Other times, a rare, large, spinning circle would appear during the game, which players could spin by pressing a sequence of alternating notes (red-blue-red-blue-...) as fast as they could, which would award further additional points to their total score (4.3c). Finally, during moments of the level in which the music was especially dramatic or eventful, the game would enter a bonus state in which successful note playing would result in additional points (4.3d).

As levels in *Osu!Taiko* were community developed by other players, specific care was taken to ensure that the content chosen was suitable for the experiment, with details to each chosen level presented in table 4.1. Because of the close cultural ties that the game has to Japanese and video game music, stimulus was chosen such that a mixture of a variety of genres was included. Furthermore, music was also chosen such that no vocals were present in any of the tracks so that participants did not experience any linguistically driven confounding effects on their experience of the game. Finally, levels were chosen specifically such that they were deemed appropriately suitable in belonging to either the easy or hard difficulty, due to the fact that rhythm games by nature often skew towards a skill ceiling that many novice players would find particularly challenging. To provide additional objectivity to stimulus selection, decisions were made with consideration of per-level parameters such as the tempo of a song (beats per minute), author subjective difficult rating (difficulty stars), total number of notes, and calculated actions per minute. Actions per minute were calculated by dividing the total number of notes n by the total duration in seconds s and multiplying by 60; $APM = \frac{n}{s} \cdot 60$. The full stimulus list chosen for the experiment is detailed in table 4.1.



(A) A large red note.

(B) The yellow drumroll string.

(C) Spinner circle with a strike counter.

(D) Bonus state shown by the purple bar.

FIGURE 4.3: Screenshots presenting examples of instances involving additional mechanics that participants encountered during play.

TABLE 4.1: Table of stimulus used in the experiment. All song names were left exactly as downloaded from the *osu!* beatmap directory.

Song	Block	Duration	Stars	Notes	BPM	APM
Trial of Thunder	Obs.	02:57	2.5	234	155	79
TBT Basic 1step	Training	00:16	0.5	5	150	19
TBT Basic 2step	Training	00:16	1	25	150	94
TBT Basic 3step	Training	00:16	1	27	150	101
TBT Basic 4step	Training	00:16	1	29	150	109
TBT Basic 7step	Training	00:16	0.5	5	150	19
TBT Basic 8step	Training	00:16	1	25	150	94
e	Training	00:39	1	42	70	65
Fun Fun Dayo	Training	01:33	1	104	128	67
Zelda Hime No Theme	Training	00:47	1.5	82	94	105
Canon Rock	Easy	02:16	2.5	161	98	71
Friends	Easy	01:21	1.5	90	140	67
Basstest	Easy	01:04	1.5	102	148	96
Time Trials	Hard	01:13	2.5	218	175	179
Kill The Beat	Hard	01:32	2.5	224	120	146
Egret and Willow	Hard	02:02	2.5	268	196	132

As previously mentioned, accuracy was also captured to test performance differences between difficulty epochs. A hard difficulty trial would presumably tax participants more heavily, and incur more mistakes from participants during the trial. A clear and objective measure of this can be found by observing the performance of participants, which could be captured by measuring the percentage accuracy of participants for each trial block. Here, percentage accuracy refers to the proportion of notes that participants were correctly able to play during a trial. This performance metric measured for each trial block was therefore hypothesised to be significantly different between each condition.

Finally, in the interests of standardising as much of the game difficulty as possible across every participant, the game was played with a setting that disabled a failure state. This prevented early and unexpected termination of the game, which was undesirable as it would have truncated electrocardiogram recordings and introduce a large amount of variability in player experience that would require a substantially larger sample size to manage as a confound.

4.4 Procedure

The experiment itself was split into six segments over the course of approximately 45 minutes.

Upon entry to the lab, participants were briefed on the procedure of the study (appendix: B.1). With the provision of consent (appendix: B.3), participants were guided through the first phase of the experiment which involved the fit of the ECG device. Participants were given the *Polar* electrocardiogram device, prepared and attached to the cleaned elastic chest strap. Any additional contextual information was provided and questions were answered pertaining to the details of wearing the strap, and then participants were provided with a private changing room in which they were able to place the device on their chest. Following this, a fit test was conducted to ensure that the ECG signal was functioning as intended.

Following this, participants were requested to stay seated in their chairs for a resting state ECG recording, and were guided to remain with their eyes shut and refrain from speaking to the investigator who remained present to monitor the recording. This block lasted for an exact five minutes, ensuring a resting state recording that matched the approximate duration of the experimental trial blocks. This resting period was included based on the discussions from the previous study, where such a baseline measurement was unavailable. Here, it served to capture a resting state physiological measurement such that the resting heart rate and resting heart rate variability of participants would be available if required at later stages of analysis. All subsequent stages of the experiment required some form of interaction or input from participants.

At all stages of the experiment including this observation block, participants were given the ability to start the level at their own volition using the

Return key, and the timing of the start of the level was synchronised from the game with the ECG epoch labels.

After the resting state recording, participants were seated at the computer and underwent the observation block of the experiment, which involved watching an AI play through a level of the game. Participants were able to ask questions during and following this observation period, but were forbidden from interacting with the keyboard during the observation session as this block was intended to be a minimally interactive and therefore minimally immersive experience. The utility of this observation block is two-fold. First, it enables participants to first observe the game in action, without immediately having to interact with it. Second, it was intended to explore the possibility of an intermediate state of engagement between resting and play. Observation might induce greater cognitive load than simply resting and doing nothing, but lower load than actually trying to play and meet the task demands of the game.

Once participants had completed the observation block, they were provided with a training block to acclimate with the experimental stimulus. The physiological and game data from this session was not included in the final analysis as it was determined that the purpose of this block was primarily for the benefit of trying to minimise differences of understanding among participants towards the task of playing the game. The TLX data from this session were used as part of the baseline comparisons with later gameplay sessions. Tracks in this section were provided in a non-randomised, progressive order as displayed in table 4.1. After completing all training levels, participants were also asked to complete the NASA-TLX as part of the training procedure. At the end of this training block, participants were given a final opportunity to ask questions or request a refit of the ECG device before continuing with the remainder of the experiment without interacting with the investigator (with the exception of requesting withdrawal from the study).

From this point forward, participants were randomly divided into one of two groups. One group would play the easier difficulty first, while the second group would play the harder difficulty first. Songs were randomised within each of these difficulty blocks per participant using the native python pseudo-random number generator and shuffler. Participants would complete three songs per difficulty block from the track list detailed in table 4.1. After completing all three levels in a block, participants then had to complete the NASA-TLX before continuing onto the next block. The separation between completing the levels and the NASA-TLX were also denoted in the ECG epoch labels such that the duration completing the NASA-TLX did was not included in the ECG epoch for that difficulty block.

At the end of the experiment, participants were then asked to complete a brief demographics questionnaire before concluding the main study. During the debrief, B.2 participants were also asked if any additional discomforts or problems were experienced during the duration of the experiment when they were not able to interact with the investigator. Additionally, demographics information was collected, including information regarding gaming experience as well as experience playing musical instruments which were included items in order to ensure the ability to control for any potential confounds on performance during the experiment.

4.5 Pre-processing

4.5.1 Psychometrics

Psychometric data in the form of the NASA-TLX was tabulated by the *Qualtrics* software platform, and then processed in python. Both raw total NASA-TLX scores as well as weighted scores were computed by computing the mean per-factor task load index. Weights were computed as instructed by the NASA-TLX manual (Hart & Staveland, 1988) and applied by multiplication with corresponding factor scores from the first half of the NASA-TLX. An overall task load score was also computed by taking the mean task load indices from a combination of each of the factors. This manner of calculating an overall score, rather than testing individual sub-components, is a common practice found across different uses of the NASA-TLX (Hart, 2006).

The reason why the mean is a valid operation to calculate the score is because dividing the score to take the average, still retains the same information as summing up the score for a total. This is because the average score still represents the spread of mental load across all factors. In addition to this, however, is the added benefit that it also becomes standardised. This line of reasoning can be demonstrated by the fact that if a hypothetical scale similar to the TLX consisted of 10 items, with 10 factors, summing up the scores and averaging would revert the score and scales back to 10 rather than 100. If one is instead interested in the prominence of certain sub-components over others, it would instead be better to additionally examine individual component scores in addition to this aggregate measure.

4.5.2 Telemetry

Behavioural telemetry was recorded in the form of key presses, and telemetry provided by the game. While player scores were originally recorded in the study, it was later determined that they were a poor indicator of overall performance due to scores being non-standardised across each track. Therefore, player accuracy was used instead as the performance statistic of interest, with accuracy defined as the percentage of notes hit by the participant during play. A mean accuracy statistic was computed from this data for each player, for each epoch.

4.5.3 Electrocardiogram

The ECG timeseries was separated by epoch labels defined during data collection. Start and end epochs of the study were truncated around the centre of their respective timeseries to 5 minutes. This was done due to the fact that the start and end periods of the experiment were variable subject to several circumstances such as any required corrections to the fit of the ECG, or the addressing of additional impromptu questions by respondents. Finally, it is noted that the periods during which participants were filling the NASA-TLX were also not included in the data collected from each block, as these periods were not considered to be part of the ECG measurement scope in the experimental design.

The ECG features were process using the *biosppy* default ECG feature extractor, the details for which are provided in the documentation and source code of the package, with the essential parameters provided below. First, an FIR bandpass filter was applied between the 3Hz and 45Hz frequencies. Following this, a Hamilton segmenter was applied and r peaks were extracted using the generated templates. This data was then passed into the *pyhrv* time domain based heart rate variability calculators, which computed HRV statistics based on the heartbeat features defined by *biosppy*. During this period, tachograms and Poincare plots were generated for each participant. The final output of this pipeline included a HRV statistic for each epoch, for each participant, in the form of the mean square of successive differences between normal heartbeats (RMSSD), which is a commonly applied form of HRV (Cowan, 1995). Henceforth, the abbreviations RMSSD and HRV should be considered to be equivalent, where RMSSD refers to the specific calculation of heart rate variability for this study.

4.6 Analysis Procedure

For the two main hypotheses, tests were conducted on the mean NASA-TLX overall score, and the RMSSD index of heart rate variability. To test the hypothesised difference in NASA-TLX scores between the two difficulties, an independent paired samples t-test was conducted on TLX scores between the two difficulty blocks across the sample. To test the hypothesised difference in heart rate variability, a repeated measures ANOVA was carried out on the RMMSD statistic across four blocks of the experiment (resting state, observation, easy, and hard). Subsequent post-hoc analyses were computed using paired samples t-tests when statistically significant results were indicated by the ANOVA.

As an auxiliary analysis to the comparison of HRV across each block, the heart rate of participants were also tested between the four blocks evaluated in the HRV ANOVA.

Exploratory analyses were conducted to confirm that the experimental manipulation functioned as intended, by comparing the average accuracy between the easy and hard difficulty block. A second exploratory test was also conducted to examine whether performance in the form of accuracy was correlated with the NASA-TLX scores, which would provide further evidence supporting the notion that the NASA-TLX was an indicator of difficulty induced mental load.

Finally, a correlation was computed between the NASA-TLX scores and RMSSD across the four experimental blocks included in the HRV analysis.

4.7 Data Suitability

As a reminder, the data suitability checks here will focus on equality of variances, rather than tests of normality. Details of the reasoning behind why are provided in Chapter 3, in section 3.8.1.

4.7.1 Performance Accuracy

Data suitability tests were conducted for the analyses on performance accuracy in section 4.8.1. Levene's tests for equal variances showed unequal variances in accuracy between the easy and hard difficulty conditions ($W = 5.29, p = 0.024$). Therefore, the t-test conducted for comparison of accuracy was adjusted by instead performing a Welch's t-test.

4.7.2 NASA-TLX Data Suitability

For the analyses comparing NASA-TLX scores in section 4.8.2, Levene's tests for equal variances indicated equal variances between difficulty conditions for both raw TLX score data ($W = 0.992, p = 0.374$), and weighted TLX-scores ($2.552, p = 0.08$).

4.7.3 Heart Rate Variability Data Suitability

For the repeated measures ANOVAs conducted on RMSSD in section 4.8.3, a Levene's test for equal variances showed equality of variance in RMSSD between conditions ($W = 0.769, p = 0.514$).

4.7.4 Heart Rate Data Suitability

For analyses on heart rate in section 4.8.4, Levene's tests showed that there was equality of variances in heart rate between conditions ($W = 0.139, p = 0.936$).

4.8 Results

4.8.1 Evaluation of Experimental Manipulation

The first hypothesis was that there would be a significant difference in performance as measured by accuracy, between the easy difficulty and hard difficulty conditions.

To test this hypothesis aiming to confirm the efficacy of the experimental manipulation, an independent paired samples t-test was conducted. A Welch's t-test showed a statistically significant difference in accuracy percentage between the easy ($M_{easy} = 80.68, SD_{easy} = 9.22$) and hard ($M_{hard} = 51.13, SD_{hard} = 14$) conditions ($p < 0.001, T = 20.99, df = 41$). This difference was also found to be relatively strong with a Cohen $d = 2.53$. This

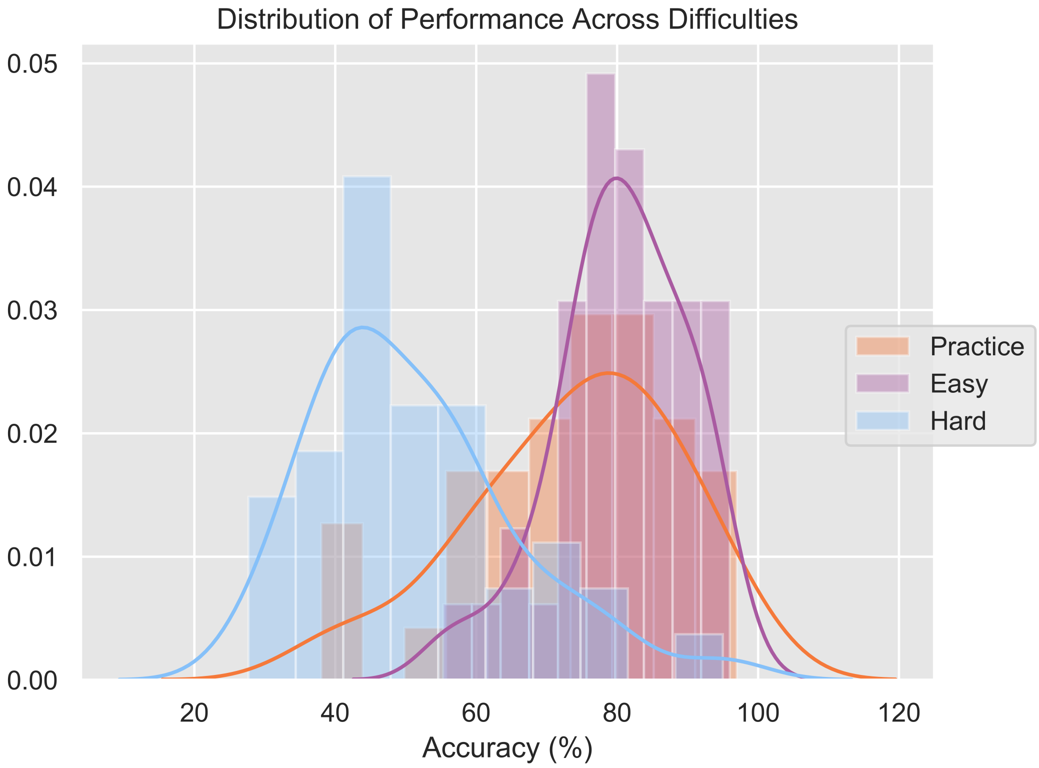


FIGURE 4.4: Histogram and Kernel Density Estimate for accuracy in each interactive block of the experiment. The practice block was included in this figure for demonstration of a theorised pre-acclimation difficulty that participants experience when learning the game.

difference is also visualised in figure 4.4, where the distribution of accuracy for the practice session was also included for further discussion below.

4.8.2 NASA-TLX Scores

The second hypothesis of this study was that there would be a statistically significant difference observed in the NASA-TLX scores between each level of challenge in the game.

Floor Effects in Frustration and Physical Load

Before proceeding with this hypothesis test, there is an obvious observation to address regarding floor effects. It is evident that floor effects exist for some the Frustration (4.6b), and Physical (4.6e) subscales of the NASA-TLX.

One interpretation of the flooring effects might be that the load of a factor was genuinely lower for easier tasks. This interpretation can be made based on the fact that the floor effect for frustration is stronger for the practice and easy conditions than the hard condition. Similarly, the floor effect for physical load is weaker as difficulty increases.

Additionally, the strong skew observed in the hard condition of the Performance subscale (4.6d) also indicates that the task difficulty might be appropriately reflected by the NASA-TLX scores here, and that the floor effect is appropriately representing a near zero amount of load on average across participants.

Hypothesis test

The NASA-TLX scores were found to have been different between the easy and hard condition irrespective of whether the scores were unweighted ($p < 0.001, T = -6.58[-15.99, -8.48], df = 41$, or weighted with the weights produced from the second half of the NASA-TLX ($p < 0.001, T = -7.91[-3.04, -1.8], df = 41$), with the only perceivable difference being the observed effect size with Cohen $d = 0.773$ for the unweighted TLX scores, and $d = 0.887$ for the weighted. The per-factor weighted scores are presented in figure 4.6, with unweighted values graphed in appendix item B. Here, the focus is placed on weighted scores due to adherence towards the originally intended use of the NASA-TLX. The overall weighted NASA-TLX load index is presented as a boxplot in figure 4.5.

4.8.3 Heart Rate Variability

The third hypothesis of this study was that there would a statistically significant difference in HRV would be observed between different levels of challenge in the game.

Upon evaluating assumptions for the ANOVA testing for differences in HRV between the four blocks of resting, observation, easy, and hard, it was found that the assumption for sphericity was not met, with Mauchy $W =$

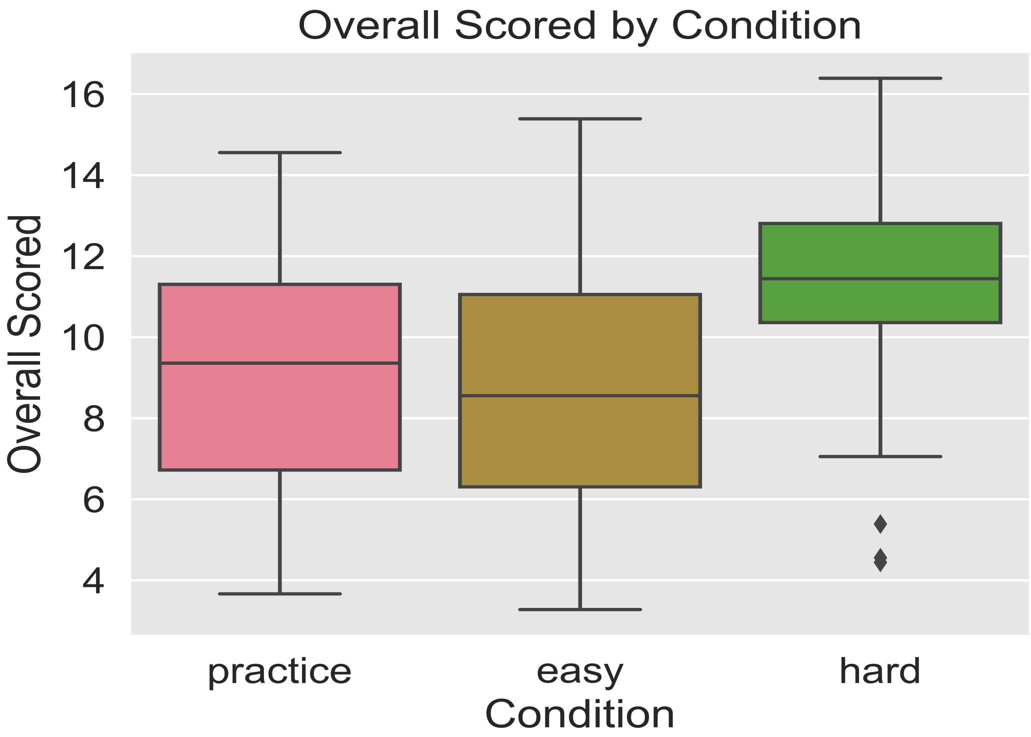
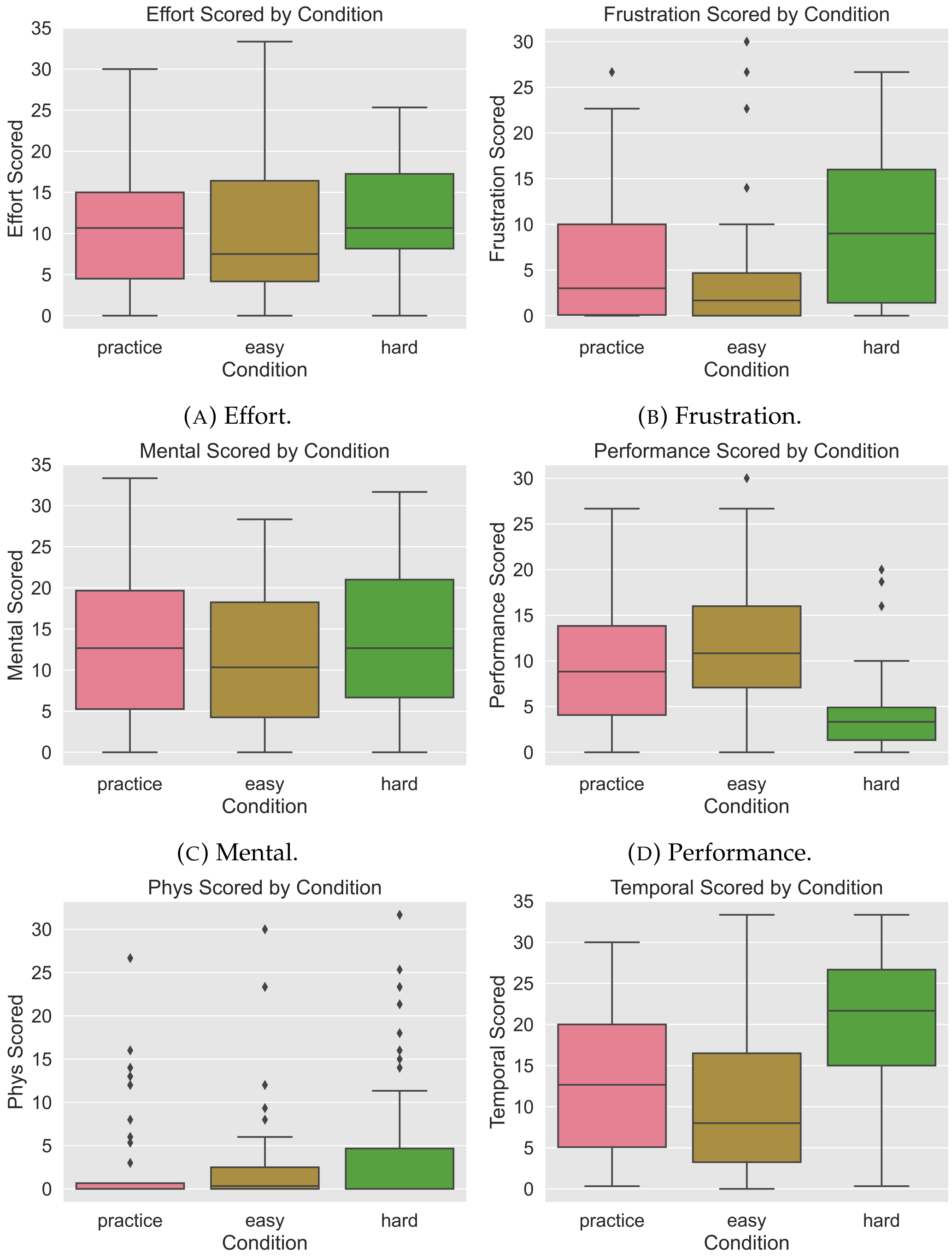


FIGURE 4.5: Overall weighted NASA-TLX task load index, computed by taking the arithmetic mean of all six factor scores.

$0.097, p =< 0.001$. While ANOVAs are often robust against violations of normality, lack of sphericity is often a concerning cause that requires correction. Therefore, the ANOVA was calculated using a Greenhouse Geisser correction. A significant difference was observed in the RMSSD between the four experimental blocks ($p_{gg} = 0.0256, df_1 = 3, df_2 = 90, F = 4.678, p_{raw} = 0.004$) with a roughly moderate effect size $\eta_{p^2} = 0.135$. This difference is visualised in figure 4.7.

TABLE 4.2: Post-hoc pairwise t-tests following the ANOVA for RMSSD.

A	B	T	df	p	Hedges g
rest	observation	1.52	30	0.142	0.275
rest	easy	1.92	30	0.064	0.37
rest	hard	2.87	30	0.008	0.528
observation	easy	0.91	30	0.371	0.088
observation	hard	3.06	30	0.005	0.269
easy	hard	3.12	30	0.004	0.196



(A) Effort.

(B) Frustration.

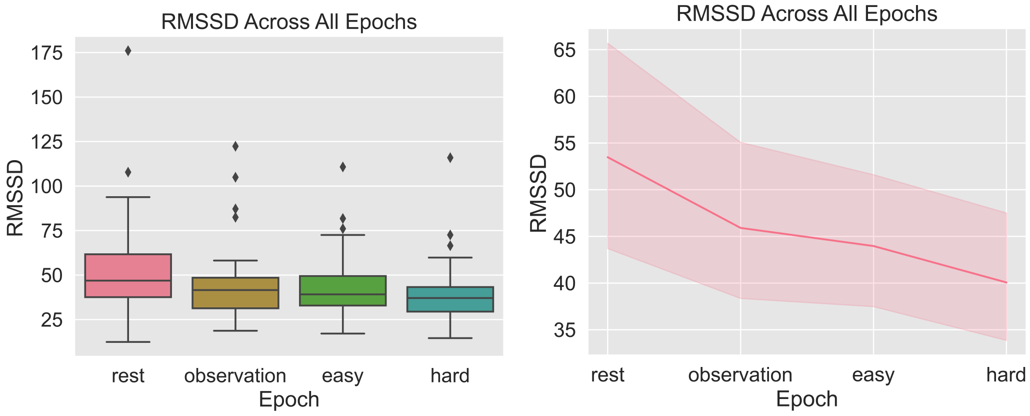
(C) Mental.

(D) Performance.

(E) Physical.

(F) Temporal.

FIGURE 4.6: NASA-TLX weighted scores for each factor, with weights computed via results from the forced choice questions in half 2 of the NASA-TLX. Note that the y axes are not standardised due to inherent variability introduced by the weighting function. Further, the practice block was included for further discussion later in this chapter.



(A) Boxplot of RMSSD across experimental blocks.

(B) RMSSD plotted as a function across all epochs.

FIGURE 4.7: Graphs of RMSSD across experimental epochs presented as a box and line plot. Although some participants played the hard difficult before the easy block, the function is presented in order of ascending difficulty to present the linear effect of task load on HRV. 95% confidence intervals were also included in the line plot.

The rather large confidence intervals are likely to be indicative of the under powered nature of the study, due to the early termination of data collection. Nonetheless, a negative trend of RMSSD is observed as the experimental task load increases, based on the observed expectation. This trend generally appears to support previous results from both the NASA-TLX and performance analyses, showing a steady change in participants' load induced by challenge as they progress through the experimental task.

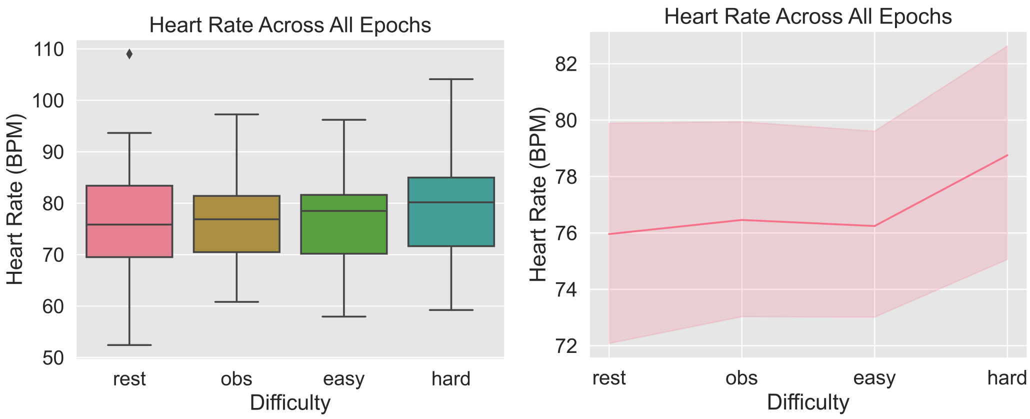
Following the results of the ANOVA, post-hoc analyses were performed with pairwise, within-groups t-tests, the results of which are tabulated in table 4.2. Among the post-hoc t-tests, it is clear that the statistically significant differences lie among comparisons between the hard difficulty epoch and the remaining epochs, which suggest that the RMSSD values did not change very much as the experiment progressed until the mental task load had ramped up substantially. A borderline result is also found between rest and easy, which is the second largest gap in task load possible between epochs in this experiment.

4.8.4 Heart Rate

To support the HRV analysis, heart rate across the four epochs of interest were also compared and these distributions are presented in figure 4.8. As with the HRV ANOVA test, the assumption of sphericity was violated in this analysis ($W = 0.44, p < 0.001$), and therefore a Greenhouse Geisser correction was adopted for this test. A statistically significant difference was observed in heart rate between the four epochs ($p = 0.019, F = 4.318, df_1 = 3, df_2 = 90$), with an approximately moderate effect size ($\eta_{p^2} = 0.126$). Also

TABLE 4.3: Post-hoc pairwise t-tests following the ANOVA for heart rate.

A	B	T	df	p	Hedges g
rest	obs	-0.646	30	0.523	-0.046
rest	easy	-0.277	30	0.784	-0.026
rest	hard	-2.382	30	0.024	-0.248
obs	easy	0.304	30	0.763	0.022
obs	hard	-2.737	30	0.010	-0.221
easy	hard	-4.296	30	<0.001	-0.242



(A) Boxplot of heart rate across experimental blocks.

(B) Heart rate plotted as a function across each epoch.

FIGURE 4.8: Graphs of heart rate across experimental epochs presented as a box and line plot. Like the HRV graphs, results are presented in order of ascending difficulty to present the linear effect of task load on HRV. 95% confidence intervals were also included in the line plot.

like the previous HRV results, there is a large confidence interval across all epoch scores which indicates that the underlying problem may be a highly variable measure.

Again, after the results of the ANOVA, post-hoc analyses were performed with pairwise, within-groups t-tests, the results of which are tabulated in table 4.3. Again, the most marked and only significant differences are found between the hard difficulty epoch and the remaining blocks of the experiment, indicating that a large increase in the load of the experimental task was required in order to induce a change in measured physiological signal.

4.8.5 Relationship Between RMSSD & NASA-TLX

The fourth hypothesis of this study was that there would a statistically significant positive correlation between HRV and NASA-TLX scores.

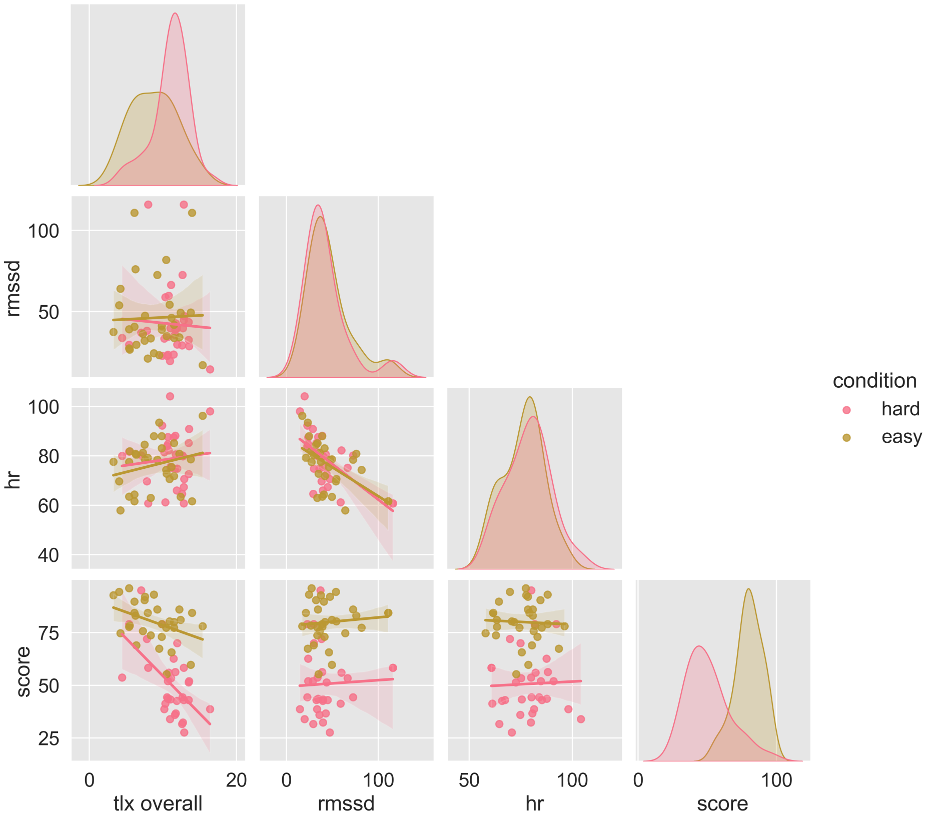


FIGURE 4.9: Correlation matrix presenting within-groups correlations between the measurements used in earlier stages of analyses.

A Pearson's correlation applied between the overall weighted NASA-TLX scores and RMSSD within epochs for all participants was found to be statistically not significant ($p = 0.798, r = 0.034[-0.28, 0.22], n = 0.62$). There was also no statistically significant correlation observed between RMSSD and accuracy scores ($p = 0.795, r = -0.034[-0.28, 0.22], n = 62$). These results are not entirely unsurprising given that there was already a fairly small effect size in the difference observed between easy and hard RMSSD ($g = 0.196$, table 4.2), whereas both TLX scores and performance had involved stronger differences between the two difficulties. This would have led to a difference between the sampling distribution of RMSSD and TLX or performance, which would explain the null result observed here.

A collection of correlations of the measurements analysed in this study are presented in the matrix figure 4.9, comparing the easy and hard difficulty conditions specifically (additional groups were omitted from this figure for reader clarity).

4.8.6 Confound Management

Finally, to verify the results calculated so far, the potential presence of any confounding effects and any subsequently required management thereof were explored.

Due to the musical nature of the game chosen for the experiment, the existence of any confounding effects from any existing familiarity with playing musical instruments was explored. Pairwise t-tests found no statistically significant differences in NASA-TLX weighted overall scores ($p = 0.28, T = 1.09[-0.75, 2.5], df = 37.4, d = 0.289$) between those who played instruments ($M = 10.2, SD = 3.06, n = 10$) and those that did not ($M = 9.32, SD = 2.96, n = 23$), with no interaction effects between playing an instrument and the easy condition ($p = 0.62, T = 0.5, df = 19.25$), or the hard condition ($p = 0.26, T = 1.16, df = 14.33$).

Another familiarity based confound was also tested in the form of any pre-existing knowledge or experience with the game used as the experimental stimulus. Again, there was no statistically significant difference in overall weighted TLX scores ($p = 0.68, T = -0.411, df = 30.33$) among those that had heard of the game before ($M = 9.68, SD = 3.5$), and those that had not ($M = 10.05, SD = 2.84$). No statistically significant interaction effects were also observed between familiarity with the game and NASA-TLX scores in either the easy ($p = 0.699, T = -0.394, df = 15.68$) or hard ($p = 0.844, T = -0.2, df = 12.32$) conditions.

In the performance of participants there were also no detected confounding effects from preexisting familiarity with musical instruments. Namely, no statistically significant difference was observed in accuracy percentage ($p = 0.943, T = -0.071, df = 32.69$) between those that played instruments ($M = 65.56, SD = 18.83$), and those that did not ($M = 65.95, SD = 21.17$). Further, no interaction effects were found between instrument playing and accuracy in either the easy ($p = 0.807, T = -0.248, df = 17.64$), or hard ($p = 0.985, T = 0.019, df = 13.42$) difficulties.

4.9 Discussion

The purposes of this study were to determine the following. First, whether the NASA-TLX would be a suitable tool to measure cognitive workload for participants playing video games. Second, whether heart rate variability as a signal modality would be indicative of this measured cognitive workload. Third, whether the short form nature of the NASA-TLX would be applicable to experiments where the signal of interest is experiential— in other words, whether a shorter psychometric test would interrupt experiences such as immersion. In general, the results provide supportive although slightly limited evidence of the alternate hypotheses.

Pertaining to the first and third objectives of the study, compelling evidence was produced via the NASA-TLX and performance scores between groups that the experimental manipulation did work as intended, and that the NASA-TLX was able to capture the difference in mental workload induced by varying game difficulty. Furthermore, the NASA-TLX was able to capture this difference in its short form factor, the implications of which are discussed further below later in this section (4.9).

Second, a small statistically significant difference was found in HRV across the different levels of interactivity within the experiment. However, on the other hand is the fact that when inspecting this difference in a post-hoc analysis, only the most diametrically opposed conditions (such as between easy-hard, and observation-hard) yielded any considerable effects. Furthermore, there was also no correlation found between RMSSD and the NASA-TLX scores. These findings lend credence to the interpretation that for some reason, not all of the RMSSD behaviours that were expected actually manifested, and no empirical relationship could be established between the psychometric and physiological measurements of interest.

Explaining the Observed RMSSD Effects

When examining RMSSD differences between conditions, something of a "gentle trend" can be observed, with a steady decline in variability score as one moves from the initial resting state condition to observation, then from observation to easy, and from easy to hard. There are a few possibilities that explain the behaviour of this trend.

First, it could be the case that this trend is a direct reflection of the degree of interactivity or immersion that participants experienced during the experiment. However, this interpretation is not well supported by the lack of any significantly observed differences in the intermediary steps between resting and hard. Re-framing these results in terms of interactivity-differences rather than interactivity might also be more appropriate. For example, it could be hypothesised that the difference in the level of interactivity between resting (i.e. not playing at all) and observing the game is far greater than the difference between observing the game and playing a very easy level. The small difference between the observation and easy conditions when compared with the larger differences between rest and observation, and easy and hard, become more consistent with this framing of interaction differences.

Supporting this is the fact that the difference in interactivity between a resting state and observing the game could theoretically be quite large, and the borderline significant result lends credence to this idea (see table 4.2). However, when considering the broader set of effect sizes and their overall small sizes, it becomes difficult to make a compelling case at all.

Another explanation could also be that RMSSD as a signal is not as sensitive to the gradual affective changes as induced by challenge based experimental design. In line with this thinking, it could even be suggested that the novelty of observing the game for the first time with anticipation of playing the game would be a fairly engaging experience almost in line with playing the game for the first time and fulfilling that anticipation, especially as participants had one interactive element as part of the observation task which was the ability to ask the investigator questions they had about the game. Nonetheless, these questions require further research to address and provide a basis for potentially designing experiments with greater and clearer differences in cognitive load or engagement between conditions.

Unmanaged Respiratory Confounds

Another reason for the challenging interpretation and small effects around RMSSD may be to do with a confound that was not included in the original experimental design, in the form of the respiratory system. Previous work such as that of Quintana & Heathers (Heathers; Quintana & Heathers, 2014) have shown that among the vast complexities of designing robust HRV experiments is the acute influence of respiration such as frequency, depth, and respiratory sinus arrhythmia upon heart rate variability. These confounds are often unreported, not mentioned, or generally inadequately managed in games research adopting such measurement techniques. Upon realising the lack of control for such a critical confounding variable, some efforts were exerted to try and extract respiratory information using the gyroscopic accelerator within the *Polar* device, based on previous published work attempting to infer respiratory system from similar hardware (Fekr et al., 2014). These efforts did not yield adequately informative signals on respiration, not to mention the fact that no strain gauge based measurement of respiration was acquired to compare as a baseline for the accelerometer based estimation performance. It is likely that substantial sources of noise such as the large variability in movement of participants during play. Thus, it is hard to justify the use of accelerometer based signals on respiration as a surrogate. A much more tenable solution would simply be to include the measurement of respiration in the form of a strain gauge respiration sensor or a similar device in another experiment following this one. In the same vein that commercial ECG hardware like the *Polar* device may have utility in research, a similar chest strap-like device with an embedded strain gauge designed for the user′s comfort might also be appropriate for games research.

It should also be noted, that there was a good justification to not have used a strain gauge in this study. One of the priorities when planning this experiment was to minimise instrument related discomforts and subsequent disruptions to the participants' playing experience. A strain gauge would have acted as an additional device worn on the skin of the participants' torso, which may have led to re-introducing such problems, especially with the potential of the resistance based measurements causing more noticeable discomfort. Additionally, the act of putting the strain gauge on would also have potentially required further interruptions during the experiment from the investigator if measurement errors were observed due to the incorrect wearing of the device. All of that is to say that the logistical challenges involved in using this additional measurement apparatus were considered to be substantial. Unfortunately, it has become evident from the findings of this study that the strain gauge is an unavoidable element of a heart rate variability based experiment, and the resource planning for any psychophysiology based games experiment using HRV as a measurement should account for the deployment of a strain gauge in addition to the pre-existing apparatus.

Apparatus Discomfort

Based on the previous *Don't Starve* study as well as the experiment discussed here, there appears to be a potentially antagonistic effect of discomfort caused by wearing signal apparatus on the immersive experience of a player. In fact, by the point that the design of this present experiment was being planned, minimisation of apparatus discomfort was already a factor (which subsequently led to the lack of a strain gauge). However, upon inspection there does not appear to be an empirical investigation that has explored the deleterious effects of wearing measurement devices on the player's experience — neither pertaining to immersion specifically, or player engagement in general. Therefore there is an impetus for future research to explore the nature of this effect (if any exists at all), as games experiments including the one in this chapter have been taking into account user comfort, despite the lack of any strong empirical evidence to motivate such considerations.

The Efficacy of the NASA-TLX

Despite its brevity and potentially limited single question per factor structure, the NASA-TLX was able to detect appreciable experimental effects in the analysis between the two gameplay difficulty conditions. The rapidity at which participants were able to complete both parts of the NASA-TLX was also noteworthy, with implications that suggest the possibility of measuring immersion with granularity via means other than physiological signals.

Qualitative observations noted that participants were able to complete the entire NASA-TLX in such a short period of time that a similarly short questionnaire designed to capture immersion might be likely to achieve similar measurement power despite the short form format. Such a short-form immersion questionnaire could be a potential alternative to measuring player experience in a more granular fashion, as an alternative to psychophysiology driven attempts at measuring experience.

A theoretical counterpoint could be made in that the use of the full NASA-TLX including the additional weighting forced choice task was a factor in the performance of the NASA-TLX in this study. However, the results of the NASA-TLX have shown that a strong effect would have been observed irrespective of whether weighted or unweighted TLX scores were used. Furthermore, even in the case that the NASA-TLX in its extended form with the forced choice task was considered to be a necessary requirement to capture a good measurement, the questionnaire would still be considerably shorter than the Immersive Experience Questionnaire, for reference. Even with the most generous of assumptions against the NASA-TLX, the demonstration of the short form questionnaire's capability to measure an interpretable signal at all is quite compelling.

Finally, it is worth considering the role of such a test in the context of the wider subject of this thesis of exploring the granular measurement of player experience. The NASA-TLX, by merit of being a rapid psychometric test, achieved granular measurements at lower costs, training, and fewer hardware requirements than the signal modality used in this study. Even if the HRV measurements in this experiment were more ideally reflective of participants' mental workloads, the question has to be asked where there is necessarily a lesser degree of information captured by the NASA-TLX in such a case. So far, the first two experiments of this thesis have only explored the aggregate level measurement of experience achievable by physiological signals, and the clearest signal difference captured in analysis was that of the largest difficulty differences. Meanwhile, this same difference was also captured by the NASA-TLX, demonstrating that the psychometric test was able to capture a comparable amount of information as the physiological signal. Of course, a signal might theoretically allow for even more granular measurements than the epoch intervals in which the NASA-TLX was answered, but this reality has yet to materialise so far in either this thesis or the wider literature at the time of writing. Based on these observations, it appears that the short form questionnaire may prove to be the more economically viable approach to a granular measurement of immersion At the very least, the results of this study have therefore provided the basis to potentially explore the question of whether a short form Immersive Experience Questionnaire may be so quick to complete that interruptions to an immersive state would be minimal.

4.10 Conclusion

This study involved an experiment using a carefully controlled video game, with stimulus curated to specifically allow for optimal conditions for obtaining physiological measurements in the form of HRV. The goals of testing the viability of the NASA-TLX and short form questionnaires in general were met with strong observed effects in NASA-TLX scores between different difficulties of the game. However, the effect of HRV change across different task loads was less pronounced, with an effect only observed between the hardest difficulty and the remaining epochs. These results suggest that particularly pre-existing strong effects might be required in order to induce any considerable contrasts in HRV, but further experimentation is required in order to confirm these suspicions. Further research is also required to ensure more careful controlling of confounds, including the measurement of respiratory behaviour which can contribute considerably to the activity of the heart.

Chapter 5

The Short Form IEQ

5.1 Foreword

As will become evident, this chapter represents a shift in the approach taken in the research of this thesis. Unfortunately, due to the COVID-19 pandemic, work had to be cut short during data collection of the previous chapter. Furthermore, the impact of lockdowns around the pandemic meant that no further in person lab work could be conducted to collect additional data using any form of physiological measurement devices.

Therefore, going forward, an approach to research that was possible through remote and online methods had to be taken. The results of the work done during this period are presented in the following chapter.

5.2 Introduction

So far, most player experience experiments using psychophysiological signals to measure affect have utilised some form of aggregation of the time-series data in order to deal with dimensionality scaling issues, including the two previous studies in this thesis. This raises the question: if a signal is only providing data in aggregate across experimental trial blocks, is there any additional information they are providing over a questionnaire applied multiple times in the same trial block? Additionally, assuming that both a physiological signal and a traditional questionnaire are capable of providing at least some information about a player's state of immersion, could a more time efficient questionnaire be developed to provide comparable information about a player's immersive experience across an experiment? If this is true, then such a tool may provide at least some of the same advantages as psychophysiological measurement techniques in measurement granularity, at a lower cost of access along with reduced statistical and methodological complexity.

As shown in the rhythm game study, the ability of participants to quickly answer the NASA-TLX over a sequence of multiple experimental trials with carefully designed interruptions would indicate that a questionnaire which is small enough to complete in a brief period of time could also be minimally interrupting while measuring the potentially delicate state of immersion. The full IEQ questionnaire has 31 items a participant needs to respond to in order to obtain the IEQ score. However, if the aim is to measure changes in immersion throughout a player's session, asking participants to fill out the 31-item questionnaire multiple times becomes prohibitive. In addition, filling out the longer IEQ questionnaire with interrupted play is more likely to break immersion and skew the measurements.

Here first steps are taken towards measuring immersion with a granular psychometric approach by developing such a tool, in the form of an Immersive Experience Questionnaire Short Form (IEQ-SF). The aim of this study is to miniaturise the IEQ into a form factor that can be completed in approximately one minute while retaining the construct of the original IEQ. Tied to this objective is necessity to validate the IEQ-SF as a satisfactory replacement for the IEQ in terms of its latent structure and power. In order to do so, the present study will test the IEQ-SF against a battery of previously acquired experimental results obtained using the IEQ as well as a novel dataset acquired solely using the IEQ-SF.

5.3 Overview

The essential challenge to developing the IEQ-SF was defining which subset of questions should be selected to constitute a short form. Such a short form had to meet two requirements: to be as full a representation of the breadth of questions of the full IEQ as possible, and to meet some optimality criteria such that there was minimal information loss from reducing the 31 item IEQ into a questionnaire less than half its original size. The process of forming this new IEQ-SF was separated into four separate studies, outlined in this section.

First, for the initial design and development of the IEQ-SF, a large survey sample dataset collected as part of a masters thesis on immersion by Perrett was used (Perrett, 2018). At this starting phase of development, the survey data was used as a basis to propose a candidate short form, resulting in an initial unifactor IEQ-SF which was proposed as a candidate short form. This short form then underwent a series of validation analyses by attempting to replicate previously published results from a series of experiments by Cutting (Cutting et al., 2020; Cutting, 2018) and Denisova (Denisova, 2016). These replications however, failed to yield adequately comparable results.

Thus, the second stage of this study involved an alternative approach to formulating a candidate IEQ-SF, through a multivariate item response theory factor analysis. This resulted in a set of multivariate candidate IEQ-SF models, which were compared to one another with a battery of replicability analyses using the same experimental data by Cutting (Cutting et al., 2020; Cutting, 2018) and Denisova (Denisova, 2016), as with the first stage attempt at an IEQ-SF. These new replicability analyses involved slight alterations to the way that immersion is scored due to the multivariate nature of the test. Therefore, additional care was taken to confirm construct validity by comparing to a new and accordingly similar version of the full IEQ. From the results of the replicability analyses, a single multivariate IEQ-SF candidate was selected for the two remaining validation studies.

The third stage study was a pre-registered analysis that sought to confirm the validity of the IEQ-SF by replicating previously published experimental results (Cutting et al., 2020). The results indicated that the final candidate of the multivariate IEQ-SF was capable of replicating previously reported results with some minor loss of statistical power, however every stage of validation to this point had been conducted with already collected data that was recorded with a full IEQ, resulting in uncertainty around the possibility of a confounding influence from simply subsampling previously collected data.

Therefore, for the final stage study, entirely new experimental data was collected using the design of Cutting et al (2019) experiment 2 (Cutting et al., 2020), in a pre-registered experiment conducted through the internet. In this study, 160 new participants were asked to complete a replication of a previous experiment by Cutting et al that was shown to have had a strong likelihood of observing a real effect. Participants were asked to complete only the IEQ-SF with no additional questions recorded from the IEQ than the short form. This data was then analysed with the new multivariate approach outlined at the proposal of the multivariate IEQ-SF and the results presented strong conclusive evidence of a working short form.

5.4 Success Criteria

In order to develop a systematic framework for the development of a IEQ-SF, a baseline criteria for what could be considered a successful short form is first established. From a survey of existing literature in HCI and Psychology, it is probable that there is not a consensus on an established or canonical evaluation strategy for selecting a short form. Ultimately, there is likely a degree of subjective judgement in the act of developing a miniaturised questionnaire from an existing one, partly because the focus is on a latent conceptual space that is inherently challenging to define, but also because there is a degree of subjectivity involved in these definitions. Therefore, there is a need to define systematic parameters to deal with this subjective judgement.

The first criterion is a requirement that any IEQ-SF factors must match the original factor structure of the IEQ. Given that the structure of the IEQ was a consequence of the original conceptual definition of immersion, any subsequent IEQ short form that does not adhere to this structure runs the risk of diverging from the initial target of measurement and may instead measure a different latent concept from immersion. Since there is inherent uncertainty in whether one is truly measuring the underlying concept at any given time, any IEQ-SF that does not match this factor structure only increases this uncertainty. Thus, there is some reliance on the fact that if the outcome matches the initial factor structure, it is more likely to be measuring something close to, if not, the original immersion factor.

The newly defined factors of the IEQ-SF should be internally reliable. As with any questionnaire, it was important for the new short form questionnaire to be consistent. However, because of the additional context in that the IEQ-SF was being developed with respects to a pre-existing questionnaire, it was also possible to define how much information loss there was by downsizing to a smaller questionnaire. In this respect, the interesting part is not just the internal reliability of the IEQ-SF, but also how much it is reliable relative to the original IEQ. Here, this was quantified with the Cronbach Alpha α, which was calculated in order to compare internal consistency against the full IEQ. As the main measure of internal reliability, an inadequately large enough α coefficient would imply that the new scale is not internally reliable. In addition to this measurement of scale consistency, correlation coefficients were also computed between the IEQ-SF scores and the remainder IEQ scores as another broad description of the representativeness of the IEQ-SF. To be exhaustively clear, the remainder IEQ was defined as a set such that all elements used in the IEQ-SF were mutually exclusive, and did not exist within the remainder set.

Finally, a successful short form would be capable of producing similar results to those previously published in experiments using the IEQ-SF. If a previous experiment produced a statistically significant result with an effect size, the new IEQ-SF should be similar to those original values. Naturally, there is an expectation that because the task is to abbreviate the original IEQ, there will be some loss in information captured by the new questionnaire, and therefore any analyses ends up being likely to report a smaller effect. With this in consideration, a baseline expectation is set such that any previously statistically significant result should remain statistically significant, and effect sizes should be no less than half of those originally reported.

These criteria were held consistent and used across every study at all stages of developing the IEQ-SF, including the multi-dimensional scale where some additional contextual judgements had to be made regarding the additional statistics involved in testing multiple factors simultaneously. In general, so long as a proposed IEQ-SF questionnaire was able to meet all of these success criteria, it would be considered fit for the purposes of measuring immersion.

5.5 Phase I. Study: The Unidimensional IEQ-SF

5.5.1 Data & Pre-processing

Provided Data

All data used in this first phase study were gathered from previous research that had used the IEQ to measure immersion, and all such experiments held immersion as the primary dependent variable of interest. Data were provided by the original authors of each respective study (Cutting et al., 2020; Cutting, 2018; Denisova, 2016; Perrett, 2018).

These data were provided in their final, clean pre-processed and tabular forms by their original authors. Therefore, care has been taken here to declare how the data were originally acquired, and how unsuitable data was determined to be removed.

Declarations of Relationships with Researchers

Relations of note were that Perrett was a masters student previously involved in the same lab, at a time before the start of this PhD research however. No prior relationship between myself and Perrett exists to be declared. Denisova and Cutting were PhD students that completed before this experiment, and their supervisors were the same as one of the supervisors for this thesis: Paul Cairns. All data used here by these authors were provided either directly by themselves, or by their previous supervisor Paul Cairns.

Main dataset description

The data used for the development stage of the IEQ-SF was a large sample acquired through an online survey by Perrett (Perrett, 2018). This data was collected in an online study that administered the IEQ to video game players in a retrospective study that asked participants to answer the questionnaire with respects to the last game that they played.

Participants

This survey contained responses from 5453 participants, of which 3539 were included in the final dataset as valid responses. These responses were all collected by recruiting volunteer respondents from online communities such as Reddit, as described in Perrett's thesis Perrett, 2018.

The selection criteria for respondents included the exclusion of responses with 4 or more unanswered questions, any responses that repeated the same multiple choice selection across the survey, any open responses found to be overtly joking such as "Cheeki Breeki" (reference to a popular internet meme) as an answer for gender, and any participants found to be under the age of consent for the study were also removed.

Demographic Data

Included in addition to their IEQ scores were respondents' playing preferences and approximate hours played, which were explored to rule out any confounding variables in the sample. Participants were asked to respond to the IEQ with respects to the most recent game that they had played, with a total of 669 different games recorded. Sample gender comprised of 2950 (83.36%) male, 512 (14.7%) female, 66 (1.86%) identifying as "other", and 11 (0.31%) non-responses. In total, participant nationalities comprised of 91 different countries, the majority of whom responded from 4 countries: 1664 (47%) from the USA, 403 (11%) from the United Kingdom, 217 (6%) from Canada, and 157 (4%) from Germany.

Additional Data used for Validation

In order to validate the IEQ-SF as a research tool, replication re-analyses of previously published experimental results were conducted. Previously published IEQ experimental results were acquired, consisting of a mixture of experimental designs previously analysed with two tailed t-tests and various ANOVA models. Additional effort was also concerted to select existing studies with data that had statistically significant experimental results as the primary concern was loss of statistical power due to miniaturisation.

For the first stage of validation, the aim was to replicate results from Cutting et al. 2019 ((Cutting et al., 2020), experiments 6 and 7 from Alena Denisova's PhD thesis (Denisova, 2016), and experiment 3 from Joe Cutting's PhD thesis (Cutting, 2018). As with the data provided by Perrett, the data here was provided in tabular form after already having been pre-processed by their original authors, as described in the cited theses and papers.

Details of the corresponding statistical tests for each of these data are provided in table 5.3.

5.5.2 Materials

The main dataset used for the development of the IEQ-SF was acquired through the online survey platform Qualtrics, and distributed through online communities on Steam and Reddit.

Statistical analyses and data processing were completed with the R programming language. Factor analyses were carried out with the *mirt* library (Chalmers, 2012), and psychometric statistics with the psych package (Revelle, 2020). For replication and validation, factorial ANOVA post-hoc analyses were carried out with the agricolae package (Mendiburu & Simon, 2015).

5.5.3 Multidimensional Item Response Theory Factor Analysis with *mirt*

It is worth briefly discussing why multidimensional item response theory with *mirt* was the chosen tool to attempt a short form IEQ. Item response theory is a rich area of psychometrics that accounts for the need to separate information of respondents from information of tests, which is something that classical test theory approaches (which factor analyses approaches like PCA would fall into) cannot do (Thomas, 2011). That is to say, item response theory is capable of treating two items in the same survey differently from one another by modelling with and for additional parameters. Otherwise, just like traditional factor analyses, the aim of the approach is to enable the measurement of some intangible or *latent* property, and the IRT approach assumes that such a latent construct exists to be measured. In the case of this chapter, that latent construct of course refers to immersion.

The challenge however, has historically been that the method was not previously applicable to questionnaires like the IEQ. This was due to the fact that the IEQ and questionnaires of its ilk consist of polytomous likert response items. To complicate matters further, multidimensionality in factor structure also leads to more complicated model specifications that unidimensional IRT was not suitable for, leading to a need for a multidimensional application of IRT. For some time, multidimensional IRT was considered to be capable of improving measurement precision (Thomas, 2011), and there was a push for the adoption of such approaches in order to reap these benefits (Borsboom, 2006).

Until more recent developments such as that of the *mirt* library, the Bayesian numerical optimisation methods required to estimate the parameters of a multidimensional item response model have had prohibitive computational costs (Chalmers, 2012). The availability of *mirt* has already provided the tools to use multidimensional item response theory for the development of player experience questionnaires, such as the Player Uncertainty in Games Scale (PUGS) (Power et al., 2018).

The approach taken with *mirt* in this chapter was similar to that used in the development of the PUGS. The suitability of items are first inspected by considering the information (in the form of factor loadings, for example), before also considering the semantic information of that item. Therefore, for all intents and purposes, the reading of factor structures would be similar to that of a traditional PCA. Then, when a factor structure is chosen, the most suitable items are also considered on the basis of what they semantically might mean.

5.5.4 Procedure Overview

The following sections in this study will describe the attempt to formulate a unidimensional IEQ-SF. The protocol for how these steps took place is described here for clarity and transparency on how these steps were planned, and how they transpired.

The first step of the planned procedure was to conduct an exploratory factor analysis. The purpose of this was to compute item loadings of a unidimensional IEQ. These item factor loadings were then used for step II of the protocol.

The second step of the planned procedure was to select the items for a candidate IEQ-SF. The item selection procedure was a combination of selecting items with the highest factor loadings, and semantic information of what items meant. The latter was incorporated in order to avoid redundancies, such as repeated items. While this might typically improve the scale reliability in a traditional questionnaire such as the original IEQ, the limited space for items here meant that breadth of conceptual coverage was assumed to be more likely to improve scale performance. At the end of the second step, when items had been chosen and a candidate short form was put forward, a second, confirmatory factor analysis was conducted. This confirmatory factor analysis was performed with only the items chosen for the candidate IEQ-SF, and the purpose of this confirmatory analysis was to examine of item factor loadings had unacceptable deteriorations in the absence of the remaining items from the full IEQ.

The third step of the procedure was to validate the candidate IEQ-SF by conducting a series of re-analyses. These re-analyses were identical to previously performed statistical analyses in other studies that had used the full

IEQ. The objective therefore, was to see if resulting analyses using the IEQ-SF produced similar enough results that it could be considered a reasonable estimation of the full IEQ.

Finally, a fourth prospective step was planned for the procedure. This potential step was to plan for the extension of the candidate IEQ-SF, if any cases were to arise where the initial candidate IEQ-SF was not performing as necessary. An example of such an outcome may be a situation where re-analyses produced radically different results from previously published findings that used the full IEQ.

In general, it was difficult to determine an a-priori threshold for what an acceptable result of re-analysis would be. Therefore, the re-analyses conducted in this chapter were to be considered qualitatively, so that a full pre-registered study could be conducted if the candidate IEQ-SF was considered adequately performant.

5.5.5 Procedure Step I - Exploratory Factor Analysis

An a-priori perspective on the factor structure

In order to determine the parameters around these requirements, the first task was to define the structure of the full IEQ. The IEQ has had over a decade of time being deployed and tested in the wild, with over 1600 citations at the time of writing this. A small but considerable portion of this was research carried out by the authors of the questionnaire themselves. Over this period in time, some observations have been noted on the factor structure initially laid out by Jennett et al. (Jennett et al., 2008). First, it appeared that the originally defined five factor structure was not always across different experiments, particularly with respect to the Challenge dimension. This problem was acknowledged and subsequent off-shoot questionnaires have been developed and published by collaborators of the IEQ to address these shortcomings as well as to further define Challenge as a measurement concept (Denisova et al., 2020). Similar concerns were also applicable for the Control factor.

Taken into a broader consideration for the work done here, an assumption of a single factor structure could also be made on the principle of simplicity, in that if the goal is to measure immersion as its whole, then treating it as a simple, single factor would be the easiest place to start. This assumption prescribes no value to any likelihood of whether a single factor structure is actually most valid, and indeed the successive studies in this chapter will illuminate the nature of the IEQ's factor structure further, beyond the unidimensional attempt here.

Therefore, based on these accumulating concerns and the established practice of how the IEQ was scored, an assumption was made that the factor structure of the IEQ was unidimensional for the purposes of an exploratory factor analysis.

To prepare for factor analysis, the dataset was split randomly into an exploratory half and a validation half in a manner similar to train-test splitting.

Only the exploratory half was used for the exploratory factor analyses carried out for the initial exploration of the factor structure, and the selection of IEQ-SF items. The validation half was used later for the sole purpose of conducting the confirmatory item response theory (IRT) factor analysis, which is detailed at the end of this section on factor analysis.

Exploratory Factor Analysis Protocol

This exploratory factor analysis was carried out using a maximum a posteriori factor analysis, with one defined factor containing all 31 IEQ items using the MIRT default graded model structure. This was estimated with the Metropolis-Hastings Robbins-Monro implementation included in the package.

As a result of conducting a factor analysis where all 31 items were considered to be a single factor, the factor that was extracted represented the whole of the IEQ. In other words, the single factor discussed below is a factor representing overall immersion as a single concept.

5.5.6 Factor Analysis Results

The resulting item-factor loadings from the exploratory factor analyses are presented in table 5.1. The model converged after 50 iterations, with an average acceptance ratio of 0.403. Because rotation is involved, the sum of squared loadings are reported in favour of the proportion of variance explained, and here $SS_{loadings} = 7.54$. In general, the average item loading with the unidimensional IEQ-SF was $M = 0.466$ with a variance of $SD = 0.028$, indicating a minimal spread.

The resulting factor loadings reported in table 5.1 were then used for the selection of items for a candidate short form in the next section, 5.5.7.

5.5.7 Procedure Step II - Item Selection

Item loadings from the exploratory factor analysis were interpreted based on a mix of quantitative and qualitative considerations to select the items comprising the new IEQ-SF. For this task, the primary concern was to fulfil the defined success criteria for a suitable IEQ-SF, and to minimise the length of the IEQ-SF as much as possible such that it could be completed in a relatively quick manner.

Initially, the items with the greatest factor loadings were chosen in a shortlist, with a nominal exclusion threshold at the .55 correlation coefficient. This threshold was chosen so as to allow us to evaluate the upper third quantile of resulting loadings, with 9 items loading above .55 and 22 items loading below this threshold.

Then, items were compared with one another to minimise conceptual redundancies. For instance, the items "The game had my full attention" (item 1) and "I felt focused on the game" (item 2) were determined to be too similar as both items addressed the concept of cognitive involvement, and both

TABLE 5.1: IEQ Factor loadings per item, under a unidimensional factor structure.

Item	Content	Loading Coef. (Exploratory)
1	The game had my full attention	0.72
2	I felt focused on the game	0.76
3	I put effort into playing the game	0.62
4	I tried my best	0.54
5	I lost track of time	0.52
6	I felt consciously aware of being in the real world whilst playing	0.38
7	I forgot about my everyday concerns	0.53
8	I was very much aware of myself in my surroundings	0.25
9	I noticed events taking place around me	0.14
10	I felt the urge to stop playing and see what was happening around me	0.29
11	I felt like I was interacting with the game environment	0.54
12	I felt that I was separated from the real-world environment	0.54
13	The game was something that I was experiencing, rather than just doing	0.66
14	The sense of being in the game environment was stronger than the sense of being in the real world	0.56
15	I found myself so involved that I was unaware I was using controls	0.48
16	I moved through the game according to my own will	0.35
17	I found the game challenging	0.36
18	There were times in the game in which I just wanted to give up	0.10
19	I felt motivated when playing the game	0.73
20	I found the game easy	0.26
21	I felt that I was making progress towards the end of the game	0.41
22	I performed well in the game	0.29
23	I felt emotionally attached to the game	0.56
24	I was interested in seeing how the game's events would progress	0.53
25	I wanted to "win" the game	0.34
26	I felt in suspense about whether or not I would do well in the game	0.33
27	I found myself so involved that I wanted to speak to the game directly	0.47
28	I enjoyed the graphics and the imagery	0.44
29	I enjoyed playing the game	0.67
30	When I stopped playing, I was disappointed that the game is over	0.48
31	I would like to play the game again	0.59

specifically referred to the amount of focus or attention drawn to the game. Consequently, from this real example, only item 2 was chosen for the short form. Further, in cases where competing items were compared, consideration for the participant's interpretation was also a factor in deciding which item to keep. In the above example, item 2 was also selected because of the non-specificity of the interpretation, whereas item 1 specifies "attention" and therefore might have been more liable to lead to signal loss once placed in the context of a short form questionnaire with only a few items.

As an extension of the aim to minimise redundancies, an additional aim was to select as varied an item set as possible from the original factor structure. This might appear contradictory to assumption that the latent space was univariate. However, it is noted that the purpose of the assumption of a univariate structure was not so much to strictly define the original factor structure before selecting a subset for the IEQ-SF. Rather, it was to avoid Cattel's bloated specific (Boyle, 1991) where focusing too strictly on item loadings in the past may have led to distorting factors that may not necessarily be distinct from one another, such as Cognitive Involvement and Emotional Involvement possibly occupying the same latent space at the same point in their respective factor hierarchy (i.e. both are simply facets of immersion, rather than sub-facets distinct from one another). It was considered entirely plausible that the IEQ was indeed not univariate, and simply chose this assumption as a starting point. The risk of a bloated specific was especially relevant in the context of formulating a short questionnaire, wheres the effects of confounding an over-specified factor with general immersion would be greater.

Confirmatory Factor Analysis of candidate unidimensional IEQ-SF

Once the item pool was defined, a confirmatory factor analysis (CFA) was conducted on the confirmatory split of the data using an item response theory confirmatory factor analysis. The item response model converged after 69 iterations. The average acceptance ratio was 0.405. Because there is rotation involved in computation of the factor structure, the sum of squared loadings ($SS_{loadings}$) is reported in favour of the proportion of variance, and here $SS_{loadings} = 2.255$. A problem arises when inspecting the factor loadings directly however. A notable deviation of the loading on item 12 was observed, with a substantially lower confirmatory loading coefficient of .33. This may have possibly been an outlier loading value for that particular factor solution, and that the proceeding validation stage of the study would help in confirming or refuting this item as a suitable choice.

All steps taken to this point as described in the factor analysis and item selection procedures produced a short form comprised of 6 items of which the specific contents, exploratory, and confirmatory loading coefficients are detailed in in the IEQ-SF outline table 5.2.

TABLE 5.2: IEQ-SF Item Pool. Note that the Item column corresponds to original item indices on the full IEQ.

Item	Content	Loading Coef. (Exploratory)	Loading Coef. (Confirmatory)
2	I felt focused on the game.	0.76	0.71
12	I felt that I was separated from the real-world environment.	0.54	0.33
13	The game was something that I was experiencing, rather than just doing.	0.66	0.65
19	I felt motivated when playing the game.	0.73	0.75
23	I felt emotionally attached to the game.	0.56	0.52
31	I would like to play the game again.	0.59	0.62

5.5.8 Procedure Step III - Validation with replication by reanalyses

With a IEQ-SF now defined, a validation of the questionnaire was required, to ensure that it was still measuring immersion as previously accomplished by the full IEQ. In order to conduct this validation, replication analyses were carried out in order to match previous findings from replication datasets, the details of which are included in the results table 5.3. Since the IEQ-SF was assumed to be a unidimensional measure like the full IEQ, all statistical analyses were conducted with the exact same procedures as their originally published methods.

This involved the calculation of a single mean immersion score as the main measure of interest. The implication here bears repeating– in this study, immersion is being treated as a single concept with a single factor structure, for the purposes of a simplest-approach-first attempt at a short form. Subscale analysis in the form of mean scores for each of the factors was not conducted as part of this re-analysis procedure. The primary reason for this was simply the fact that some of the sub-scales did not exist anymore.

Following the calculation of IEQ-SF scores for each of the studies, analyses procedures were then replicated as closely as possible to their originally reported procedures in their original publications. For example, Cutting et al performed a series of t-tests for each of their experiments (Cutting et al., 2020), and therefore, t-tests would also be conducted with the IEQ-SF. Additional care was also taken to ensure that the same sum of square methods and interpretation of the same effect size statistics were used, as in their original studies.

The purpose of these re-analyses were to compare the results, to explore the power of the IEQ-SF in detecting the same effects as the original IEQ. If the IEQ-SF is an adequate representation of the original IEQ, then the p values and detected effect sizes should in theory, be close to to the original IEQ.

Because it was difficult to determine a-priori what to expect from this procedure, a threshold for an effect size ratio was not determined. Instead, the results were to be considered qualitatively before the candidate short form was to be applied in a proper, pre-registered validation study.

Results from the replication analyses using the IEQ-SF are presented in table 5.3. Note that there are some discrepancies between the reported results here and those from their original publications, which were usually due to some minor tabulation or calculation errors which were found while corroborating with previous authors and these differences were mainly caused by differences in how statistical packages counted observations with a missing observation, which were observed to be negligible. Here, differences were recorded between *p* values associated with each analysis, and more importantly, the comparative effect sizes between the IEQ and the new IEQ-SF.

Additionally, the newly calculated *p* values were not calculated with any corrections for multiple comparisons. The reason for this was an existing expectation that the IEQ-SF may face issues with power and sensitivity, and therefore any corrections might hinder its performance. More importantly, the data originally collected were sourced from studies published independently of one another, insofar as the sample data were concerned. Since the aim of this procedure was to conduct analyses as closely as possible to their original publications, no corrections were applied unless stated otherwise in their originally reported results.

As defined in the success criteria, metrics were also computed for internal consistency in the form of the Cronbach α, as well as the short-form and remainder score correlation.

5.5.9 Validation Results

The majority previous results were replicated with the IEQ-SF which suggested that the IEQ-SF is able to fall approximately close to results produced by the full IEQ. Among the studies that were statistically significant in their original publications, Denisova (2018) experiment 7 yielded comparable effect sizes for both levels of the factorial design, and Cutting et al. (2018) experiment 3 also had a comparable effect size with what was originally reported.

However, if only significant results were to be considered, approximately half of these analyses could have been considered to be failures to replicate. In particular, there were large deviations from the original results produced in experiment 1 of Cutting et Al. (2020) (Cutting et al., 2020) and experiment 6 of Denisova (2018) (Denisova, 2016) where effect sizes appeared to more than halve, and results that were previously well within the threshold of statistical significance were no longer within the acceptance range.

For assessment of internal consistency, the estimated Cronbach α was .67 (95% CI = [0.64, 0.69], IEQ score M = 3.8, $SD = 0.56$). The Cronbach α of the full IEQ was calculated, with the split used for the exploratory factor analysis as a point of comparison. This produced an α of .76 (IEQ M = 3.6, $SD = 0.34$; α 95% CI = [0.74, 0.77]). As defined in the success criteria, correlations were

TABLE 5.3: Comparison of replication tests with the IEQ-SF against results obtained with the original IEQ. The test column informs what type of test was run, where the test chosen is aligned with the test from the original study. $p(IEQ)$ refers to the p value originally published, and $p(IEQ - SF)$ refers to the p value produced using the IEQ-SF candidate. Similarly, $EffectSize(IEQ)$ refers to the originally published effect size, while $EffectSize(IEQ - SF)$ refers to the effect size produced by the candidate IEQ-SF.

Study	Test	N	p (IEQ)	p (IEQ-SF)	Effect Size (IEQ)	Effect Size (IEQ-SF)
Cutting et Al. (2020) Experiment 1	t test	35	0.0384	0.503	$d = 0.729$	$d = 0.229$
Cutting et Al. (2020) Experiment 3	t test	40	0.396	0.295	$d = -0.271$	$d = 0.336$
Cutting et Al. (2020) Experiment 4	t test	157	0.569	0.503	$d = -0.09$	$d = 0.107$
Cutting (2018) Experiment 3	One-way ANOVA	48	0.017	0.024	$\eta^2 = 0.17$	$\eta^2 = 0.15$
Denisova (2018) Experiment 6	One-way ANOVA	120	< 0.001, 0.008	< 0.001, 0.078	$\eta_p^2 = 0.16$, $\eta_p^2 = 0.08$	$\eta_p^2 = 0.11$, $\eta_p^2 = 0.04$
Denisova (2018) Experiment 7	One-way ANOVA	60	0.003, 0.008	0.009, 0.005	$\eta_p^2 = 0.14$, $\eta_p^2 = 0.12$	$\eta_p^2 = 0.11$, $\eta_p^2 = 0.13$

calculated between IEQ-SF scores and scores from the remaining items in the full IEQ after removing those used for the IEQ-SF. A Pearson's correlation produced a statistically significant correlation ($p < .001, t = 37.302$) with a moderately strong positive correlation between the two scores ($r = 0.678$, 95% CI = [0.652, 0.704]).

5.5.10 Validation Interpretation

In the case of experiment 1 from Cutting et al. (2020), there was a possibility that the effect originally reported result was simply a false positive, which is always a possibility under the framework of null hypothesis testing. However, if it were the case that the original experiment was simply a false positive, it would follow that this result should, in theory, also produce a false positive result in line with the result produced by the full IEQ, with a comparable (if slightly smaller) effect size. This was not the case, so instead it was considered that this failure to replicate was more likely to be due to the possibility that a small set of questions were responsible for such a large proportion of the effect detected by the IEQ that their removal had all but eliminated this signal. If this were the case, it would imply that the original structure of the IEQ was likely biased by a subset of its items for this particular experiment.

First, it was possible that there was inadequate sampling from the original IEQ in what may have been a case of a particularly unfavourably biased experiment. To address this, a second series of replication analyses were run to explore this possibility. On the subject of an inadequate item pool, guidance provided by Kline (Kline, 2000, 2014) outlined issues with small questionnaires, and suggested (very generally) that a 10 item scale would be optimal. While empirical evidence since then has shown that small scale questionnaires with fewer than 10 items have been considerably successful in their measurement goals, the IEQ-SF may not necessarily fit the criteria necessary for success. Tools like the NASA-TLX have demonstrated that under the right circumstances, it is possible to build a small psychometric tool that is still capable of capturing information consistently (Hart, 2006).

In the context of a small item pool, another reason for the failure to replicate could also be the arguably low factor loadings of certain items chosen for the IEQ-SF. For instance, Item 12 was selected for qualitative properties despite its lower loading coefficient of .54. While subjective interpretation of the questionnaire items from the researcher's perspective may provide justification to the value of a particular item, the lower loading coefficients might support the suggestion that there are simply an inadequate number of items if one also desires a greater breadth of coverage.

Most importantly, it appeared that the tool being made developed here was less statistically powerful than initially expected. Combined with a selection process that involved subjective assessments of the item contents, the use of previous data as a basis to validate through replication may have been inherently liable to introduce bias to any resulting IEQ-SF that was developed. To deal with these potential biases, the options were to either introduce further experimental data to the replication pool, or add more items to the item pool in an effort to avoid overfitting the IEQ-SF to a small sample of experimental results. In this case, the chosen approach was to first experiment with a new variant of the IEQ-SF which was extended to explore the degree of difference made by adding further items in the hopes that what was observed may have just been larger variances that could be reduced through this method.

5.5.11 Procedure Step IV - Extending the IEQ-SF Item Pool

Following the first round of validation, additional items were added to the IEQ-SF pool to test whether the small size of the questionnaire was the reason for its poor performance in replicating previous results. Items were selected using the same process as in the previously described procedure, and a subsequent confirmatory factor analysis was conducted again with the extended item pool.

Worryingly, a poor confirmatory factor loading was produced again, this time for item 7. However, given the exploratory nature of this extended IEQ-SF, the decision was made to pursue this a little further to see what might happen. The expectation was that another failure to replicate prior results using an extended IEQ-SF would indicate that there were likely to be deeper issues than simply the item pool size in the short form questionnaire. The subsequent IEQ-SF model is reported in table 5.4, where 2 additional items were added to the IEQ-SF to form an 8 item questionnaire that was chosen for another attempt to validate previous results.

5.5.12 Extension Results

Validation results using the extended IEQ-SF are reported in table 5.5. This table is initially presented without a direct comparison for the first IEQ-SF for the reason that the motivating origin for the IEQ-SF was to develop a tool that performed comparably to the original IEQ.

In general, a mixture of differences in p values provide a picture of an outcome that does not fully represent a good short form of the original IEQ. These differences therefore, require discussion in order to decide the next step.

5.5.13 Discussion of Study I

Based on the two sets of replication results from using the univariate IEQ-SF and the extended variant of this IEQ-SF, it was concluded that the currently proposed univariate IEQ-SF did not hold up under scrutiny, especially when focusing on experimental power using solely statistically significant results from existing analyses. Neither variants of the proposed IEQ-SF candidates

TABLE 5.4: 8 Item IEQ-SF Pool. Note that the Item column corresponds to original item indices on the full IEQ, and new items added to the previous iteration of the IEQ-SF are separated at the bottom.

Item	Content	Loading Coef. (Exploratory)	Loading Coef. (Confirmatory)
2	I felt focused on the game.	0.76	0.70
12	I felt that I was separated from the real-world environment.	0.54	0.33
13	The game was something that I was experiencing, rather than just doing.	0.66	0.65
19	I felt motivated when playing the game.	0.73	0.75
23	I felt emotionally attached to the game.	0.56	0.52
31	I would like to play the game again.	0.59	0.62
7	I forgot about my everyday concerns.	0.53	0.39
24	I was interested in seeing how the game's events would progress.	0.53	0.56

TABLE 5.5: Comparison of replication tests with the newly extended IEQ-SF against results obtained with the original IEQ.

Study	Test	N	p (IEQ)	p (IEQ-SF)	Effect Size (IEQ)	Effect Size (IEQ-SF)
Cutting et Al. (2020) Experiment 1	t test	35	0.038	0.77	$d = 0.729$	$d = 0.099$
Cutting et Al. (2020) Experiment 3	t test	40	0.396	0.619	$d = -0.271$	$d = -0.159$
Cutting et Al. (2020) Experiment 4	t test	157	0.569	0.187	$d = -0.09$	$d = -0.212$
Cutting (2018) Experiment 3	One-way ANOVA	48	0.017	0.024	$\eta^2 = 0.17$	$\eta^2 = 0.15$
Denisova (2018) Experiment 6	One-way ANOVA	120	< 0.001, 0.008	< 0.001, 0.068	$\eta_p^2 = 0.16$, $\eta_p^2 = 0.08$	$\eta_p^2 = 0.12$, $\eta_p^2 = 0.05$
Denisova (2018) Experiment 7	One-way ANOVA	60	0.003, 0.008	0.002, 0.001	$\eta_p^2 = 0.15$, $\eta_p^2 = 0.17$	$\eta_p^2 = 0.11$, $\eta_p^2 = 0.13$

could produce convincingly convergent effect sizes to values that were originally reported, and there was a failure to meet or confirm the assumption made regarding the possibility of a small item pool being the cause of failures to replicate.

Upon inspecting the full array of results again, it was noted that in the exploratory factor analysis, the factor loadings across the full univariate IEQ seemed to have a relatively low average loading ($M = 0.47$, $SD = 0.028$) with only a third of items loading greater than .55. This was considered to be a consequence of squashing the entire item pool of the full IEQ onto a single dimension, which was a indication that the assumption assumption in a univariate factor structure may also not have been met.

If the items chosen for this IEQ-SF are also considered semantically, it is clear that even if challenge may be a sub-factor that has varying amounts of prominence in IEQ scores in previous studies, it still may be a critical enough component of the overall immersive experience that a challenge item ought to be included in any short form. The lack of a single challenge item chosen in the IEQ-SF candidate in this study may have contributed to the overall poor replication of previous analyses of the full IEQ.

One consideration to improve the candidate questionnaire was to try to develop a more robust and systematic process by which one could iteratively, or even exhaustively, test any number of combinations of subsets drawn by sampling items from the full IEQ and running validation analyses for each produced model. It was decided not to do this, as this approach included several issues including a significant risk of over-fitting to the large dataset from a single study, as well as being too focused on internal consistency and high item loadings.

Instead, the approach chosen was to explore the possibility that a deeper problem caused the failures to replicate previously reported results. One of the core assumptions of the approach thus far was that the IEQ was a unidimensional scale. Even though there was reasonable justification for the basis of this assumption outlined in section 5.5.5, it was also acknowledged to be a cursory assumption to be used as a starting point, and it may be the case that this assumption may have been incorrect. However, having now also noted observations that the previously reported factor structure of the IEQ may not entirely be robust, it would be unwise to simply return to the original factor structure without exploring this possibility. To explore this possibility, a new factor analyses was carried out to explore the dimensionality of the original IEQ as well as the structure of its latent space without regarding any previously held beliefs from the original IEQ publication.

5.5.14 Conclusion of Study I

All results suggested that the IEQ was not likely to exist on a unidimensional domain, and that subsequent exploration of the latent space was required to formulate a more appropriate structure for the short form.

5.6 Phase II. Study: The Multidimensional IEQ-SF

Taking a different approach to formulating the IEQ-SF, in this study the aim was to re-examine the factor structure of the original IEQ in order to inform the item selection process for the short form. As with the first phase study, there was a focus on a mixture of factor analyses results and qualitative assessments of item contents. Because of the additional dimensionality adding complexity to the questionnaires, in this study multiple candidate short forms were compared before proceeding with a single chosen candidate questionnaire to conduct validation analyses on.

5.6.1 Participants

The same dataset used for the generation of the IEQ-SF in the previous section was used again here for exploring the factor structure of the IEQ. By using the same dataset as in the univariate attempt, there would be a clear point of reference to the results of the previous attempts at developing the IEQ-SF. In order to protect against overfitting like in the previous factor analysis, the data was split into exploratory and validation halves, and the same seed was used as the previous univariate study for the pseudo-random number generator splitting the data. The exploratory data analysis was conducted on the first half of the split and the confirmatory factor analyses were conducted on the second half.

Also similar to the prior section, the same experimental data was used as from the first attempt, in Cutting et al. 2019 ((Cutting et al., 2020), Denisova 2016, (Denisova, 2016), and and cutting 2018 (Cutting, 2018). Details of these data were provided in 5.3.

5.6.2 Materials

As with previous factor analyses, the mirt package was used for structure exploration (Chalmers, 2012), in the R programming language. The agricolae (Mendiburu & Simon, 2015), and psych (Revelle, 2020) packages were used to conduct statistical evaluations for scale consistency as well as validation analyses.

5.6.3 Factor Analysis Procedure

Due to the difference in assumptions about the number of dimensions in the original publication of the IEQ compared to newer research, the approach taken here to explore the factor structure aimed to cover a breadth of different possible factor structures. Five exploratory factor analyses were carried out, with each analysis having a number of assumed factors ranging from 1 to 5. Factor loadings for each of these models were then evaluated in a similar fashion with the previous univariate study.

TABLE 5.6: Multidimensional exploratory factor analysis results. The numbers in the column headers refer firstly to the number of specified factors, and then to the specific factor within each model, e.g. the column 4F-F2 refers to Factor 2 of the 4 Factor model. Within a model, rows in bold represent cross-loaded factors. Loading coefficients are highlighted in red or green based on direction of loading past thresholds of .35 and -0.35. A more readable, full page variant of this table can be found in the appendix D.

	1F-F1	2F-F1	2F-F2	3F-F1	3F-F2	3F-F3	4F-F1	4F-F2	4F-F3	4F-F4	5F-F1	5F-F2	5F-F3	5F-F4	5F-F5
IEQ1	0.72	-0.58	-0.30	-0.50	-0.32	0.17	**-0.56**	**-0.35**	**0.11**	**-0.02**	**-0.23**	**-0.35**	**0.09**	**-0.01**	**0.41**
IEQ2	0.76	-0.73	-0.13	-0.65	-0.16	0.18	-0.72	-0.20	0.12	0.00	-0.28	-0.21	0.08	0.02	0.54
IEQ3	0.62	-0.68	0.02	-0.58	0.01	0.27	-0.63	-0.01	0.22	0.01	-0.13	0.00	0.10	-0.06	0.61
IEQ4	0.54	-0.66	0.11	-0.56	0.11	0.27	-0.66	0.06	0.20	0.07	-0.13	0.07	0.11	-0.01	0.61
IEQ5	0.52	-0.29	-0.42	-0.25	-0.44	0.07	-0.13	-0.37	0.09	-0.26	-0.17	-0.36	0.06	-0.22	0.07
IEQ6	0.38	0.07	-0.80	0.09	-0.79	-0.01	0.03	-0.79	-0.03	-0.03	0.03	-0.80	0.00	0.00	-0.04
IEQ7	0.53	-0.30	-0.43	-0.28	-0.46	0.00	-0.23	-0.42	-0.01	-0.16	-0.24	-0.44	0.02	-0.05	0.08
IEQ8	0.25	0.18	-0.73	0.21	-0.71	0.04	0.08	-0.74	0.01	0.08	0.12	-0.75	0.06	0.06	-0.07
IEQ9	0.14	0.22	-0.59	0.24	-0.56	0.08	0.00	-0.65	0.02	0.28	0.22	-0.65	0.04	0.17	0.04
IEQ10	0.29	-0.14	-0.29	-0.13	-0.29	0.00	-0.26	-0.34	-0.05	0.13	-0.12	-0.39	0.02	0.21	0.11
IEQ11	0.54	-0.48	-0.16	-0.52	-0.22	-0.15	-0.16	-0.08	-0.08	-0.54	-0.59	-0.12	-0.01	-0.22	-0.13
IEQ12	0.54	-0.17	-0.68	-0.15	-0.70	-0.02	-0.06	-0.64	0.00	-0.23	-0.08	-0.63	-0.04	-0.20	0.08
IEQ13	0.66	-0.59	-0.17	-0.57	-0.21	0.01	-0.27	-0.09	0.06	-0.50	-0.57	-0.11	0.10	-0.24	-0.01
IEQ14	0.56	-0.22	-0.62	-0.21	-0.66	-0.08	-0.05	-0.56	-0.04	-0.34	-0.16	-0.55	-0.08	-0.27	0.03
IEQ15	0.48	-0.21	-0.49	-0.19	-0.51	0.00	-0.01	-0.42	0.04	-0.33	-0.12	-0.40	-0.02	-0.29	0.03
IEQ16	0.35	-0.43	0.09	-0.48	0.06	-0.12	-0.33	0.10	-0.11	-0.22	-0.46	0.05	-0.03	0.03	0.05
IEQ17	0.36	-0.38	0.00	-0.13	0.07	0.83	-0.19	0.06	0.80	0.04	-0.06	0.03	0.79	0.03	0.16
IEQ18	0.10	-0.15	0.05	-0.32	0.02	-0.47	**-0.37**	**-0.03**	**-0.51**	**0.08**	**-0.47**	**-0.12**	**-0.31**	**0.39**	**-0.06**
IEQ19	0.73	-0.79	0.00	-0.75	-0.04	0.09	-0.62	-0.01	0.08	-0.23	**-0.46**	**-0.03**	**0.06**	**-0.09**	**0.38**
IEQ20	0.26	-0.21	-0.08	0.04	-0.02	0.83	0.01	-0.01	0.83	0.00	-0.06	-0.04	0.89	0.03	-0.06
IEQ21	0.41	-0.52	0.11	-0.52	0.07	-0.03	-0.41	0.09	-0.04	-0.17	-0.30	0.08	-0.06	-0.07	0.27
IEQ22	0.29	-0.41	0.13	-0.53	0.09	-0.34	**-0.58**	**0.03**	**-0.40**	**0.07**	**-0.17**	**0.05**	**-0.45**	**0.06**	**0.48**
IEQ23	0.56	-0.47	-0.20	-0.46	-0.24	-0.02	-0.17	-0.13	0.02	-0.46	**-0.35**	**-0.11**	**-0.03**	**-0.36**	**0.07**
IEQ24	0.53	-0.58	0.03	-0.55	-0.02	0.04	-0.28	0.09	0.08	-0.43	-0.51	0.07	0.10	-0.19	0.04
IEQ25	0.34	-0.34	-0.03	-0.21	-0.02	0.37	-0.35	-0.07	0.32	0.14	0.28	-0.01	0.12	-0.12	0.63
IEQ26	0.33	-0.25	-0.15	-0.06	-0.12	0.54	-0.01	-0.08	0.56	-0.12	0.15	-0.04	0.40	-0.26	0.23
IEQ27	0.47	-0.31	-0.31	-0.27	-0.33	0.04	0.06	-0.18	0.11	-0.53	-0.09	-0.11	-0.06	-0.62	0.07
IEQ28	0.44	-0.53	0.08	-0.57	0.05	-0.10	-0.36	0.11	-0.08	-0.32	-0.56	0.07	-0.01	-0.04	0.05
IEQ29	0.67	-0.83	0.16	-0.85	0.12	-0.05	-0.71	0.15	-0.06	-0.24	-0.76	0.09	0.06	0.09	0.20
IEQ30	0.48	-0.34	-0.27	-0.31	-0.28	0.07	-0.20	-0.26	0.06	-0.22	-0.21	-0.27	0.06	-0.17	0.09
IEQ31	0.59	-0.66	-0.04	-0.64	-0.07	0.05	-0.52	-0.02	0.02	-0.24	-0.53	-0.08	0.11	-0.04	0.18

Because the goal was to inspect whether or not a specific factor structure was feasible, closer attention was paid to cross loadings in order to assist in the comparison between each factor model. In addition, subjective assessments of the coherence of item collections in each factor structure were made, and these are reported in the results. Lastly, the scree plot of eigenvalue loadings of a principal components analysis were generated, to provide additional information on the number of factors in the latent structure.

5.6.4 Exploratory Factor Analysis Results

Full results of the exploratory factor analysis for each model and corresponding factors are presented in table 5.6, along with labelled instances of cross-loadings and loadings past a minimum threshold (-0.35 and .35). Here, cross-loadings simply refer to instances where one item is loaded beyond the specified threshhold into more than one factor, which indicates that either the item or factor structure is inadequately specified for the latent concept.

Cross-loadings were observed in the 4 and 5 factor models, with 3 cross-loaded items in 4F and 5 cross-loaded items in 5F. Specifically, items 1, 18, and 22 in model 4F, and in addition to these, items 19, and 23 in model 5F. Notably, 4 of the 5 cross-loaded items in model 5F were loaded in opposite directions (signs). No cross-loadings were observed for models 1F, 2F or 3F, although several borderline cases were observed for the same set of items in models 2F and 3F which indicated that these items were likely to be causes of construct noise rather than the factor structures themselves.

5.6.5 Interpretation

From the results of the exploratory factor analysis, the initial viable candidates were chosen on the basis of factor loadings alone. Each factor model was interpreted one by one which revealed key shortcomings of the 5F model, before then comparing differences between the remaining candidates. Immediately, it was found that based on the clearer and more informative factor structures with the 2F, 3F, and 4F models, as well as results from the previous section attempting to build a unidimensional IEQ-SF, 1F was ruled out as a viable candidate for the IEQ-SF.

The most notable quality of the 5F model were the number of cross-loadings, which was the highest among all the models. While it is somewhat expected that the number of cross-loaded items would increase with the number of factors in a given finite pool of items, the cross-loadings still indicate that these items were contributing to multiple factors, often in antagonistic ways between factors in the particular case of model 5F. This inherently makes it difficult to interpret the meaning of each factor, and in the particular case of model 5F, it was observed that the within-factor item pools did not match the originally reported 5 factor structure of the IEQ as it was originally published. Finally, it was observed that some factors in model 5F were remarkably small, for instance 5F-F3 consisted of 4 items, and 5F-F4 only consisted of 3 items of which 2 were cross-loaded which suggested it was redundant. Based on these observations, the model 5F was eliminated from the list of viable candidates for formulating an IEQ-SF.

While model 4F also contained a number of cross-loaded items, the same items were borderline cross-loaded cases in the 2F and 3F models, with coefficients just barely below the specified threshold for defining cross-loadings. Therefore, value was instead drawn from the item pool contents. In comparison to model 3F, the factor structure of model 4F actually almost matched 3F identically with the core difference being that the large factor 3F-F1 was divided in half within 4F. When comparing with the original factor structure of the original IEQ, it was shown that 4F-F4 had a mixture of items from 4 factors from the original IEQ. This made it quite challenging to pinpoint a conceptual definition for the 4F model.

In comparison, 3F was a bit clearer to interpret with subjective judgements. It appeared to us that model 3F was similar to the original IEQ's factor structure with 2 large, but quite clear and obvious differences. First, it appeared that the Cognitive Involvement and Emotional Involvement factors of the original IEQ had merged into one in model 3F. As noted above, 3F-F3 was also identical to 4F-F3, and when compared to the original IEQ structure, this factor was shown to be the Challenge subscale. The 3F-F2 factor was then matched back to the original Real World Dissociation factor from the original IEQ, and this factor was also found to be almost identical in model 4F.

Lastly, model 2F appeared to be less informative than both models 3F and 4F as it was observed that items from 3F-F3 and 4F-F3 simply did not load as strongly in model 2F, with some items not loading at all, such as items 18, 20, and 26 all containing loading coefficients lower than .25. From this it was concluded that model 2F just simply did not contain any information about Challenge as a construct. This gave us a justification on which model 2F could be rejected as insufficiently informative to use as a basis for building an IEQ-SF.

Therefore, the key discerning difference between models 3F and 4F were whether or not a conceptual definition for 4F-F4 could be made, which provided additional information to the questionnaire. With model 3F, there would be a general "Involvement" factor that was a combination of the original Cognitive and Emotional involvement IEQ factors. Alternatively, an approximate reformulation of the two Involvement factors would be found in choosing 4F as the basis for a multidimensional IEQ. One consideration made here were the practical implications of these options for the purposes of a IEQ-SF. One could make the case that both 3F and 4F were valid structures for the IEQ, however,in going with a 4 factor structure, there consequently be a need to build a slightly larger item pool for the IEQ-SF, in order to match the power of each individual factor with those from a model with fewer factors. On the other hand, there would also be an issue if the decision was made to use 4F over 3F without sampling an equal number of items from each factor. In this case, there would be a risk of reduced information within each factor in the 4F short form. Given the risks and trade-offs involved in building a short form questionnaire, it was determined that a slightly less informative single factor with more detection power and scale reliability across the questionnaire was more desirable, and thus concluded that a 3 factor IEQ was the most suitable model to work with as a basis for a short form.

5.6.6 Confirmatory Factor Analysis

Having chosen a 3 dimensional factor structure as the basis for a newly defined IEQ structure from which to build a short form, the next step was to produce a confirmatory factor analysis. Similar to the first study, this confirmatory factor analysis was calculated using the second half split of the data.

As a reminder, note that the IEQ score represents the overall immersion score, Involvement refers to the nature in which players feel engrossed within the game, real world dissociation (RWD) is the factor that captures the experience of disconnecting from the real world and losing track of time, and challenge is the factor that captures the difficulty and task demands of playing a game.

In addition to a multidimensional item response theory confirmatory factor analysis, a confirmatory bi-factor analysis was also conducted, which provided additional information in the form of a h^2 communality value. The h^2 value provides information on how much each item also correspondingly relates to the main latent concept of interest, such as immersion. For this bi-factor analysis, items were assigned to their respective factors on the basis of greatest factor loading coefficients, in addition to being assigned to a general latent IEQ construct. No secondary factors were defined, and h^2 communality coefficients were recorded in addition to loading coefficients in the bi-factor confirmatory factor loadings. These results are reported in table 5.7.

The confirmatory factor analysis did not raise any alarming results for any items in the chosen 3 factor structure, and smaller coefficients were considered to provide additional guidance on which items were more appropriate to be used in a IEQ-SF, during the short form selection stage of this study.

5.6.7 Primary Multi-factor Candidate

After choosing to work with the 3F model, selection criteria were defined to determine the pool of items to be used in this iteration of the short form. Again, a combined qualitative and quantitative approach was used to determine the most suitable items. As with the item pool selection process these decisions were not clear or simple, so several additional candidates were nominated as viable short forms, and the subjective judgements that led to sticking with the main candidate were documented for discussion.

The largest of these candidates serves as a benchmark to compare the performance of each short form model with, and is referred to as the Benchmark-11 IEQ-SF. As per its name, Benchmark-11 is comprised of 11 items which were approximately sampled equally from each of the 3 factors of the IEQ. This was done in order to maintain the factor structure of the IEQ based on the exploratory factor analysis. Thus, the Benchmark-11 was originally a 12 item questionnaire but validity concerns with question 18 in particular did not allow us to keep it in the item pool due.

During the process of evaluating the fitness metrics for the Benchmark-11, it was noted that the performance of the questionnaire fluctuated a lot more significantly when Item 18 was present in the questionnaire than compared to when it was excluded. Item 18 contained the question "There were times in the game in which I just wanted to give up.", which was a member of the Challenge factor.

While 11 items is larger than the 6 or 8 item reductions of the IEQ in section 5.5.5, results showed that it was still a significant enough reduction that it achieved the primary goal of forming an IEQ short form to be a brief enough questionnaire that it could be repeatedly applied over an experiment. At a cursory glance, it can be seen that Item 18 did have loading coefficients as strong as the other items in the Challenge factor except for Item 25. This

TABLE 5.7: Multidimensional item response theory confirmatory factor analysis results for a 3 factor defined structure. Note that IEQ stands for the overall Immersive Experience Questionnaire score, RWD stands for the Real World Dissociation factor. The involvement, RWD, and Challenge factors are all based on their original factors in the full IEQ. h^2 communality coefficients are reported with respects to within defined factors combined with the general IEQ construct. Coefficients not displayed are presumed to be 0 as items were only assigned one factor in the model definition.

	IEQ	Involvement	RWD	Challenge	h^2
IEQ1	0.63	0.23			0.45
IEQ2	0.67	0.36			0.57
IEQ3	0.51	0.43			0.45
IEQ4	0.48	0.41			0.40
IEQ5	0.53		0.24		0.34
IEQ6	-0.26		-0.75		0.63
IEQ7	0.46		0.36		0.34
IEQ8	-0.05		-0.74		0.55
IEQ9	0.03		-0.66		0.44
IEQ10	-0.20		-0.25		0.10
IEQ11	0.48	0.30			0.32
IEQ12	0.52		0.54		0.55
IEQ13	0.60	0.27			0.43
IEQ14	0.56		0.43		0.49
IEQ15	0.46		0.28		0.30
IEQ16	0.20	0.43			0.23
IEQ17	0.29			0.78	0.70
IEQ18	-0.04			0.35	0.13
IEQ19	0.53	0.52			0.56
IEQ20	-0.14			-0.77	0.61
IEQ21	0.26	0.40			0.22
IEQ22	0.17	0.40			0.19
IEQ23	0.51	0.16			0.29
IEQ24	0.42	0.35			0.30
IEQ25	0.25			0.27	0.14
IEQ26	0.36			0.47	0.35
IEQ27	0.53				0.28
IEQ28	0.33	0.49			0.35
IEQ29	0.40	0.78			0.76
IEQ30	0.48				0.23
IEQ31	0.36	0.65			0.55

TABLE 5.8: Results comparing effect sizes between Challenge short form scores with, and without, question 11.

Study	Test	Variable	p	d
Cutting (2018) Exp. 3	3 Items	Difficulty	0.00009	0.339
	4 Items	Difficulty	0.0002	0.315
Denisova (2018) Exp. 6	3 Items	Information	0.005	0.735
	3 Items	Adaptation	0.008	0.349
	4 Items	Information	0.001	0.941
	4 Items	Adaptation	0.009	0.289
Denisova (2018) Exp. 7	3 Items	Information	0.017	0.097
	3 Items	Adaptation	0.012	0.108
	4 Items	Information	0.033	0.078
	4 Items	Adaptation	0.002	0.156

alone would not have been problematic, and it was indeed kept in as part of the 4 items selected from the Challenge pool for the IEQ-SF. However, it was found that Cronbach α coefficients were quite low for a 4 item Challenge pool when compared to a 3 item Challenge pool without question 18 ($\alpha = -0.18[-0.26, -0.1]$ with item 18, $\alpha = -0.8, [-0.93, -0.66]$ without item 18), which was a greater cause for concern with the Challenge subscale. This difference in Cronbach's α was also visualised in figure 5.2. Finally, during the validation stages of analysis later in this study, concerns were further exacerbated by the fact that effect sizes appeared to fluctuate, sometimes quite radically. To demonstrate this, results were selected from three experiments which specifically involved manipulations surrounding Challenge and difficulty, and the results of these specific comparisons are presented in table 5.8. While acknowledging that p values aren't necessarily meaningful when read absolutely without context, and d estimates typically involve a high degree of variance, when these statistical observations were considered alongside the Cronbach's a impact of item 18, as well as the factor loading, additional efforts were required to determine whether to keep the question or not.

Given the lack of conclusive evidence from a quantitative assessment of item 18, the approach was taken to assess the meaning of the content of the question. In this, one interpretation was that the question lacked definitive clarity, which may have been a reason behind the variation introduced by the item found in the statistics. Players could have given up for a variety of reasons other than challenge, which would explain the low factor loading. Furthermore, it was also noted that the item might not actually have a linear relationship with the latent construct at all. If one considers the definition of flow which requires an optimal balance within challenge to reach an optimal state (Csikszentmihalyi, 1990), it is plausible that the (mathematical) process of "giving up" in playing a game is not strictly linear. One possibility considered was a piece-wise function with an increase in immersion as challenge increases, until the end of the scale at which immersion experiences a steep decline. Meaning that as players experience greater and greater difficulties, their immersion may in fact be increasing, but past an optimal threshold, any further increases of challenge may actually be penalising players' immersions. Of course, this was simply a hypothesis but the thought exercise permitted the conclusion that there was now an impasse where there was no longer confidence that the item was measuring what it was designed to measure. When combined with the statistical instability and especially with the α values without the item, it was finally decided that item 18 was to be removed from the questionnaire.

Following the selection of Benchmark-11's item pool, two confirmatory factor analyses were conducted. First, a multidimensional item response theory CFA was computed, with the three specified factors and their corresponding component items. Following this, a Bi-Factor analysis was computed in order to assess h^2 communality. Benchmark-11, its item pool, and factor loadings from the exploratory and confirmatory factor analyses are detailed in table 5.9. Acknowledging some expected variation in loading coefficients between the exploratory and confirmatory factor analyses, no large enough differences were observed that led to the belief that the chosen model was inappropriate. Therefore, these were deemed to be satisfactory and so a subsequent evaluation was made of the semantic meaning of this proposed model.

The new model represented a smaller proportion of the original factors of the IEQ. Arguments have been made earlier in this section on why the Control factor was not included in this short form. Similarly, the two forms of involvement have now been compressed into a single factor. Qualitatively, an argument can be made that immersion is still being represented across most of its originally conceived factors. Information on a player's involvement (in both forms), real world dissociation, and experience of challenge are still being captured. Qualitatively evaluating these factors' items case by case also leads to the suggestion that they still represent their original form.

Involvement still contained an item around a player's cognitive engagement with the game: *"I felt focused on the game."*, as well as their emotional involvement with their game: *"I enjoyed playing the game."* The slight skew towards emotional involvement could also be considered to be not too concerning as the other heavily loaded cognitive involvement item appears to be a slightly redundant, and somewhat more absolute variant of the one that was chosen: *"The game had my full attention".* Therefore, capturing more of the breadth of emotional involvement made sense, in hindsight.

The real world dissociation factor was represented by a similarly broad (at least in a relative sense) number of 4 items. These items captured the conscious awareness of the player: "I felt consciously aware of being in the real world whilst playing.", and their disengagement from their life and environment: "I forgot about my everyday concerns; I felt that I was separated from the world-world environment". Although the latter of these two statements seems to be almost tautological, it was also the highest loaded item in this subscale with an exploratory and confirmatory factor analysis loading both

TABLE 5.9: Benchmark-11. Note that the Item column corresponds to original item indices on the full IEQ, with each factor sectioned by a line.

Item	Content	Expl.	Conf.	Bi-Factor	h^2
	Involvement				
2	I felt focused on the game.	0.73	0.69	0.30	0.57
13	The game was something that I was experiencing, rather than just doing.	0.60	0.59	0.28	0.66
19	I felt motivated when playing the game.	0.80	0.81	0.54	0.59
29	I enjoyed playing the game.	0.75	0.76	0.79	0.76
	Real World Dissociation				
6	I felt consciously aware of being in the real world whilst playing.	-0.68	-0.65	-0.65	0.46
7	I forgot about my everyday concerns.	0.53	0.56	0.43	0.34
12	I felt that I was separated from the real-world environment.	0.82	0.82	0.71	0.66
15	I found myself so involved that I was unaware I was using controls	0.58	0.53	0.40	0.31
	Challenge				
17	I found the game challenging.	0.88	0.83	0.76	0.72
20	I found the game easy.	-0.83	-0.77	-0.76	0.59
26	I felt in suspense about whether or not I would do well in the game.	0.52	0.54	0.40	0.31

of 0.82. What is also interesting is that one of the chosen items for real world dissociation actually mentions controls: *"I was unaware I was using controls."*, which further lends credence to the suggestion that control as a factor did not make much sense in the original conception of the IEQ.

The challenge factor also contained similar items phrased in a manner semantically close to the subject of the factor: *"I found the game challenging."* and *"I found the game easy"*. Both of these items loaded more strongly than *"I felt in suspense about whether or not I would do well in the game."* The difficulty (or ease) of the game is something that appears to have been reliably and easily determined by previous participants answering the IEQ, at least in the contexts of whatever games the IEQ was taken with. The suspense and uncertainty of one's performance is also an informative item, least of which because of the fact that uncertainty is now an area of growing interest in immersion research with a scale of its own (Power et al., 2017).

In general, it does appear at least on the surface that the original IEQ is still being largely represented by the new candidate short form. The next step then, was to proceed with formulating alternative variants from the Benchmark-11 IEQ-SF.

5.6.8 Alternative Multi-factor Candidates

It was also considered that a reliable measurement could be obtained from an even smaller reduction of the IEQ- especially given the close statistical output from downsizing Challenge from 4 to 3 items. Thus, from the Benchmark-11 model, several smaller variants were also derived to be compared in order to check this possibility.

The first set of these variants include a 9 item variant titled All-Factors-9 due to the fact that it still follows the 3 factor structure of the IEQ and Benchmark-11 item pools. Similarly, an All-Factors-6 variant was also formed. With all of these smaller variants, items were selected from the Benchmark-11 item pools based solely on their explanatory power as defined by their factor loading coefficients, without any context from subjective item meanings. The decision was made to do this, due to the fact that all items in the stripped down 11-item factor were considered to be conceptually relevant and valid, and so the main metric of interest was how much information they provided to the scale, which was inferred using loading coefficients.

A second set of smaller but more disparate IEQ-SF candidates were also formed from the Benchmark-11 model. The critical difference between these candidates and the previously mentioned Benchmark-11 and All-Factors-X variants are observed in the fact that these candidates do not include the Challenge subscale of the full IEQ. There was instead the additional possibility that challenge itself was entirely an unstable factor, partly due to the aforementioned problems with item 18, but also due to the fact that challenge had already been historically marked as a potentially problematic factor of the IEQ (see previous section 5.5.5). Therefore, the aim was also to test whether removing the 3 items of the challenge subscale from Benchmark-11 might still produce a comparably reliable immersion score. Two short form candidates with this 2-Factor structure were constructed, with an 8 item and a 6 item variant titled RI-8 and RI-6.

Details of each alternative multi-factor short form model are provided in the appendices (appendix item D).

5.6.9 Validation of a Multi-factor Structure

To further substantiate a 3 factor structure in addition to the confirmatory factor analyses, the parallel scree plot of the Benchmark-11 model was interpreted and presented in figure 5.1. A visual inspection of the scree plot would reveal that three factors have significantly larger eigenvalues than the remaining factors. While factors 2 and 3 do not pass the conventional (if arbitrary) line at 1, the elbow of the plot does appear to corroborate with the structure retaining 3 factors.

Furthermore, the Cronbach α was calculated per factor, across variations of those factors that differed in their number of items, for example comparing a 4 item Involvement factor with a 3 item Involvement variation, and a 2 item Involvement Variation. The graph visualising these comparisons is presented in figure 5.2. The resulting plot further supported the suggestion that the Challenge factor was a point of concern, as the α of this factor in its full

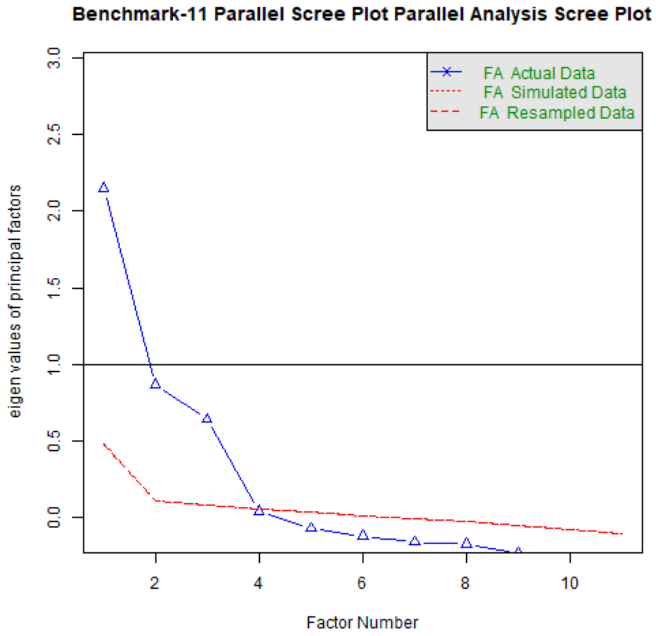


FIGURE 5.1: Scree plot of the Benchmark-11 candidate, based on the second half sample of the dataset.

form is comparatively low (.32) with its full item pool of 4 questions. This further supported the decision to reduce Challenge down to 3 items even in the largest Benchmark candidate for the IEQ SF, where α was in more acceptable regions for 3 items (.67) and 2 items (.72).

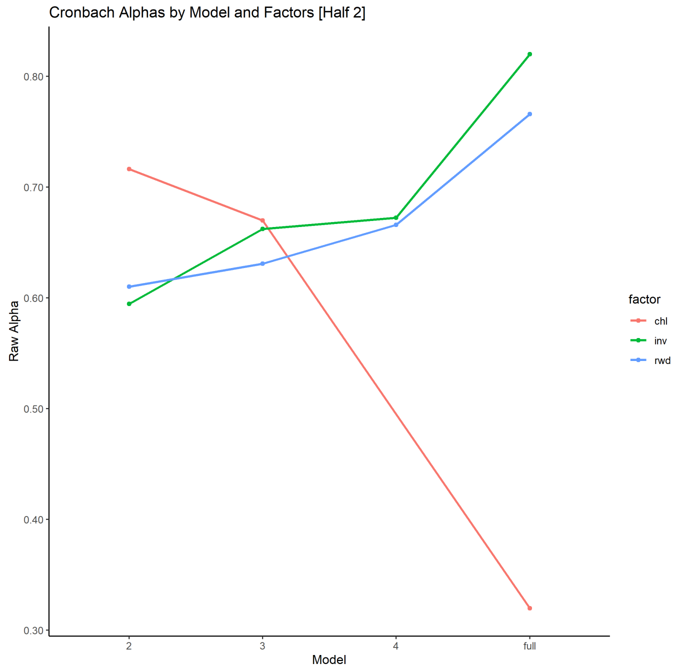


FIGURE 5.2: Graph of Cronbach alphas per factor, across N-Item variations of each factor.

With respects to the Involvement and Real World Dissociation factors, there is a notable steady decline in α as the number of items are reduced, with the largest difference observed between their full number of items and the 4 item variants used in the Benchmark-11 model. This is closer to what would be expected when considering a reduction of the number of items from a typical questionnaire. It also suggests that the relative information provided by individual items are more equitable, although it should be noted that the items with greatest loadings are retained across all variants graphed here.

5.6.10 Validation Results between Candidate Short Forms

To select the fittest of the multi-factor IEQ-SF candidates, the same series of statistical tests were conducted for validation as with the uni-factor IEQ-SF in section 5.5.2. However, as there are now multiple factors in each IEQ-SF candidate, the analysis has been altered to account for this and therefore the proceeding tests can not be strictly considered as replications. Instead, the analyses involved conducting a test for each of the factors independently, meaning that there were 3 analyses and 3 subsequent resulting p values and effect sizes for a given application of the IEQ-SF, whereas there would have simply been one in the past for a single, full IEQ score. This meant that it was no longer a task of simply comparing a previous p value to a newly computed short form p value in order to compare the short form to the original IEQ. Instead, more subjective judgements were required in addition to the typical interpretations of effect sizes.

This also meant that there was no longer a simple one-to-one comparison of these results with the original values reported from the previous studies, though interpretations can still be provided with some caution and care. All tests were conducted with $\alpha = 0.05$ and effect sizes were interpreted in conjunction with significance values with additional care due to the new method of analysis. To accommodate these changes, details of new analyses conducted with the Benchmark-11 candidate have been provided in table 5.10. As previously mentioned, this comparison is not a simple side-by-side of one p value with another, so in lieu of this, additional commentary surrounding each study is provided.

For the comparison of different IEQ-SF candidates, the graphs in figure 5.3 that show comparative effect sizes and p values for each factor in the results for Cutting (2018) Experiment 3, and Denisova (2018) experiment 7 (Cutting, 2018; Denisova, 2016). These particular studies were chosen specifically due to the convenience of their particularly strong effects allowing us to interpret the relative detective power of the different short forms.

In general, there was some variance in the efficacy of different short forms in producing the strongest effect sizes between different studies, and this fluctuation was considered to be consequence of a combination of noise, as well as the particular applicability of specific questions to specific experimental tasks. Therefore, a combination of quantitative and subjective, qualitative assessment of results was employed.

TABLE 5.10: Results for validation analyses conducted with the Benchmark-11 IEQ-SF candidate. Note that for the two experiments in Denisova (2018), the order of results are presented by levels in the factorial design: Adaptation, and then Information. Previous effect sizes correspond to reported effects for full IEQ scores.

Study	Test	N	Factor	Previous Effect Size	p	Effect Size
Cutting et Al. (2020) Exp. 1	t test	35	Involvement	$d = 0.72$	0.413	$d = 0.28$
	t test	35	RWD		0.899	$d = 0.043$
	t test	35	Challenge		0.058	$d = 0.663$
Cutting et Al. (2020) Exp. 3	t test	40	Involvement	$d = -0.27$	0.234	$d = -0.382$
	t test	40	RWD		0.377	$d = 0.282$
	t test	40	Challenge		0.341	$d = -0.305$
Cutting et Al. (2020) Exp. 4	t test	40	Involvement	$d = -0.091$	0.657	$d = -0.071$
	t test	40	RWD		0.1271	$d = -0.245$
	t test	40	Challenge		0.045	$d = 0.072$
Cutting (2018) Experiment 3	One-way ANOVA	48	Involvement	$\eta^2 = 0.17$	0.049	$\eta^2 = 0.125$
	One-way ANOVA	48	RWD		0.119	$\eta^2 = 0.09$
	One-way ANOVA	48	Challenge		<0.001	$\eta^2 = 0.339$
Denisova (2018) Experiment 6	One-way ANOVA	120	Involvement	$\eta_p^2 = 0.16$ $\eta_p^2 = 0.08$	0.002, 0.006	$\eta_p^2 = 0.079$, $\eta_p^2 = 0.083$
	One-way ANOVA	120	RWD		0.036, 0.006	$\eta_p^2 = 0.037$, $\eta_p^2 = 0.085$
	One-way ANOVA	120	Challenge		0.349, 0.735	$\eta_p^2 = 0.008$, $\eta_p^2 = 0.005$
Denisova (2018) Experiment 7	One-way ANOVA	60	Involvement	$\eta_p^2 = 0.14$ $\eta_p^2 = 0.12$	0.017, 0.033	$\eta_p^2 = 0.096$, $\eta_p^2 = 0.077$
	One-way ANOVA	60	RWD		0.306, 0.062	$\eta_p^2 = 0.018$, $\eta_p^2 = 0.059$
	One-way ANOVA	60	Challenge		0.011, 0.016	$\eta_p^2 = 0.108$, $\eta_p^2 = 0.097$

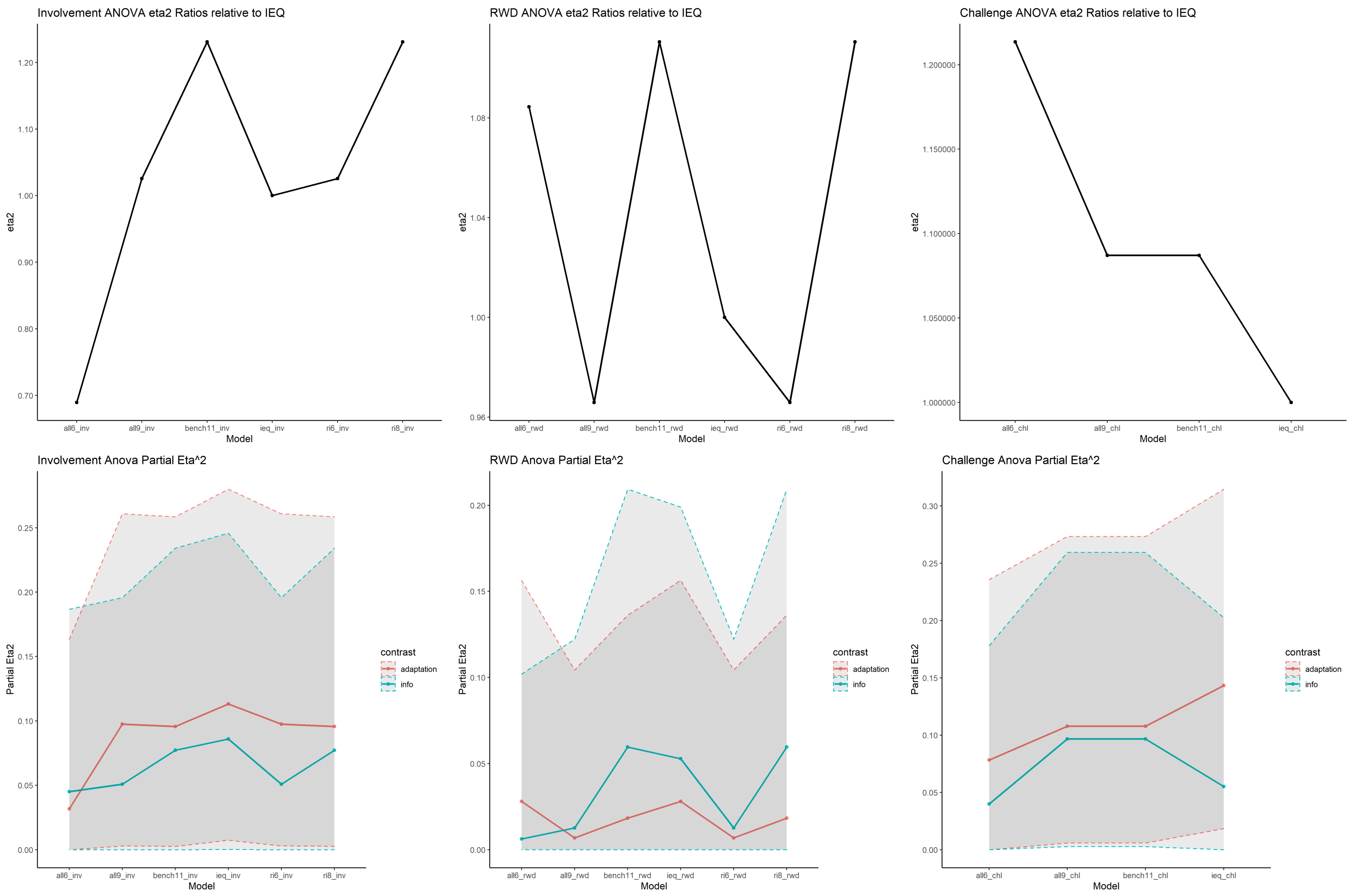


FIGURE 5.3: Raw effect sizes for every candidate multi-factor IEQ-SF.
First row: Cutting (2018) Experiment 3. Second row: Denisova (2018) Experiment 7.

One of the quantitative approaches taken to assess these results was to define a very simple fitness metric for the greatest ratio of an effect size relative to the effect size produced using the original IEQ with the new factor structure. This would approximately standardise across different effect size estimators between the pool of studies, and allow us to select the model which was most probable to detect the strongest effects. Formally, this value is a simple division of a candidate IEQ-SF factor score d_{sf} with the full factor score d_{full}, which produces the effect size ratio D; $D = \frac{d_{sf}}{d_{full}}0$. Then, all D values were summed across the pool of studies to find the candidate item pool with the largest $D0$. The resulting D ratios are tabulated in table 5.11.

Resulting D ratios demonstrated that the absolute effect size ratios were only moderately variant across each model in every study, with an average trimmed variance of approximately $SD = 0.07$ across all factors in every study from the pool. These results provided a different story than the instability visualised by the graphs in figure 5.3 told. Rather, the models performed approximately close to one another, and in the cases of extreme differences such as that observed in Cutting et Al. (2020) Experiment 1, the large ratios were due to the fact that the effects originally produced were near zero, therefore leading to any deviations being relatively large multiplicands of the score produced by the full factor.

The models which were most likely to produce the greatest effect sizes relative to the full IEQ were either Benchmark-11 with an average D of .101, or All-6 with an average D of 1.11. This difference was considered to be negligible in light of the typical variance and noise expected from a psychometric tool.

The internal reliability of the two most viable candidates were lower than expected, with a Cronbach $\alpha = 0.45$ for Benchmark-11 and $\alpha = -0.31$ for All-Factors-6, when computed using the full dataset that was used in the factor analyses stages of this study. However, when computing alpha on a per-factor basis, the results were more in line with expectations and values considered to be internally consistent, with Benchmark-11 producing $\alpha = 0.68, 0.45, and - 0.73$ for Involvement, Real World Dissociation, and Challenge, respectively. The supplementary measurement of consistency also showed that the scores produced by the short form were strongly correlated with the remaining items from the full IEQ with a Pearson's $r = 0.870$.

When comparing the relative effect sizes produced by each candidate, it was worth reminding that the full IEQ itself was likely to be a measurement tool which was not sensitive enough to suitably capture the kinds of granular changes that were of interest, particularly with the approach of computing relative effect size ratios. Even without this brutish approach of comparing effect size ratios, the wide confidence intervals shown in figure 5.3 and published in the studies used in the pool would suggest that the IEQ did not produce particularly discriminating measurements. Therefore, given the noisy nature of psychometric tools and the primitive means of quantifying detective powers used here, it was deemed that the most viable candidate could not be determined in these statistics alone. Thus, other considerations among the defined criteria for a successful short form were evaluated.

TABLE 5.11: Effect sizes produced by each candidate short form, expressed relative to full scores for each factor as a ratio.

Study	Factor	All-9	All-6	RI-8	RI-6	Benchmark-11
Cutting et Al. (2020) Exp. 1	Inv.	0.93	0.79	0.40	0.93	0.40
	RWD	-11.47	-3.86	-1.49	-11.47	-1.49
	Chal.	0.73	0.99			0.73
Cutting et Al. (2020) Exp. 3	Inv.	1.58	1.14	1.34	1.58	1.34
	RWD	1.38	1.34	2.37	1.38	2.37
	Chal.	0.82	0.90			0.82
Cutting et Al. (2020) Exp. 4	Inv.	0.56	-0.11	0.56	0.56	0.56
	RWD	1.30	0.49	2.11	1.30	2.11
	Chal.	1.31	1.54			1.31
Cutting (2018) Exp. 3	Inv.	1.03	0.69	1.23	1.03	1.23
	RWD	0.97	1.08	1.11	0.97	1.11
	Chal.	1.09	1.21			1.09
Denisova (2018) Exp. 6	Inv.	0.72	0.65	0.54	0.72	0.54
		1.24	0.93	1.51	1.24	1.51
	RWD	0.43	0.84	0.76	0.43	0.76
		0.83	0.29	0.95	0.83	0.95
	Chal.	0.27	0.24			0.27
		2.54	14.21			2.54
Denisova (2018) Exp. 7	Inv.	0.86	0.28	0.85	0.86	0.85
		0.59	0.52	0.90	0.59	0.90
	RWD	0.25	1.00	0.65	0.25	0.65
		0.24	0.12	1.13	0.24	1.13
	Chal.	0.75	0.55			0.75
		1.75	0.72			1.75

Beyond the quantitative approach of comparing effect sizes, one obvious, vital difference between the Benchmark-11 and All-Factors-6 candidates were the number of items in each short form. A shorter questionnaire is more desirable in that they are simply quicker to complete, fulfilling one of the core purposes of building a short form at all. However, it should also be considered that a shorter questionnaire has a smaller margin of error. If items are not responded to correctly, the accuracy of the resulting measurement suffers immensely. Given this trade-off between speed of application and scale reliability, the safer option was considered to be Benchmark-11, given its larger item pool. Additional justification for choosing Benchmark-11 was also the fact that once acclimated to the questionnaire, data from the experiment in the previous chapter using the NASA-TLX had shown that participants took less than half a minute on average to complete the first part of the TLX which contained 6 items, much like the All-Factors-6 short form. If an assumption was made that the time taken to complete a questionnaire was a linear function of the number of items, then the Benchmark-11 questionnaire would take a minute to complete. This still fulfilled the specified success criteria for speed of application.

5.6.11 Discussion of Study II

Results from exploratory factor analyses revealed that the latent structure of the IEQ was multidimensional, most likely with 3 factors: Involvement, Real World Dissociation, and Challenge, which was supported by confirmatory item response theory and bi-factor factor analyses. This meant, therefore, that the initial concerns that the IEQ was not actually univariate were justified.

The primary candidate short form was able to produce statistical results that were approximately in line with results previously published under the full IEQ. Where statistically significant effects were found in a unidimensional IEQ score, similar results were produced by the Benchmark-11. Most of the other candidates were also able to produce similar results. This alone was a large improvement upon the first, unidimensional attempt at formulating an IEQ short form.

The lacklustre Cronbach α computed by treating the Benchmark-11 holistically as a single score was initially concerning, but upon inspecting the α of each factor internally, as well as the strong correlation between the short form score with the remaining IEQ, this was interpreted as a natural loss of some internal reliability from downsizing the item pool so much, combined with the complicated structure that arises from having to treat the new questionnaire as a truly multi-dimensional construct.

Based on the results from this comparison of candidate short forms, it was decided that the Benchmark-11 IEQ-SF was the most viable short form, and this variant was therefore chosen to be the newly defined short form to be used in future validation studies.

5.6.12 Conclusion of Study II

The IEQ was confirmed to be multidimensional, based on factor analyses of the latent structure. Of the various candidate short forms nominated, based on the ability to closely reproduce previously published results that used the IEQ, and subjective evaluations of content meaning and relevance, the Benchmark-11 candidate was most the most suitable candidate for an IEQ short form. This was decided due to the still-fast-enough speed of application of an 11 item questionnaire, and the benefit of reduced risk by having a questionnaire almost twice the size of the next most viable candidate.

However, while these results are statistically significant, they often have a reduced effect size compared to their original publications. Given that this was an exploratory analysis to test the applicability of the newly derived IEQ-SF, there was not yet sufficiently convincing evidence of the suitability of the IEQ-SF as measurement tool capable of being used in place of the IEQ. Therefore, the new questionnaire was now ready to be tested and validated in further studies in more rigorous tests.

5.7 Phase III. Study: Pre-Registered Validation Analysis

The third phase study was a validation analysis that was pre-registered on the open science framework with the id bf8qn (Aung et al., 2021a). This study involves the reproduction of previously published results from the second experiment of (Cutting et al., 2020), using the newly defined IEQ-SF. The central central purpose for this study was to provide a validation test with data that had not been involved in any exploratory work during the development process of the IEQ-SF, and was therefore independent not yet contaminated by any potential risks of overfitting to a selected sample.

5.7.1 Participants

The dataset was drawn from Cutting et al (2019) experiment 2. The sample of 180 participants was collected in an online experiment from the *Prolific* platform, in which participants were tasked with playing a session of the game *Two Dots*. The experiment was a between subjects design and participants belonged to either a higher or a lower immersion condition, in which they played a version of the game that would induce the condition's according level of immersive experience. Each condition had exactly 80 participants, and the gender demographics were split between 77 male, 81 female, and 2 non-binary participants between the ages of 18-40 years old.

Following the main experimental task of playing the game, all participants then completed a forced choice test on 30 images for a distractor recognition task not pertinent to the hypotheses of this replicability study. They then completed an Immersion Experience Questionnaire to measure their experience of the game. Full details of the experimental procedure are available in the original publication (Cutting et al., 2020).

5.7.2 Materials

All statistical analyses were conducted with the *R* programming language, as with all previous studies. Cronbach α were calculated with the *psych* package (Revelle, 2020), and graphs were plotted with *ggplot2*.

5.7.3 Design & Hypotheses

The original experiment involved a distractor recognition task that was an independent measure from the IEQ scores. Here, the focus was on the more relevant dependant variable which was the IEQ-SF score, derived from the original IEQ scores that were obtained from participants completing the questionnaire after playing the game. The independent variable was the version of the game that was played, either the high immersion or low immersion condition of the game.

The main hypothesis for the original experiment was that the immersion score would be significantly different between the low and high immersion condition. However, as the new IEQ-SF no longer uses a unitary immersion metric, the hypothesis and therefore the analyses here are not strictly the same. For this reason, the previous hypothesis in question was extrapolated into a hypothesis with multiple outcome measures.

Here, it was hypothesised that there would be a significant difference in level of individual immersion factor scores reported between each game as measured by the IEQ-SF. The high immersion game was hypothesised to have a statistically significantly higher scores for Involvement, Real World Dissociation, and Challenge, as measured by the IEQ-SF.

A second hypothesis was also formed in order to test if the close relationship between the scores of the IEQ-SF with the original IEQ did indeed exist as implicated in the second phase study. Therefore, the hypothesis was that there would be a significant correlation between an immersion score computed using remaining items from the original IEQ, and IEQ-SF scores within each condition. This hypothesis was constructed with correlations in comparison to the full remaining set rather than remaining items on a per-factor basis as some factors were already very small to begin with, and correlations were unlikely to be meaningful given the small remaining item pools (1-2 items in the case of the Challenge subscale). As a supplementary measure, a correlation was computed between the full IEQ and a single immersion score using items from the IEQ-SF.

5.7.4 Analysis Procedure

IEQ-SF scores were calculated using the original data provided by Cutting et al., with one score for each corresponding factor of the new structure.

TABLE 5.12: Results for the reproducibility t-test analyses conducted with Cutting et al. (2020) experiment 2, with effect sizes expressed in Cohen's *d*0. Mean scores (*M*) for each factor are also provided per condition, as well as *SD* standard deviations.

Factor	*p*	*t*	Effect Size	Effect Size CIs	Low Immersion	High Immersion
Involvement	<0.001	-8.38	$d = -1.325$	[-1.670, -0.980]	$M = 3.153$ SD = 0.752	$M = 4.031$ SD = 0.559
RWD	<0.001	-3.977	$d = -0.629$	[-0.949, -0.309]	$M = 2.828$ SD = 0.845	$M = 3.328$ SD = 0.742
Challenge	<0.001	-5.269	$d = -0.833$	[-1.158, -0.508]	$M = 2.471$ SD = 0.967	$M = 3.196$ SD = 0.761

For the first hypothesis, a series of two tailed t-tests were calculated for each factor of the IEQ-SF. Corresponding effect sizes were also computed through Cohen's *d*0.

For the second hypothesis comparing the scores between the full IEQ scores and the IEQ-SF, correlation coefficients were computed for each factor score with a set of items considered to be a remainder from the full IEQ after all IEQ-SF items were removed.

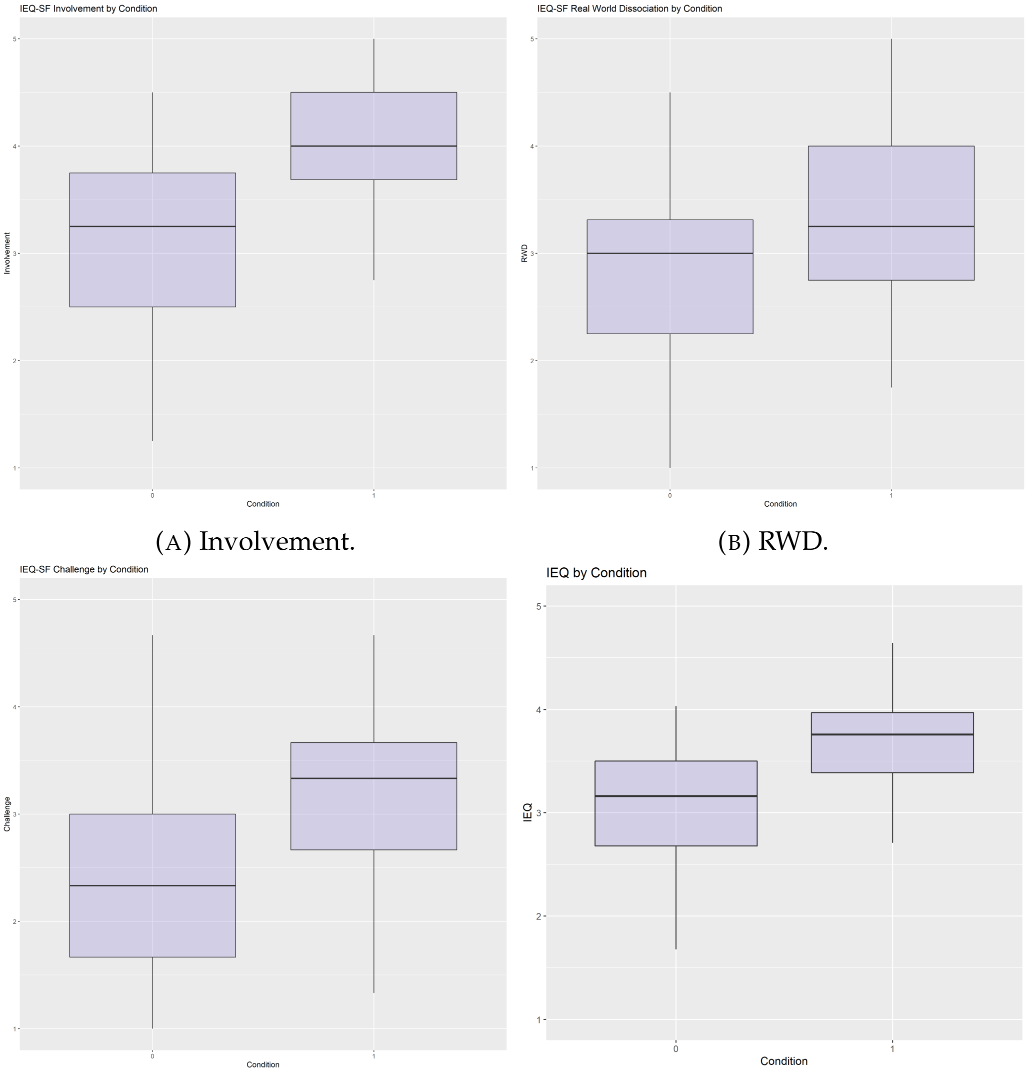
As per the success criteria, Cronbach α coefficients were also calculated for each factor of the IEQ-SF in this study.

5.7.5 Results

There was a statistically significant difference observed between conditions for every factor of the IEQ-SF, the results of which are presented in table 5.12. For reference, conducting an analysis of this data with the original formulation of the IEQ resulted in a statistically significant difference between the two conditions (*df* = 158, $t = -7.854, p < 0.001$) with a substantial effect size ($d = -1.242$, $[-1.583, -0.9]$). Boxplots corresponding to each factor of the IEQ-SF as well as the originally published IEQ scores are provided in figure 5.4, showing a strong observed effect irrespective of factor or form.

All factors of the IEQ-SF correlated significantly with the remainder full IEQ, and a single score derived from all IEQ-SF items also correlated significantly with the remaining items on the full IEQ. The results of these analyses are presented in table 5.13.

For the final metric of scale reliability, Cronbach α was found to be $\alpha = 0.72$ for involvement, $\alpha = 0.69$ for real world dissociation, and $\alpha = 0.71$ for challenge, which were all deemed to be within the threshold for reasonably reliable (α = 0.65).



(A) Involvement.

(B) RWD.

(C) Challenge.

(D) Full IEQ.

FIGURE 5.4: Boxplots showing IEQ-SF scores across each factor between the two experimental conditions, as well as the original full IEQ scores as reference.

TABLE 5.13: Correlation results between each IEQ-SF factor score and a remainder IEQ score derived by removing all IEQ-SF items from the full IEQ. An additional full IEQ-SF score was also computed due to the limited nature of correlating with remainder sets.

Factor	p	r	Conf. Intervals
Involvement	<0.001	0.85	[0.802, 0.889]
RWD	<0.001	0.64	[0.533, 0.720]
Challenge	<0.001	0.55	[0.435, 0.652]
Full IEQ-SF	<0.001	0.84	[0.793, 0.884]

5.7.6 Discussion of Study III

This study was an independent validation analysis using the new IEQ-SF, of a previously published statistically significant result that had used the original IEQ. Given the strong effect observed in the original publication, the results found here using the IEQ-SF are reassuring. The effect sizes observed with the IEQ-SF range from roughly moderate ($d = -0.629$) to very strong ($d = -1.325$), which are comparable to the originally reported effect size of $d = 1.24$). These results show that the IEQ-SF is a viable alternative to the full IEQ, with some minimal loss to power.

Evidence was also found in favour of the second alternate hypothesis, with moderate to strong correlations observed between each IEQ-SF factor and a score computed from the remaining items of the full IEQ. Furthermore, Cronbach α coefficients for each factor appeared to indicate that even in a smaller and more representative sample for a real world use of the IEQ-SF, the questionnaire is still adequately reliable.

However, a limitation of this study was the fact that the IEQ-SF was simply derived from data that had already been collected using the full IEQ. Therefore, the case could be made that there exists some quality in the process of completing the full IEQ that may have been transferred to these IEQ-SF scores that would otherwise not exist were the IEQ-SF to be the sole questionnaire completed by participants. Perhaps some items in the full IEQ are able to inform participants about others, either by a cueing effect or by task familiarisation. This limitation discredits the ecological validity of the IEQ-SF, and requires new experimental data in order to be addressed.

5.7.7 Conclusion of Study III

The IEQ-SF was shown to be able to successfully reproduce a previously published effect on immersion and its sub-components in an independent validation analysis. However, because this was accomplished with previously collected data, there could be no certainty about the true applicability of the IEQ-SF in real world research. Therefore, further evidence was needed showing the success of the IEQ-SF on entirely its on merits.

5.8 Phase IV. Study: Pre-Registered Replication Experiment

The final phase study of this series concludes with a replication study that aimed to collect entirely new experimental data while using only the IEQ-SF as the sole measuring apparatus. This was done with the aim of being a final validation study that applied the IEQ-SF in a real world experimental context. The experiment in this study is a replication experiment based on the one used in Cutting et al.'s study which was used in the previous phase validation study. The same experimental task and procedures were replicated, with the crtical difference being the use of the IEQ-SF in place of the full IEQ. This study was also separately pre-registered on the Open Science Framework, with the id 3pgv2 (Aung et al., 2021b).

5.8.1 Participants

The participants for this experiment were recruited through the Prolific online experiments platform, and were financially compensated with £6 for their time. All participants were English speakers and had not completed any similar similar game experiment before.

The recruitment target was to obtain sample of 160 participants, which is based on a mixture of previous research that this study sought to replicate (Cutting et al., 2020), as well as newly acquired results from the previous phase validation study.

In the prior work by Cutting et al., a lab experiment with $n = 36$ participants detected an effect with size $\eta^2 = 0.126$, and two separate online pilot studies reported similar effect sizes of $\eta^2 = 0.124$ and $\eta = 0.2620$. A power calculation conducted on the lowest effect size from these reported values suggested that each condition should have 66 participants in order to produce a power of .96, totalling 132 participants. A pilot study was also conducted for this study and pilot sample of $n = 30$ observed effects that were in line with previously reported effect sizes (Involvement: $p < 0.001, d = -1.64$, RWD: $p = 0.12, d = -0.99$, Challenge: $p = 0.015, d = -0.95$).

The effect sizes found in the offline validation analysis in the previous phase study observed Cohen d effect sizes of .1.325, 0.629, and .833 for each factor. With the smallest of these effects ($d = 0.629$ for the Real World Dissociation factor), a two tailed a-priori power analysis revealed that a total of $n = 140$ participants would be required to achieve a power of .96.

Based on the power analyses above, a total of 180 participants were deemed to be adequate for the sample size, where an additional number of participants were included to compensate for an anticipated possibility that there may have been a loss of power due to the use of the IEQ-SF rather than the IEQ.

Participants who did not complete the whole study, self-reported any form of colour blindness, or reported any bugs or glitches with their experience of the task were all removed from the sample and were replaced with new participants before any analysis was conducted. Additionally, participants who paused for longer than 20 seconds between any two moves were also excluded as they were deemed to be inattentive in completing the experimental task. Participants who did not complete the tutorial section of the experimental task were also removed from the sample as they did not complete the main task and therefore were not exposed to the experimental manipulation. Any participants with incomplete or unrecorded data due to technical issues were excluded. Finally, a dummy question was included in the IEQ-SF and participants who did not answer this question correctly were also removed for failing an attention check.

In total, 206 participants were recruited for the main experiment, of which 7 participants were omitted for colour blindness, 1 participant was omitted for not completing the tutorial, 14 participants were omitted for taking a break during the game, 6 were omitted for failing the attention check dummy question, 12 participants were omitted due to faulty behaviour of the study (bugs and glitches, etc), and 2 participants' data were lost due to technical issues. The final remaining sample was $N = 164$ participants, with 83 participants in the low immersion condition and 81 participants in the high immersion condition.

The demographic composition of the remaining participants that met the selection criteria are as follows. Participants' gender composition were made up of 80 identifying as Females, 83 as Males, and 1 as Non-Binary. The mean age of participants was 27.1 years old, with a standard deviation of 6.26 years and a median of 26 years. The youngest participant was 18 years old, and the oldest participant was 41 years old.

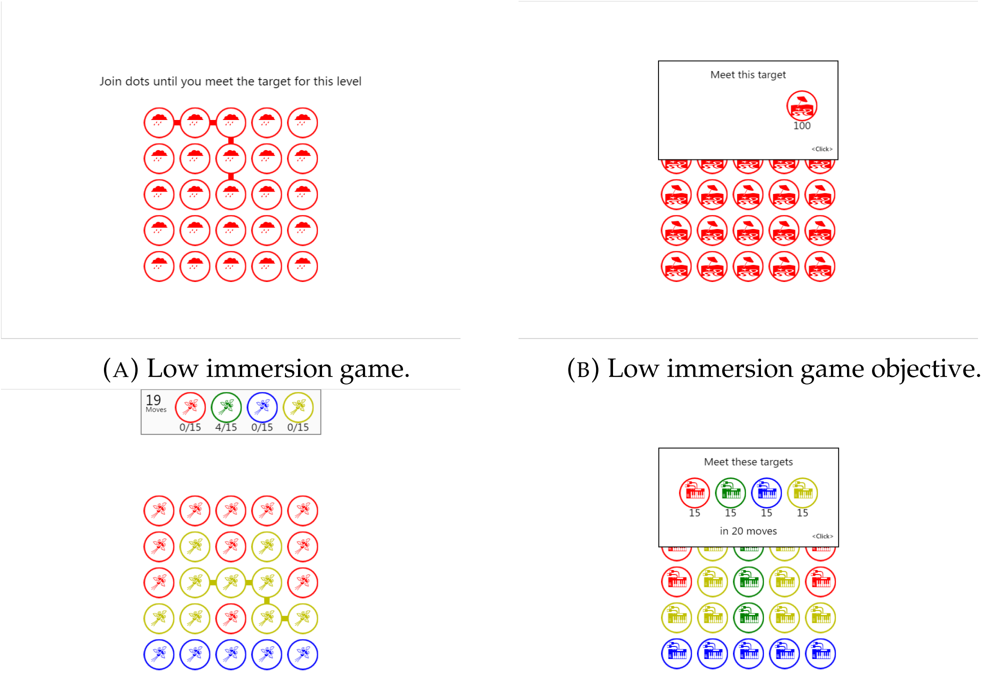
5.8.2 Materials

The main experimental task was the game Two Dots, with different variants based on the experimental manipulation. Apart from this difference on the game's immersion, both games were designed to involve similar visual stimuli and similar motor actions. There were two variants of *Two Dots*, one for each experimental condition. *Two Dots* is a simple, self-paced puzzle game that is engaging and can be learnt quickly. The game is played on a grid containing different coloured dots, and the objective of the game is to join adjacent dots of the same colour while meeting a target number of joins within a set number of moves.

The high immersion variant of the game was a direct clone of *Two Dots*. The second, reduced immersion level variant of the game, was the same game except all dots were rendered in the same colour. By making all the dots identical, the game was designed to be made less engaging, which was intended in order to reduce the level of immersion recorded by participants. This was accomplished despite the fact that all participants would be performing the same activity as with the high immersion variant of the game. An additional difference between the two variants of the game was the removal of the goal number of joins and moves remaining display in the low immersion variant of the game, which was designed to further reduce participant immersion in this condition. Finally, both variants of the game had the distractor images rendered within the dots using the *Webdings* typeface.

The game was programmed in Javascript so that it would run in a web browser for an online study. A version of the experiment that does not save data and allows choice of conditions can be demoed online.

Statistical analyses were conducted with the *R* programming language, and the package *psych* was utilised for the calculation of Cronbach α reliability statistics. Finally, power analyses were conducted using the *pwr* package.


Join dots until you meet the target for this level

Meet this target

Meet these targets

in 20 moves

(A) Low immersion game.

(B) Low immersion game objective.

(C) High immersion game.

(D) High immersion game objective.

FIGURE 5.5: Screenshots of the experimental task from each condition, of both the game and the objective targets provided to players at the start of each level.

5.8.3 Design & Hypotheses

The experiment was a between participants study with two conditions. The main experimental manipulation was the level of immersion present in the game, which was determined by the presentation of the game itself. Screenshots of the different verisons of the game are presented in figure 5.5.

There were four hypotheses generated for this study. First, it was hypothesised that there would be a significant difference in the level of Real World Dissociation reported between each of the games/conditions, as measured by the IEQ-SF. The high immersion game was hypothesised to have the higher Real World Dissociation scores. Second, there would be a significant difference in the level of involvement reported between each of the games as measured by the IEQ-SF involvement factor, and the high immersion game would have the higher involvement score. Third, there would be a higher challenge score reported by players via the challenge factor of the IEQ-SF, and players in the high immersion game would report higher challenge scores than players in the low immersion condition. Fourth, there would be a significant difference in the number of distractor images recognised between the high immersion and low immersion game, and participants in the low immersion condition would recognise more distractor images. All hypotheses were statistically tested with two tailed tests at a significance threshhold of $\alpha = 0.050$.

Because the full IEQ was no longer being used, scores could not be computed for internal reliability using the remainder to IEQ-SF score correlations, so this test which was conducted and reported in the previous chapter was omitted here. Additionally, the time to complete the IEQ-SF was also measured in order to confirm expectations that the questionnaire could be completed in a minute or less.

5.8.4 Procedure

Participants were allocated to the high or low immersion conditions randomly, in such an equal manner that 90 participants were allocated for each experimental condition. Following the acquisition of informed consent and an instruction and practice session for the experiment, participants would play the condition dependant version of the game.

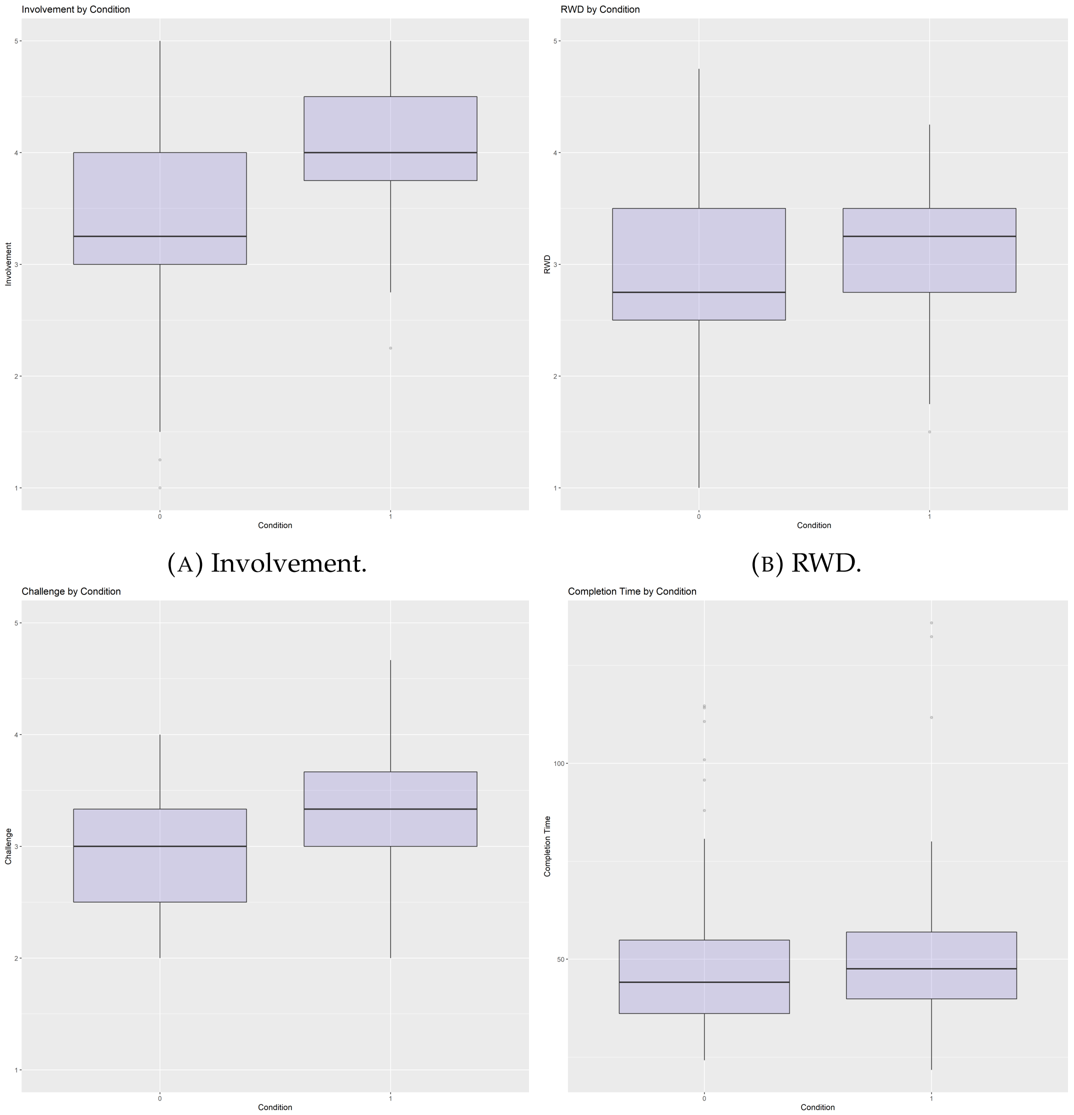
During game play, each participant was shown 60 distractor images in a random order. These images were randomly chosen from a pool of 90 images. Immediately after the end of the main experimental task, all participants then completed a forced choice test on 30 images that they had been shown during the experimental task, in order to measure how many images they recognised. The questions in this test were displayed in a random order determined using a Fisher-Yates shuffle algorithm. Random numbers for this algorithm were generated by a linear congruential generator algorithm built into the experimental source code which followed best practises to maximise the randomness of the output.

Finally, participants were then asked to complete the IEQ-SF to measure their experience of the game, which also included the dummy question intended to check participant attention.

5.8.5 Results

A t-test was conducted for each factor of the IEQ-SF to determine any difference between the low and high immersion games. Table 5.14 presents the results of these analyses. In summary, a statistically significant difference was observed in IEQ-SF scores between the two experimental conditions for every factor, after a Holms-Bonferonni correction for multiple comparisons was applied. Particularly noteworthy is the fact that the 95% confidence intervals for the Cohen d effect sizes in every factor did not include 0 within its bounds, indicating a high probability that an effect exists for every factor even if the point estimators for their magnitudes are not accurate. An accompanying set of boxplot graphs for each factor's IEQ-SF scores can be found in figure 5.6 sub-figures (A)-(C).

A t-test conducted to determine any difference in distractor image recall found a statistically significant difference ($p < 0.001, t = 4.399, df = 162$), with an approximately moderate effect ($d = 0.69, 95\%CI[0.369, 1.004]$). A greater number of items were recalled by the low immersion condition $M = 19.75, SD =$ than the high immersion condition $M = 17.57, SD =$0. Pearson's correlations were calculated to determine the influence of each factor



(A) Involvement.

(B) RWD.

(C) Challenge.

(D) Completion Time

FIGURE 5.6: Boxplots showing IEQ-SF scores across each factor between the two experimental conditions, and time taken to complete the IEQ-SF questionnaire at the end of the experiment. One participant was omitted from this graph in (D) due to taking longer than 200 seconds to complete the IEQ-SF, but was retained for the statistical analysis.

TABLE 5.14: Results for score differences between the low and high immersion conditions, within each factor of the IEQ-SF. Note that the p values provided in the columns p_{holms} are corrected for multiple comparisons using the Holms-Bonferroni procedure.

Factor	p_{holms}	t	Effect Size	Effect Size CIs	Low Immersion	High Immersion
Involvement	<0.001	-6.22	$d = -0.97$	[-1.29, -0.64]	$M = 3.38$ SD = 0.67	$M = 4.08$ SD = 0.34
RWD	0.006	-2.78	$d = -0.44$	[-0.747, -0.123]	$M = 2.9$ SD = 0.48	$M = 3.19$ SD = 0.35
Challenge	<0.001	-4.88	$d = -0.76$	[-1.08, -0.44]	$M = 2.96$ SD = 0.25	$M = 3.35$ SD = 0.27

TABLE 5.15: Correlation results between each IEQ-SF factor score and the number of distractors recalled. p values adjusted by the Holms-Bonferroni correction are provided in the column

p_{holms}

to control for family-wise error rate.

Factor	p	p_{holms}	r	Conf. Intervals
Involvement	0.031	0.062	-0.168	[-0.31, 0.889]
RWD	0.005	-0.22	0.015	[-0.358, -0.066]
Challenge	0.85	0.85	-0.015	[-0.168, 0.139]

on image recall, and these results are reported in table 5.15. After a Holms-Bonferroni correction for multiple comparisons, the real world dissociation factor had a significant correlation with the number of items recalled by participants, while the involvement and challenge scores did not.

Pertaining to the success criteria of being rapidly applicable, it was found that the IEQ-SF was completed in $M = 51.3$ seconds on average, with a standard deviation of $SD = 23.7$ seconds. Additionally, a t-test was conducted to determine if the condition played had any influence on the time to complete the IEQ-SF. There was also no statistically significant difference in time to complete the IEQ-SF ($p = 0.8$, $t = 0.253$, $d = 0.039$) between the control condition ($M = 51764.13$ms, $SD = 27327.15$ms) and experimental condition ($M = 50824.48$ms, $SD = 19565.18$ms). A boxplot of times to complete the IEQ-SF is provided in sub-figure (D) of figure 5.6.

Finally as measures of internal reliability, the Cronbach's α was computed for each factor of the IEQ-SF using per-question scores from the full sample. An $\alpha = 0.77$ was observed for the involvement factor, $\alpha = 0.24$ for RWD, and $\alpha = -0.89$ for challenge. Because of the low α on the RWD factor, per-omission values were recorded which produced $\alpha = 0.75$ were the first item

TABLE 5.16: Covariance matrix between each item of the Real World Dissociation factor from participants completing the IEQ-SF.

	RWD1	RWD2	RWD3	RWD4
RWD1	1.39	-0.33	-0.70	-0.50
RWD2	-0.33	1.54	0.93	0.61
RWD3	-0.70	0.93	1.29	0.63
RWD4	-0.50	0.61	0.63	1.48

of RWD to be removed, corresponding to the question "I felt consciously aware of being in the real world whilst playing". Similarly, the removal of item 2 would also have substantially improved the consistency to $\alpha = -0.57$, which was the question "I forgot about my everyday concerns". When splitting the sample by condition and examining the α per group, low values were still observed ($\alpha = 0.32$ for the low condition, and $alpha = 0.09$ for high immersion). Therefore, it was decided that the assumptions of the Cronbach α statistic should be checked. Here, it was found that the covariances between each item were not equal, and the covariance matrix can be found in table 5.16 which shows that there was unequal covariance between the items of the RWD factor, suggesting that the Cronbach α was not a suitable metric for this sample.

As an additional measure, scale reliability was re-computed using Revelle's omega total ω_t, based on guidance by Revelle (Revelle & Zinbarg, 2009) showing it to be a more accurate estimation of reliability than both Cronbach's α as well as another alternative approach of calculating the greatest lower bound. Here, estimations of internal consistency were higher for all three dimensions, with involvement $\omega_t = 0.82$, real world dissociation $\omega_t = 0.84$, and challenge $\omega_t = 0.74$, all of which suggest a greater and more acceptable internal consistency for the IEQ-SF than Cronbach's $\alpha 0$.

5.8.6 Discussion of Study IV

This was a confirmatory experiment which demonstrated that the IEQ-SF was solely able to produce similar measurements and subsequent results to those previously published by Cutting et al with all four null hypotheses being rejected in favour of the alternate hypotheses. The results in this study provide encouraging evidence that the IEQ-SF as a measurement tool is a viable alternative to the IEQ while remaining able to measure the same latent constructs.

The IEQ-SF was also found to be completed quickly by the overwhelming majority of participants in under a minute, which means that the new questionnaire fulfils one of the defined essential success criteria for a successful short form. That there were no differences in completion times between either conditions would also indicate that participants were able to complete the IEQ-SF irrespective of how immersed they were during the experiment.

Compared to results from the phase III study, effect sizes were somewhat smaller despite the slightly larger sample size. It is difficult to reason why this may be the case, and there is certainly a considerable probability that this may simply be nothing more than simple variance between two different experiments. However, it should also be acknowledged that there is a possibility that completing the full IEQ may involve some other additional benefit than simply having more items. For instance, one reason this final experiment was conducted was due to the uncertainty around any existing biases on measurements from priming or cueing effects occurring due to a higher number of opportunities for participants to process different questions, i.e. points of measurement on the latent construct of immersion. This matter would be a point for further research, but ultimately does not change the conclusion here that the IEQ-SF was capable of replication previous findings published under the full IEQ.

The low Cronbach α for the real world dissociation factor was also a cause for concern, and despite the still relatively significant and strong results of the experiment, a scale with a poor consistency metric required addressing. The result in itself was somewhat surprising, given that the $/alpha$ for the real world dissociation factor fluctuated around the .65 mark in previous examinations of the IEQ-SF's internal consistency. It's plausible that similar to the reduction in effect sizes, there existed some effect driven by the relationship between items within factors that provided participants with better clarity or information when answering the full IEQ, and that this effect no longer existed in the IEQ-SF. This also eliminated the possibility that the scale consistency was manipulated by the experimental condition itself, which was the state of immersion induced by the task. Since there was a failure to meet one of the core assumptions of the Cronbach α, the pre-defined success criteria for internal consistency was no longer adequate in this case and further research is required to establish the reliability of this scale.

5.8.7 Conclusion of Study IV

Despite issues with internal consistency and the slightly lower effect sizes, the results of this experiment indicate that the IEQ-SF as a sole measurement device could adequately replicate previous findings captured with the full IEQ with a minimal loss of power and measurement validity. Further work examining the extraneous effects of reducing the number of items in a questionnaire is warranted given the effect sizes observed here.

5.9 Discussion of findings across all studies

This series of studies have developed, produced, and validated a functioning short form alternative to the IEQ originally published by Jennett et al., with minimal loss to power or validity. Based on the pre-defined success criteria in section 5.4, the final formulation of the IEQ-SF was able to meet almost all the set targets. Based on results from the phase 4 study, the IEQ-SF could be completed in less than a minute by the majority of participants completing the questionnaire. Furthermore, other than a few select instances, the IEQ-SF was also able to produce relatively acceptable levels of internal consistency as measured by the Cronbach α, and the series of validation analyses and additional study applying the IEQ-SF have shown that results previously produced with the full IEQ could be replicated with the IEQ-SF. These all indicate that the IEQ-SF is a viable measurement apparatus in its own right, and comparable to the original IEQ in function. However, there were several caveats involved in reducing the IEQ to this new form factor, including alterations to the original factor structure, concerns surrounding the internal consistency of the questionnaire, and moving from the original single immersion score to a set of per-factor scores.

The most noticeable change in the new IEQ-SF is that unlike the original IEQ or first attempt at the IEQ-SF, an overall Immersion score by taking the mean across all the factors is now no longer the way of using the questionnaire. While this would still technically be possible to have a single measurement of immersion under the IEQ-SF, this approach was considered to provide little benefit in comparison to assessing each factor individually. One of the greater justifications for this comes from the fact that the factor structure of the IEQ-SF no longer matches the original IEQ, and because of this the single IEQ-SF score is no longer comparable to the full IEQ. From the perspective of developing a tool whose users are researchers and academic practitioners, strongly separating the IEQ-SF from the original IEQ by splitting into per-factor scoring provides the intended end user with a clear conceptual property that differentiates these two questionnaires. Furthermore, because the IEQ-SF has fewer items in total, there is additional value in inspecting each factor individually where the convenience of a single score is traded with the added information in a granular inspection of player experience. Finally, there have been arguments made by other games researchers developing psychometric tools that a multi-dimensional latent construct should be measured multi-dimensionally, without a single score. Such an approach has already been applied in recent work on the Player Experience Inventory (Abeele et al., 2020) which treats the multi-dimensional scoring and analysis of its factor structure as immutable for its users. Given the somewhat pre-theoretical nature of many player experience measurement tools, it would certainly be more beneficial for researchers to be more concerned with the granular details of a player's experience. For instance, it is far more valuable to know that participants who completed the experiments by Denisova did indeed experience the greatest experiential differences within the challenge factor, which is reassuring since these experiments primarily focused on the notion of player beliefs on challenge and adjusted difficulty. This is also particularly useful since the overall IEQ score itself never contained interpretable meaning within its raw values. Knowing that two participants might record different IEQ scores of 3 and 4 did not provide precise information about the extent to which they were immersed, but rather relative information regarding an approximate state of immersion from one another. It was also never a practical possibility that participants would score at or near the floor of possible IEQ scores, and studies using the IEQ simply reported on differences in immersion levels between experimental conditions. With all these limitations of the original IEQ scoring system in consideration, it was concluded that moving away from the single score and focusing on comparing individual factors instead would allow researchers to focus on more meaningful information captured by the scales. Therefore, it was determined that the new approach of treating the scoring and analysis of the IEQ-SF should be considered to be an advantageous change from the previous method of scoring the IEQ with a single value.

One of the success criterion defined early in the study was the necessity for the IEQ-SF to follow the original IEQ factor structure. Strictly speaking, this criterion was not met as an entirely new factor structure was derived in order to achieve a better measurement of player immersion. One could argue that the new IEQ-SF factor structure in fact matches the factor structure of an entirely new IEQ defined by a new 3 factor structure, which does accurately describe the developmental process of the IEQ-SF. However, this would also require a new re-definition of the full IEQ to be used in a similar manner to the IEQ-SF described in this chapter, and to address that the original IEQ has an incorrectly specified factor structure and scoring system. This is out of the scope of the set of studies described here, but it is worth acknowledging that taking such steps with re-defining the original IEQ would be the logical conclusion of accepting the IEQ-SF. Most importantly, the new IEQ-SF was still able to produce results statistically comparable to those that were previously published, which itself would indicate that the broad definition of immersion and its sub-components were adequately specified in both the original IEQ and the IEQ-SF.

Another essential criterion for the success of the IEQ-SF was the ability to be rapidly applicable and be completed in a brief period of time by participants. The development of short form questionnaires is a common practice in applied psychometrics, but one worry with the IEQ-SF and player experience research in general was the possibility that the act of completing a questionnaire would break or interrupt the state of immersion experienced by players. In making a questionnaire that can be quickly completed, this risk is minimised, but it should also be noted that this risk is not entirely removed. In the phase IV study, the IEQ-SF was simply completed once as the main focus was to confirm the construct validity of the IEQ-SF and its ability to replicate a previously published result. One potential case could be that immersion is a fragile state that is maintained tenuously, and simply drawing cognitive focus and resources away from the game in order to answer a questionnaire in an entirely different task would impede upon the immersion of a player irreparably, at least within the temporal scope of a games experiment. Thus, additional research would therefore be necessary to confirm that the use of the IEQ-SF over multiple periods of play would not hinder the immersive state of a participant in a manner consequential to the goal of accurately measuring player experience without intervening in the process.

The capability of the IEQ-SF to be rapidly applied, if confirmed, would challenge the notion that physiological measures of player experience are the best means of capturing granular information of player immersion. Previously, physiological measurement methods were primarily suggested due to their ability to capture player experience measurements without interrupting said experiences. However, results from the Don't Starve study in chapter 1 have shown that participant comfort might potentially have a profound impact on their experiences. Furthermore, at the time of writing, I could not find any published applications of psychophysiological measurements that informatively describes engagement or immersion at a granular level. Studies that apply physiological measures of experience are still limited to aggregate comparisons of relative player experience between experimental groups, rather than providing information of player experience within the players' range of experiences themselves. This limitation of psychophysiological measurements are quite comparable to the limitations described earlier of the original single IEQ score, and comparatively, treating the IEQ-SF as a multi-dimensional latent construct with individual metrics on sub-components of immersion is in fact much more informative than a physiological correlate with an existing psychometric score. In this sense, the IEQ-SF has fulfilled its fundamental goal of providing a granular yet informative measurement of player immersion, potentially in a manner more useful than what can be achieved with the current state of the art in applied psychophysiological research.

Additionally, the utility of the IEQ-SF as a questionnaire that can be quickly answered by participants means that the work currently carried out to make advances in applied psychophysiological research would benefit from the ability to compartmentalise and granular measure the different constituent constructs of immersion in a temporally granular manner. One of the limitations of the previous works in this thesis was the fact that a continuous and large series of measurements were required to be correlated with a single measurement produced by the original IEQ. If this single score was instead a timeseries (even if discrete) of scores, the task of meaningfully extracting information regarding the relationship between a signal timeseries with a psychometric timeseries would be a substantially easier task. So, at worst, were the IEQ-SF to be shown to be a valid measurement method over multiple periods of play, the short form questionnaire would still be an indispensable stepping stone to the continuous granularity that psychophysiological research practitioners are seeking to achieve.

One of the limitations of the study was the fact that the criteria for selecting the winning candidate factor structures, and the suitable items to be kept after reduction, were somewhat arbitrary. While hypotheses could be easily formulated for whether a pre-registered replication study, it was harder to define clear hypotheses for replication analyses. On one had, it would have been simple and easy enough to simply state that analyses would be the same. However, in this approach there would be difficulty in navigating the intermediate grey area of an arbitrary number of failures to replicate. Furthermore, this only applied for analyses conducted after the process of reviewing factor structures and selecting the most appropriate factor structure or most appropriate item. At the point of selecting these items, hypotheses are less relevant compared to sound understanding of the domain of interest, which in this case was immersion. Still, it would be better to consider ways to build more formal methods for stating and confirming expected outcomes of factor analyses in cases like this.

5.10 Conclusion

In this chapter, a short form variant of the IEQ was conceived, developed, and validated in a series of studies that involved both previously obtained experimental data, as well as a new sample of experimental data captured solely with the IEQ. Considerable care was exercised to define the space of the latent construct of immersion, and several validation approaches were taken to confirm this factor structure. Results from validation studies in this chapter have shown that the IEQ-SF is a valid, reliable, and adequately powerful tool to measure player immersion, at least relative to previous uses of the original IEQ. Intended as a psychometric tool that could be rapidly applied, the IEQ-SF fulfilled the specified success criteria for a short form immersive experience questionnaire, and therefore is ready to be tested for its primary intended use case of repeated applications across multiple stages of play, in future research.

Chapter 6

Discussion & Conclusion

The objective of this thesis was to answer the question *"Can immersion of real world games be measured non-disruptively at the same time as a video game is being played?"* To answer this question one psychometric and two psychophysiological studies were conducted.

Results of both planned and exploratory analyses from the psychophysiological studies were mixed, leading to a need for more nuanced interpretations in establishing findings of these studies. On the other hand, the combined results of all four phases of the psychometric study provided both quantitative and theoretical bases for the development and use of a reduced version of the IEQ. For brevity, repeated statements of results from each experiment are minimised and instead effort is drawn to discuss the core essence of conclusions drawn from each study, and then the thesis overall.

6.1 Findings

To navigate the collective set of results in this thesis, the findings in this section can be divided into broad categories:

- Findings that directly provide evidence pertaining to the measurement of immersion using psychophysiological techniques.
- Findings that provide insight into designing experimental procedures for research involving the use of psychophysiological techniques.
- Findings towards the development of the IEQ Short Form to be used in future work answering this research question.

For each of these findings, limitations constraining the interpretation of relevant results are also discussed in order to provide a comprehensive outlook of the contributions of this thesis.

6.1.1 Differences in pupil diameter may not be caused by differences in immersion

In Study I, statistically significant differences in pupil diameter were detected between different in-game times of day. Given the gameplay differences between times of day, it is not unreasonable to interpret this as indicative of a difference in cognitive activity between times of day. However, even though environmental luminosity was controlled for, times of day also included a luminosity response from the stimulus that significant limited such an interpretation, thus requiring a further experiment controlling for this confound. In general however, it is also hard to take a an average of a set of highly variable measurements over a longer period of play and expect the information in the temporal domain would be retained in any aggregate analysis. The in-game daytime epoch alone is 8 minutes long, and includes a variety of in-game events that may likely induce different pupillary responses that cannot be so simply aggregated. Hence it would be more reasonable to include analyses at more than one level of temporal granularity in a study using this modality.

6.1.2 Heart rate metrics may be capable of measuring cognitive workload

In Study II, a significant difference in RMSSD between resting and hard difficulty play may have been caused by the greater cognitive workload of the high difficulty condition. Unlike study I, the stimulus presentation and experimental design was better controlled, lending credence to interpretations of analyses of activity, at least at the aggregate level. The nuance here was instead drawn to the amount of interactivity specifically— no differences were observed between the rest and observation conditions, or the observation and easy conditions. Instead, it was the jump to actually playing that resulted in differences between RMSSD in conditions. This could be explained by the jump from no interactivity at all in observation and rest, to having to actually play in easy and hard. As always, improvements could have been made and this time the most important follow-up would have been an experiment controlling for any potential respiratory confounds.

6.1.3 Short form psychometric scales can be designed to measure over time

In Study II, the NASA-TLX illuminated both stability in responses, as well as the sensitivity to detect differences in workload between the different experimental conditions. The counterbalanced design of condition order supports this interpretation, as the TLX score differences between conditions were present despite counterbalancing, which accounts for order and fatigue effects in the NASA-TLX as a scale participants had to respond to. Based on the successful implementation of the NASA-TLX in Study II, the IEQ-SF was developed as the final contribution of this thesis.

6.1.4 It is important to test assumptions and common practices when shortening a questionnaire

In the development of the IEQ-SF during Study III Phase I (section 5.5), it became clear that checking of assumptions became especially important. Having worked with the IEQ for a number of years, and given the IEQ's presence within games research for the past decade, an initial assumption was held about its factor structure. This assumption eventually turned out to be false, as demonstrated by the much improved reliability of the multidimensional IEQ-SF in Phase II of the study (section 5.6). Therefore, when given with a task to reducing any given questionnaire, it is essential to test the assumptions of the underlying structure of a questionnaire even when there may exist substantial practice and evidence to existing assumptions. Moreover, it is especially important to recognise that even if a questionnaire is often analysed at an aggregate level, such as overall immersion as defined by the IEQ being taken as an appropriate description of the scale's structure.

Testing assumptions of structure also revealed that by taking the aggregate immersion score of the IEQ as its primary outcome measurement (as it is often treated in practice), issues with particular subscales of the questionnaire may become shrouded over time. Common practice adopted in research often transforms into a sometimes mistaken assumption that they are best practice. In the case of immersion, the fact that three of the sub-scales were far more reliable than the remaining factors illustrates that old assumptions of the IEQ should have been, and were challenged in the process of designing the IEQ-SF. Best put, part of reducing a questionnaire should be to conduct a review of the scale in its entirety, before attempting to extract the most valuable items for a smaller variant.

Similarly, when checking for the internal consistency of a scale or subscale, it is also important to verify whether the assumption of equal covariance between items is met. If not, and arguably even in cases where it is, it is preferable to use alternative methods such as the ω_t or ω_h.

6.1.5 The IEQ Short Form can be used standalone

In practice, the final multidimensional IEQ-SF model chosen could replicate previously published differences in IEQ scores across multiple papers. This was accomplished with both replications of analyses from previous studies, as well as implementation of the IEQ-SF on its own in a replication experiment. Particularly in cases of slightly larger samples, the IEQ-SF provided estimates of effect sizes that aligned with previous findings. However, even in smaller studies where the IEQ-SF was likely to have lost power, the *D* effect size ratio still indicated that the proportion of effect sizes was not so much worse than previously published results to have been concerning.

6.1.6 Short questionnaires can be deployed rapidly, with minimal interruption to experience

Results from study IV have shown that once familiar with the questionnaire, participants were able to complete the IEQ-SF in short order. Similarly, the NASA-TLX was completed during intermittent periods between play in study II. Both of these results suggest that immersion, or at least cognitive load levels are able to be maintained or recovered after small interruptions. This is a promising finding not just for the IEQ and immersion research, but any games research seeking to capture a measurement of latent experiences over time. The limitation here, is of course the fact that there was not an opportunity seized in this thesis to deploy the IEQ-SF in its intended use case of repeated administration, and such an application should be the most immediate subsequent piece of work.

6.2 Limitations

6.2.1 The IEQ may have reached the current limits of current theory

An important reflection to be made, particularly after the final chapter, is that the theory of immersion in its current form appears not to have changed significantly since its original conception. Although there is increasing interest in considerations for the psychological functions of immersion Cairns et al., 2014a, the theory of immersion is currently still very much based on its original conception through grounded theory.

It is also worth asking if work in measuring immersion is actually achievable with the current framework available. In this thesis, the original goal was to measure immersion with more granularity, and this goal was kept the same throughout the course of this thesis. However, consider the possibility that immersion as it is presently defined does not have a clear enough theoretical framework for truly granular measurement. In this situation, the current approach taken so far in this thesis may be confined to the limits set by this possibility. The question then, might be what specifically would be the correct approach?

One consideration is to look at adjacent areas to immersion that have clearer and more established theory. For example, work by immersion researchers such as that of Cuttings Cutting, 2018 initially began with a consideration of immersion but then took a path towards a more specific and psychologically focused area, which in this case was attention. The thinking here may very well be the right approach forward: rather than continuing to explore measurement approaches using immersion in its present form, perhaps work should instead be done by exploring measurement of psychological processes that one might believe are strongly tied to immersion.

In fairness, the use of the NASA-TLX in this thesis was one such instance where a step was taken towards this direction. Although this endeavour was cut short by the COVID-19 pandemic, it was a line of inquiry that could yet lead to better insights on the nature of how one gets immersed with games.

Nonetheless, there is an ever-growing argument for stepping back from further studying the measurement of immersion in its current form and to instead turn towards a more psychologically informed theory of immersion.

6.2.2 The IEQ-SF was not fully validated for its intended use

One of the goals of developing the IEQ-SF was to develop a tool capable of being administered multiple times over the course of a single play session. There is an assumption and a hypothesis that this use of the IEQ-SF and the resulting interruption of gameplay would not cause unacceptable amounts of deleterious effects on a player's immersion. This assumption was ultimately not tested in this thesis due to time constraints and the subsequent impact of the COVID-19 pandemic.

Although promising results indicate that the IEQ-SF could be used with minimal interruption of immersion, this thesis can only make such a statement specifically for cognitive workload. The NASA-TLX was able to measure sustained cognitive workload between sessions of participants playing *Osu!*.

The limitation of course is that cognitive load is only a potential component of immersion, and while perhaps cognitive load is resistant to interruptions, immersion in its totality may not be. To confirm whether this problem may exist, it would be paramount to conduct an experiment in which the IEQ-SF is in place of the NASA-TLX. Similarly, any researcher seeking to develop and deploy a reduced questionnaire should aim to apply their short form questionnaire repeatedly if their goal is to achieve time granular measurements.

6.3 Further Work

There are several experiments that immediately come to mind as low-hanging fruit for further work, based on both limitations from the studies as well as necessary follow-ups to findings in this thesis.

First, an eye tracking study in an experiment using the stimulus designed in Study II would have accounted for the lack of adequate confound controls in study I. In fact, this was the premise on which study II was designed until the apparatus became unavailable, and the modality was changed to heart rate variability. Such a study would ideally have also involved a pre-play session resting state measurement of pupil diameter against a no-stimulus fixation target, allowing for a more robust comparison of pupil diameter with normalised measurements. Comparisons of pupil diameter as a measure of cognitive effort across different difficulties could also be accompanied by analysing aggregate activity through stimulus locked pupil diameter measurements. This could be accomplished with the use of levels with planned peaks and troughs in level of activity throughout a single song. In doing so, the imbalanced sample of combat instances across in-game times of day could have been alleviated.

Second to this would also include a follow-up study using heart rate variability. Here, a strain gauge would be deployed to manage the confound of respiration. Additionally, like with the study suggested above, a combination of two experiments could be conducted. First, using stimulus blocks of levels designed to sustain a constant level of immersion. This would be followed by a second experiment using levels designed to induce variances in immersion through changes in interactivity.

Third, an experiment using the IEQ-SF with repeated responses over the course of a single session of play would be necessary to truly test the premise that immersion can be measured over time with a questionnaire. Of the potential future work suggested in this thesis, this would perhaps be the most obvious and fruitful next step. Even in the process of a failure to reliably measure immersion granularly with the IEQ-SF, there would at least be a clear signal that such an approach would be unsuitable.

Outside the measurement of immersion itself, it would also be valuable to determine how different measurement approaches influence the experience of immersion. For example, if the use of physiological hardware did in fact interrupt the immersion of participants playing *Don't Starve*, then is this interruption similar to, lesser than, or greater than interruptions caused by simply asking participants to answer a very short questionnaire? Conducting experiments that enquire about participant comfort and conscious awareness of wearing physiological signal apparatus, and experiments that more closely investigates of interruptions to play would contribute a great deal to this end.

More broadly, research efforts should also be made to more closely tie measurements of physiological signals and psychometrics together. This would ideally be accomplished with the use of short, repeatedly applied questionnaires, enabling the correlation of signals to psychometric scores. Ideally, this would be achieved by the use of a short questionnaire providing measures of immersion over the course of a play session, allowing for more granular aggregates of a signal to use in correlations.

Finally, a note should also be made for all future work in that a great deal of effort must be put into adequately planning and reporting analyses and pre-processing of data in this area of research. In particular, it is important to specify why certain pre-processing stages are conducted, such as the normalisation of a physiological signal (or lack of), or the use of one reliability metric over another. In both the development of experiments with psychometric scales, as well as physiological signals, a great number of decisions are made on how data is handled. It is only through the transparency of this conduct that collective progress can be made towards improving the measurement of immersion, and indeed any latent experience.

6.4 Closing Thoughts

To put it simply, measurement of a latent experience is a tricky task, and measurement of immersion with any kind of granularity over time is challenging. This problem is especially relevant to physiological measurements, where issues can arise from an innumerable number of sources. These include confounds caused by physiological responses to an unaccounted element in a stimulus, or even confounds caused by one physiological process in the body influencing the physiological process being measured. Research teams must wield a near encyclopedic knowledge collective knowledge of anatomy, statistics, and domain knowledge in games in order to escape unscathed by such unexpected confounds. In documenting some of the considerations required to manage physiological experiments, this thesis provides clear evidence that such an approach to researching measurement must be taken with care. However, despite the presence of such confounds, results from the physiological experiments here have also provided hope that there might indeed exist a basis on which we can infer some measurement of immersion from physiological signals.

It has also become more evident over the course of this thesis that the pre-existing modality of measuring immersion is possibly more robust and adaptable than previously believed. The IEQ can, for example, be reduced, and a small form variant has been shown to work in place of the IEQ should the need arise. This alone may provide a means to measure immersion granularly over time while the work to establish more objective signals continues. The process of reducing the questionnaire can also be taken and applied to other experiential questionnaires, and arguably it is this contribution that provides the best approximation and answer to the original research question. It does appear after all that it may be possible to achieve a measurement of immersion that is at least slightly more granular over time using a short questionnaire, even if the signal is only aggregated over blocks of levels or trials. Ultimately, up until this point in the field of games research, aggregating signals over blocks of play has also been the common practice in the analysis of physiological signals. So while psychometrics may not be as objective as physiological measurements, with only a little more validation research, they may be experientially unintrusive and temporally reliable enough to measure immersion over time.

Appendix A

Chapter 3 Appendices - Investigating the relationship between pupillometry, eye-tracking, and the IEQ

A.1 Participant Information Sheet - Control Condition

PID: E ___________C0

Participant Information Sheet

Overview

This experiment will consist of completing a video game session, prior to which you will be given a tutorial which will explain in detail how to play the game. The aim of this experiment is to evaluate player experience when playing a survival video game. We will also collect psychophysiological measurements in the form of eye tracking, and electrodermal activity (skin conductance) as additional data to the experiment.

Procedure

During this short session, you will become familiar with the gameplay and the controls of an survival video game, 'Don't Starve'. As a in typical survival game, the main character, Wilson, will have to collect and build objects in order to survive. During the trial session, your character will appear in a randomly generated world with objects that you will have to collect and monsters that you will have to avoid. Your character has a number of needs – you will have to feed him with berries, meats, eggs and carrots, so that he doesn't die of starvation. The heart shows your character's health – as long as nothing attacks him and he doesn't starve, he will be pretty much fine. To recover his health, he can eat flowers. Lastly, the brain shows his sanity – if Wilson enters a graveyard, walks in the darkness or doesn't shave, his sanity will go down, and as a result of that, it will be more likely that he'll get attacked by imaginary monsters. However, you can pick flowers to bring his sanity back up. The objects you collect are self-explanatory, but feel free to ask about any of them during the tutorial session. The creatures you will come across in the world can either attack you, protect you if you feed them, or become your dinner. Rabbits and birds can be caught and either eaten raw or cooked. Pigs are harmless unless you attack them – you can also befriend them if you give them meat. Most of the other creatures are likely to be deadly so it would be a good idea to run away if they spot you. The aim of the game is to survive – that is, your result will be composed of how well you do in the game, together with how many days you last. Your character is afraid of darkness, so you better light some sort of fire before it gets dark to keep him happy, dry and warm.

Main Experiment

The main part of the study consists of one gaming round during which you will be playing the game you have just tried during the tutorial. In the main part of the study, just like in the tutorial part, you will be playing in a randomly generated game world. Upon completion of this gaming session you will be asked to fill in a questionnaire about your gaming experience.

Recording Apparatus

During this experiment, several apparatus will be connected to you in order to acquire physiological information of the effects of this experiment. We will collect pupillometry using a head mounted eye tracker, which will record your eyes. We will also collect electrodermal activity, by placing two electrodes on your wrist. In order to collect these data, the experimental is required to place them on your body, with your permission.

Questionnaire

After playing the game, you will be asked to complete a player experience questionnaire. This questionnaire is comprised of 31 questions, for which you must provide answers on a scale of 1 to 5, where 1 is selected if you strongly disagree with the item, and 5 is chosen if you strongly agree. While you are instructed to answer every question to the best of your ability, you retain the right to leave any question unanswered should you wish to.

There will also be a brief questionnaire pertaining your opinion of the game you just played, and a demographics survey following the main questionnaire, to gather data on your game playing experience.

Questions

If you have any questions regarding this experiment, please feel free to ask them at any point to the investigator or following the experiment by contacting mta510@york.ac.uk or paul.cairns@york.ac.uk and alex.wade@york.ac.uk.

A.2 Participant Debrief Document - Placebo/Treatment Condition

PID: E ___________P1

Participant Information Sheet

Overview

This experiment will consist of completing a video game session, prior to which you will be given a tutorial which will explain in detail how to play the game. The aim of this experiment is to evaluate player experience when playing a survival video game with adaptive artificial intelligence (AI). We will also collect psychophysiological measurements in the form of eye tracking, and electrodermal activity (skin conductance) as additional data to the experiment.

What is adaptive AI?

All video games have a decision-making process that controls opponents and objects, which is called game artificial intelligence (AI).

Typical game AI controls the events and occurrences in the virtual world of the game – the number and location of opponents, strength of equipment that can be found on different levels, or even the skills that can be obtained from levelling-up; while a more effective game AI would make the gameplay more realistic by making the characters and environment inside the game able to reason effectively.

One of the possible ways to moderate the challenge levels for each person is to make the game AI adaptable to player behaviour. Dynamic game difficulty balancing involves helping players avoid getting stuck, adapt gameplay more to an individual's preference and taste, or even detect players cheating in the game. Adaptation is used to learn about a player in order to respond to the way they are playing, for example adjusting opponents' speed and accuracy in order to present a more appropriate challenge level.

Some modern video game developers create game AI capable of adapting to the player behaviour. You are about to test one of these projects yourself. In the present game, we use adaptive AI to modify the world as you play.

Procedure

During the initial practice session, you will become familiar with the gameplay and the controls of an survival video game, 'Don't Starve'. As a in typical survival game, the main character, Wilson, will have to collect and build objects in order to survive. During the trial session, your character will appear in a randomly generated world with objects that you will have to collect and monsters that you will have to avoid.

Your character has a number of needs – you will have to feed him with berries, meats, eggs and carrots, so that he doesn't die of starvation. The heart shows your character's health – as long as nothing attacks him and he doesn't starve, he will be pretty much fine. To recover his health, he can eat flowers. Lastly, the brain shows his sanity – if Wilson enters a graveyard, walks in the darkness or doesn't shave, his sanity will go down, and as a result of that, it will be more likely that he'll get attacked by imaginary monsters. However, you can pick flowers to bring his sanity back up.

The objects you collect are self-explanatory, but feel free to ask about any of them during the tutorial session. The creatures you will come across in the world can either attack you, protect you if you feed them, or become your dinner. Rabbits and birds can be caught and either eaten raw or cooked. Pigs are harmless unless you attack them – you can also befriend them if you give them meat. Most of the other creatures are likely to be deadly so it would be a good idea to run away if they spot you.

The aim of the game is to survive – that is, your result will be composed of how well you do in the game, together with how many days you last. Your character is afraid of darkness, so you better light some sort of fire before it gets dark to keep him happy, dry and warm.

Main Experiment

The main part of the study consists of one gaming round during which you will be playing the game described above. During this session the game AI will adapt to your behaviour depending on your gaming style and the choices you make in the game. Adaptive AI implemented in this game will be collecting and using the information about you as a player throughout the session, and will be learning from your behaviour as a player in order to keep the game balanced and challenging. The AI will modify aspects of the world generation such as enemies and resources you encounter. Upon completion of the session you will be asked to fill in a questionnaire about your gaming experience.

Recording Apparatus

During this experiment, several apparatus will be connected to you in order to acquire physiological information of the effects of this experiment. We will collect pupillometry using a head mounted eye tracker, which will record your eyes. We will also collect electrodermal activity, by placing two electrodes on your wrist. In order to collect these data, the experiment is required to place them on your body, with your permission.

Questionnaire

After playing the game, you will be asked to complete a player experience questionnaire. This questionnaire is comprised of 31 questions, for which you must provide answers on a scale of 1 to 5, where 1 is selected if you strongly disagree with the item, and 5 is chosen if you strongly agree. While you are instructed to answer every question to the best of your ability, you retain the right to leave any question unanswered should you wish to. There will also be a brief questionnaire pertaining your opinion of the game you just played, and a demographics survey following the main questionnaire, to gather data on your game playing experience.

Questions

If you have any questions regarding this experiment, please feel free to ask them at any point to the investigator or following the experiment by contacting mta510@york.ac.uk or paul.cairns@york.ac.uk and alex.wade@york.ac.uk.

A.3 Participant Consent Document

Informed Consent of Participation

The purpose of the form is to tell you about the study and highlight features of your participation in the research.

Who is running this?

The study is being run by Myat Aung, who is a PhD student in the Departments of Computer Science and Psychology at the University of York. The principal supervisors of this research are Dr.Paul Cairns, and Prof.Alex Wade.

What is the purpose of the study?

The aim of this study is to investigate how people experience playing digital games. To do this, some equipment will also be used to track your eyes and your skin conductance (electrodermal activity).

Confidentiality - Who will see this data?

Your results are anonymous, private, and confidential – only the researchers will see your results. They will compile the data from all participants into a secure database that will be used to analyse the data. At this point of, all data will be anonymised, and you will not be identifiable.

Right to Withdrawal - Do I have to do this?

Your participation is completely voluntary and even after signing this form, you are not required to complete the experiment if you do not want to. You can therefore withdraw from the study at any point, and if requested your data can be destroyed.

Can I ask a question?

Do ask any questions you may have about the procedure that you are about to follow. However, during the main part of the study, please refrain from talking to the experimenter, and save any non-urgent questions you may have until the end of the test. If you have any questions about the purpose or background of the experiment, please wait until the end of the experiment.

Consent

Please fill and sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that Myat will be obliged to treat your data as described.

Please circle either YES or NO as appropriate.

I, the undersigned, confirm the following:

1. I have read and understood the information provided on the Participant Information Sheet for this experiment.

   YES / NO

2. I have been given the opportunity to ask questions about the study and my participation in this study.

   YES / NO

3. I voluntarily agree to participate in this study

   YES / NO

4. I understand my right to withdrawal at any time during the experiment, without having to provide a reason and without penalty for withdrawal.

   YES / NO

5. I understand the use of data for research and publications as explained in the Participant Information Sheet

   YES / NO

6. The confidentiality of data has been explained, in particular that all data will be anonymised and I will not be identifiable by the data.

   YES / NO

Participant Name:

Participant Signature:

Name of Researcher: Myat Aung

Researcher Signature:

Date:

PID: E ____________

A.4 Immersive Experience Questionnaire

1. The game had my full attention
   - Strongly Disagree (1)
   - Disagree (2)
   - Neutral (3)
   - Agree (4)
   - Strongly Agree (5)

2. I felt focused on the game
   - Strongly Disagree (1)
   - Disagree (2)
   - Neutral (3)
   - Agree (4)
   - Strongly Agree (5)

3. I put effort into playing the game
   - Strongly Disagree (1)
   - Disagree (2)
   - Neutral (3)
   - Agree (4)
   - Strongly Agree (5)

4. I tried my best
   - Strongly Disagree (1)
   - Disagree (2)
   - Neutral (3)
   - Agree (4)
   - Strongly Agree (5)

5. I lost track of time
   - Strongly Disagree (1)
   - Disagree (2)
   - Neutral (3)
   - Agree (4)
   - Strongly Agree (5)

6. I felt consciously aware of being in the real world whilst playing
   - Strongly Disagree (1)
   - Disagree (2)
   - Neutral (3)
   - Agree (4)
   - Strongly Agree (5)

7. I forgot about my everyday concerns
    ○ Strongly Disagree (1)
    ○ Disagree (2)
    ○ Neutral (3)
    ○ Agree (4)
    ○ Strongly Agree (5)

8. I was very much aware of myself in my surroundings
    ○ Strongly Disagree (1)
    ○ Disagree (2)
    ○ Neutral (3)
    ○ Agree (4)
    ○ Strongly Agree (5)

9. I noticed events taking place around me
    ○ Strongly Disagree (1)
    ○ Disagree (2)
    ○ Neutral (3)
    ○ Agree (4)
    ○ Strongly Agree (5)

10. I felt the urge to stop playing and see what was happening around me
    ○ Strongly Disagree (1)
    ○ Disagree (2)
    ○ Neutral (3)
    ○ Agree (4)
    ○ Strongly Agree (5)

11. I felt like I was interacting with the game environment
    ○ Strongly Disagree (1)
    ○ Disagree (2)
    ○ Neutral (3)
    ○ Agree (4)
    ○ Strongly Agree (5)

12. I felt that I was separated from the real-world environment
    ○ Strongly Disagree (1)
    ○ Disagree (2)
    ○ Neutral (3)
    ○ Agree (4)
    ○ Strongly Agree (5)

13. The game was something that I was experiencing, rather than just doing
    - o Strongly Disagree (1)
    - o Disagree (2)
    - o Neutral (3)
    - o Agree (4)
    - o Strongly Agree (5)

14. The sense of being in the game environment was stronger than the sense of being in the real world
    - o Strongly Disagree (1)
    - o Disagree (2)
    - o Neutral (3)
    - o Agree (4)
    - o Strongly Agree (5)

15. I found myself so involved that I was unaware I was using controls
    - o Strongly Disagree (1)
    - o Disagree (2)
    - o Neutral (3)
    - o Agree (4)
    - o Strongly Agree (5)

16. I moved through the game according to my own will
    - o Strongly Disagree (1)
    - o Disagree (2)
    - o Neutral (3)
    - o Agree (4)
    - o Strongly Agree (5)

17. I found the game challenging
    - o Strongly Disagree (1)
    - o Disagree (2)
    - o Neutral (3)
    - o Agree (4)
    - o Strongly Agree (5)

18. There were times in the game in which I just wanted to give up
    - o Strongly Disagree (1)
    - o Disagree (2)
    - o Neutral (3)
    - o Agree (4)
    - o Strongly Agree (5)

19. I felt motivated when playing the game
    - o Strongly Disagree (1)
    - o Disagree (2)
    - o Neutral (3)
    - o Agree (4)
    - o Strongly Agree (5)

20. I found the game easy
    - o Strongly Disagree (1)
    - o Disagree (2)
    - o Neutral (3)
    - o Agree (4)
    - o Strongly Agree (5)

21. I felt that I was making progress towards the end of the game
    - o Strongly Disagree (1)
    - o Disagree (2)
    - o Neutral (3)
    - o Agree (4)
    - o Strongly Agree (5)

22. I performed well in the game
    - o Strongly Disagree (1)
    - o Disagree (2)
    - o Neutral (3)
    - o Agree (4)
    - o Strongly Agree (5)

23. I felt emotionally attached to the game
    - o Strongly Disagree (1)
    - o Disagree (2)
    - o Neutral (3)
    - o Agree (4)
    - o Strongly Agree (5)

24. I was interested in seeing how the game's events would progress
    - o Strongly Disagree (1)
    - o Disagree (2)
    - o Neutral (3)
    - o Agree (4)
    - o Strongly Agree (5)

25. I wanted to "win" the game
   - o Strongly Disagree (1)
   - o Disagree (2)
   - o Neutral (3)
   - o Agree (4)
   - o Strongly Agree (5)

26. I felt in suspense about whether or not I would do well in the game
   - o Strongly Disagree (1)
   - o Disagree (2)
   - o Neutral (3)
   - o Agree (4)
   - o Strongly Agree (5)

27. I found myself so involved that I wanted to speak to the game directly
   - o Strongly Disagree (1)
   - o Disagree (2)
   - o Neutral (3)
   - o Agree (4)
   - o Strongly Agree (5)

28. I enjoyed the graphics and the imagery
   - o Strongly Disagree (1)
   - o Disagree (2)
   - o Neutral (3)
   - o Agree (4)
   - o Strongly Agree (5)

29. I enjoyed playing the game
   - o Strongly Disagree (1)
   - o Disagree (2)
   - o Neutral (3)
   - o Agree (4)
   - o Strongly Agree (5)

30. When I stopped playing, I was disappointed that the game is over
   - o Strongly Disagree (1)
   - o Disagree (2)
   - o Neutral (3)
   - o Agree (4)
   - o Strongly Agree (5)

31. I would like to play the game again
   - o Strongly Disagree (1)
   - o Disagree (2)
   - o Neutral (3)
   - o Agree (4)
   - o Strongly Agree (5)

A.5 Perception of Adaptive AI Questionnaire

1. The game was generating content according to my behaviour in the game.
   o Strongly Disagree (1)
   o Disagree (2)
   o Neutral (3)
   o Agree (4)
   o Strongly Agree (5)

2. New content in the game appeared based on my decisions as a player.
   o Strongly Disagree (1)
   o Disagree (2)
   o Neutral (3)
   o Agree (4)
   o Strongly Agree (5)

3. The game matched the challenge to my skills and abilities as a player.
   o Strongly Disagree (1)
   o Disagree (2)
   o Neutral (3)
   o Agree (4)
   o Strongly Agree (5)

4. The behaviour of the game changed when I was doing too well or too badly.
   o Strongly Disagree (1)
   o Disagree (2)
   o Neutral (3)
   o Agree (4)
   o Strongly Agree (5)

5. The game was generating content based on the needs of my character at that point in the game.
   o Strongly Disagree (1)
   o Disagree (2)
   o Neutral (3)
   o Agree (4)
   o Strongly Agree (5)

6. The game was not responding sensibly to my actions as a player.
   o Strongly Disagree (1)
   o Disagree (2)
   o Neutral (3)
   o Agree (4)
   o Strongly Agree (5)

Appendix B

Chapter 4 Appendices - More Granular Measurements of Mental Load with Heart Rate Variability in a Rhythm Game

B.1 Participant Information Sheet

Participant Information Sheet

Thank you for agreeing to take part in our study. Please ensure that you read the details of the experiment below. If you have any questions prior to beginning the experiment, please do not hesitate to ask the investigator, ***Myat Aung***.

Overview

This experiment will consist of playing a rhythm video game called *osu!taiko*. Below are instructions on how to play, and prior to beginning the main study, you will be given a tutorial which will explain in detail how to play the game.

The aim of this experiment is to evaluate player experience and task load while playing a rhythm video game. This will be done through questionnaires as you play. We will also collect psychophysiological measurements in the form of heart rate measurements, which will require the placement of a heart rate tracker on your torso. You will have been informed of this in an information email before arriving, but if you are uncomfortable with this procedure at the last moment before starting the experiment, please don't hesitate to tell the experimenter that you wish to cancel or leave.

Procedure

In this experiment, you will play a sequence of levels from the rhythm game *osu!taiko*. Before the main experiment, you will get to watch how the game is played. Then you will be instructed how to play this game, and will practice several times before the main experiment. Between each level, you will be asked to complete a brief psychometric test on the 'task load' of the level you just played. This test will be completed digitally. In total, the experiment will not exceed 45 minutes.

Recording Apparatus

During the experiment, your heart rate will be recorded. This is achieved through the use of three electrodes, placed at specific locations on your torso. This is an important part of the experiment, but we understand if you are not comfortable with this requirement. The electrodes will be placed on you with sticky gelled patches that are easily removable after the experiment. We will also provide sanitary alcohol wipes if you wish to use any.

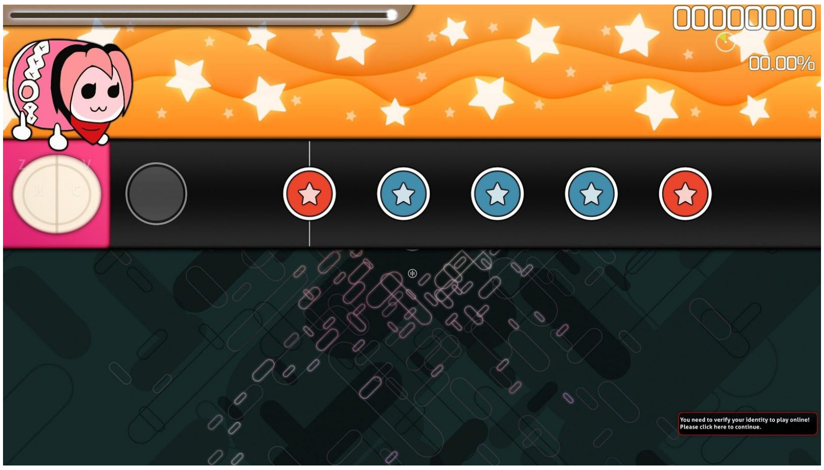
During the experiment, we will also record your heart rate using a commercial sports tracker known as the *Polar-H10*. This is a simple device that is worn on your torso with an adjustable elastic strap. More information on this device is provided below. If you wish to put this device on yourself, you are welcome to do so and will be instructed by the experimenter how to before they leave the room to ensure your privacy. If you feel uncomfortable at any point during the study, please do not hesitate to tell the experimenter or to otherwise state your intention to abort the experiment.

While you play, data will be recorded from your response to the game, as well as telemetry from each of your play sessions. Please be advised that we are not interested in your personal performance, nor do we judge you based on your ability to play the game. We collect this information to get a better understanding of how to appropriately design a game stimulus for our experiments.

How to play: *osu!taiko*

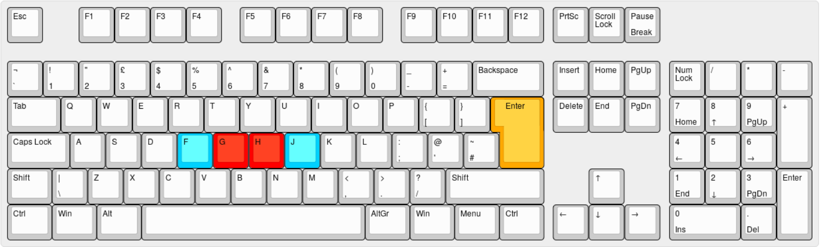
osu!taiko is a version of a popular arcade music game in Japan. In *osu!taiko* your goal is to try and play some drums in rhythm with music in the game. You do this by pressing buttons on the keyboard corresponding to 'notes' of a song that scroll across your screen. On the top left of your screen, there is a bar that represents what percentage of the notes in the song you have hit. On the top right of the screen, your score is displayed, as well as your accuracy. When you press a button, the beige drum on the left side of the screen will show which note you hit, and the drum mascot will respond. **We know that this is a lot of information**, so you will be taught how to play the game and get several chances to practice. Images below will also demonstrate the game. You will also begin by watching how the game is played before you try yourself.

As notes scroll across the conveyor, your goal is to try and time each button press to timed as accurately as possible when it enters the grey circle on the left.

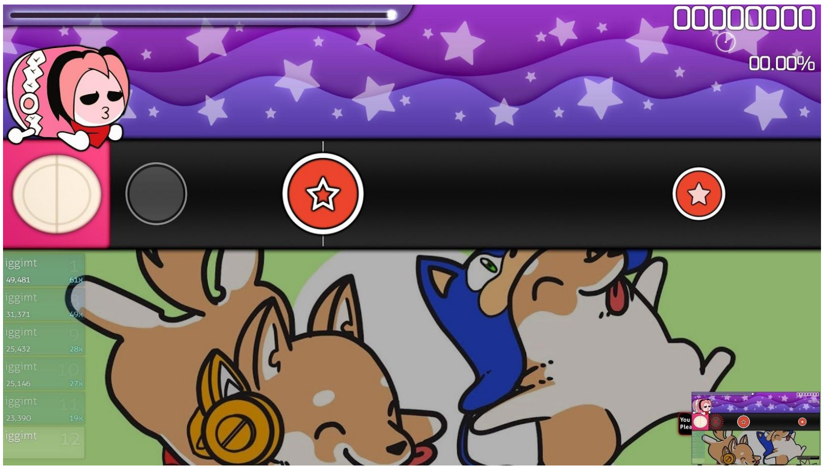


You need to verify your identity to play online!
Please click here to continue.

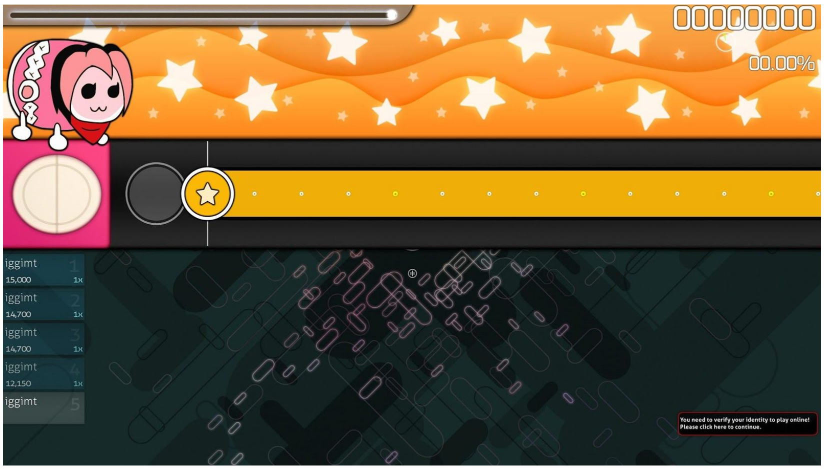
You will notice that notes come in two colours, red and blue. *osu!taiko* can be played entirely with the keyboard. There are four buttons, **F G H J**. The inner two buttons (**G H**) correspond to the red notes, hitting inside of the drum. The outer two buttons (**F J**) correspond to blue notes, or the sides of your drum. A map of the keyboard layout is shown below. How you lay your fingers out is up to you, but we at least recommend using each hand for each side of the keyboard. Before each level, the experimenter will check if you are ready to proceed. You can begin a level by pressing Enter when instructed to do so.


Esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12 PrtSc Scroll Lock Pause Break
Backspace Insert Home PgUp Num Lock
Tab Q W E R T Y U I O P Enter Delete End PgDn
Caps Lock A S D F G H J K L
Shift Z X C V B N M Shift
Ctrl Win Alt AltGr Win Menu Ctrl

Some notes will be larger. You have the option to hit both buttons of that colour for these notes, to score double points. If you want, you can also just press one button as usual, but you will not receive the extra points.

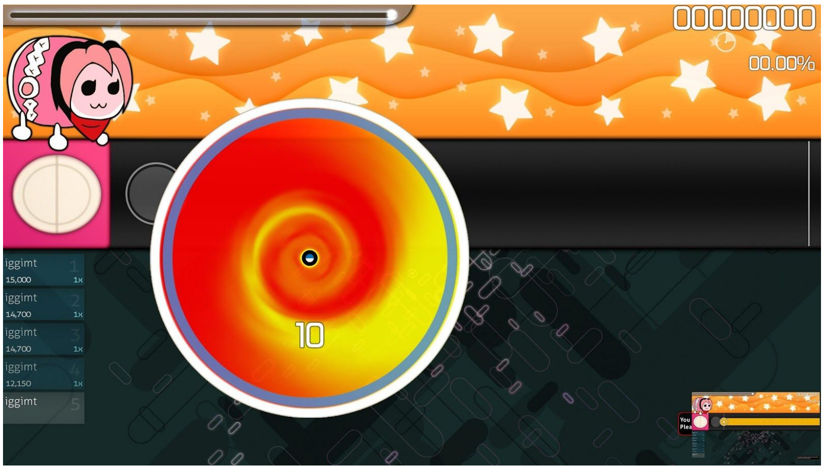


Sometimes, you will see longer notes. These represent drumrolls. Here, you should press notes of the same colour in rhythm with the song, for as long as the note is on the screen (pictured below). You can choose either red or blue, but once you press a colour, you must stick with it for the drumroll.

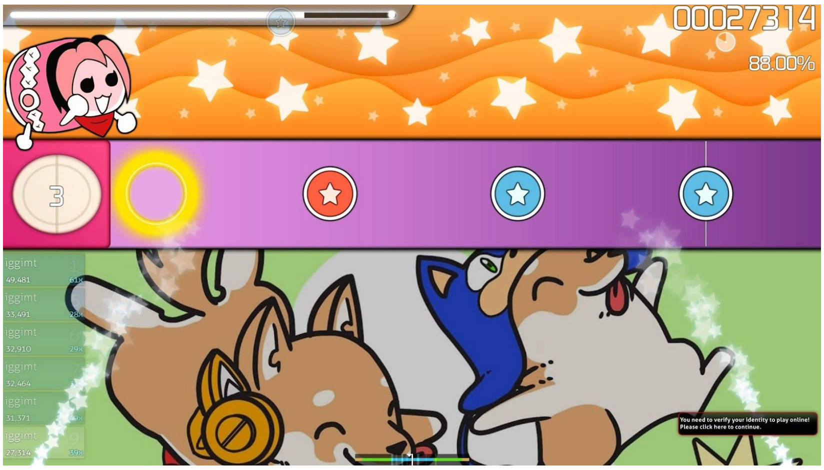


You need to verify your identity to play online!
Please click here to continue.

You will sometimes see a big multi-coloured note. For these, you must alternate buttons until the spinning circle disappears. The number tells you how many more times you must press the buttons to continue.



Finally, sometimes your game will become more colourful. In these cases, you don't have to do anything additional, but you get extra points for hitting notes.



You need to verify your identity to play online!
Please click here to continue.

This is all the information required for the game. We understand this is a lot to take in, so we will give you chances to practice and the experimenter will also instruct you how to play during an initial training level. You will also begin by observing how the game is played, before you try yourself.

In the main experiment, you will play a sequence of six levels, each of which should take about 2 minutes.

Goal

Your aim is to score as high as possible by hitting the correct notes at the correct times, minimising the number of notes that you miss, and chaining lots of hit notes together. The more notes you chain, the higher your score will be.

At the end of each level, your score will be recorded, though we <u>do not judge</u> you based on your performance, as we are only interested in your experience of playing the game. We wish to get a better understanding of how to design the difficulty of the game for our future experiments.

Questionnaires

Task Index I: Workload Test

The short workload test is comprised of two parts.

In the first part, you will simply be asked to rate six different aspects of the experimental task on a likert scale. Each workload test will not take more than a minute or two to complete, they are designed to be short and easy. We compel you to take your time considering each entry for this test.

Rating Scales

Rating scales will be described on the page every time you are asked to complete this test. They are also described below, if you wish to read them now.

Mental Demand: How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?

Physical Demand: How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slack or strenuous, restful or laborious?

Temporal Demand: How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow and leisurely, or rapid and frantic?

Overall Performance: How successful were you in performing the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?

Frustration Level: How insecure, discouraged, irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?

Task Index II: Sources of Load

In the second part of the workload test, you will be given pairs of factors from the scales above. Your task here is simply to select the option from each pair that contributed more to the workload of the experimental task (playing the level). You can do this by simply clicking the option that you want.

You will also get a chance to practice these questionnaires before the main experiment.

Questions & Concerns

If you have any questions regarding this experiment, please feel free to ask them at any point to the investigator, or if you have questions following the experiment, you can contact the researchers involved at ***mta510@york.ac.uk***, ***paul.cairns@york.ac.uk***, ***alex.wade@york.ac.uk***.

B.2 Participant Debrief Document

Participant Debrief Sheet

Thank you for participating in this study. The following is for your information, and you are welcome to take this document with you on your departure.

As described in the information sheet, we sought to measure your mental load and experience as you played *osu!taiko*. The goal of our study is to explore the possibility of existing physiological signal patterns corresponding to differences in the demand and mental load of video game play.

As we stated earlier, we are not interested in your ability to play the game. We are however interested in understanding how to design a game that people can play in a short experiment like the one you just completed. To this end, we collected your performance data to see if our game was too easy, or too hard.

In the event that you have any questions or concerns, you are urged to contact the experimenter by email (mta510@york.ac.uk), or the principal supervisors (paul.cairns@york.ac.uk, alex.wade@york.ac .uk).

Again, we thank you for your participation in this study. If you know of any friends or acquaintances that are eligible to participate in this study, we politely request that you not discuss it with them until after they have had the opportunity to participate or decline to do so.

B.3 Participant Consent Document

Informed Consent of Participation

The purpose of the form is to tell you about the study and highlight features of your participation in the research.

Who is running this?

The study is being run by *Myat Aung*, who is a PhD student in the Departments of Computer Science and Psychology at the University of York. The principal supervisors of this research are *Dr.Paul Cairns*, and *Prof.Alex Wade*.

What is the purpose of the study?

The aim of this study is to investigate how people experience playing a digital rhythm game. To do this, some equipment will also be used to track your heart rate using a commercial heart tracker (Polar-H10) that is worn on your chest with an elastic strap. If you are not comfortable with the experimenter placing this device on your body, you will be provided with instructions on how to do so yourself, and the experimenter can leave the room to give you privacy.

Confidentiality - Who will see this data?

Your results are anonymous, private, and confidential – only the researchers will see your results. They will compile the data from all participants into a secure database that will be used to analyse the data. At this point of, all data will be anonymised, and you will not be identifiable.

Right to Withdrawal - Do I have to do this?

Your participation is completely voluntary and even after signing this form, you are not required to complete the experiment if you do not want to. You can therefore withdraw from the study at any point, and if requested your data can be destroyed.

Can I ask a question?

Do ask any questions you may have about the procedure that you are about to follow. However, during the main part of the study, please refrain from talking to the experimenter, and save any non-urgent questions you may have until the end of the test. If you have any questions about the purpose or background of the experiment, please wait until the end of the experiment.

Consent

Please fill and sign below that you agree to take part in the study under the conditions laid out above. This will indicate that you have read and understood the above and that ***Myat*** will be obliged to treat your data as described.

Please circle either YES or NO as appropriate.

I, the undersigned, confirm the following:

1. I have read and understood the information provided on the Participant Information Sheet for this experiment.

YES / NO

2. I have been given the opportunity to ask questions about the study and my participation in this study.

   YES / NO

3. I voluntarily agree to participate in this study.

   YES / NO

4. I understand my right to withdrawal at any time during the experiment, without having to provide a reason and without penalty for withdrawal.

   YES / NO

5. I understand the use of data for research and publications as explained in the Participant Information Sheet

   YES / NO

6. The confidentiality of data has been explained, in particular that all data will be anonymised and I will not be identifiable by the data.

   YES / NO

Participant Name:

Participant Signature:

Name of Researcher: *Myat Aung*

Researcher Signature:

Date:

PID:

Appendix C

IEQ-SF Usage Manual

C.1 Overview

The Short Form Immersive Experience Questionnaire (IEQ-SF) is a short questionnaire designed to measure player experience following a game play session that can be completed in approximately a minute. It is a multi-dimensional scale based on a 5-point likert scale system of response.

C.2 Questionnaire Description

The IEQ-SF consists of 11 items drawn from the original IEQ Jennett et al., 2008, where items are presented in a 5-point likert scale with 1 corresponding to "Strongly Disagree" to 5 being "Strongly Agree". The questionnaire structure is split into 3 factors, Involvement, Real World Dissociation, and Challenge. This factor structure is detailed below in table C.1.

Involvement is defined as the degree of involvement from cognitive and emotional facets during game play, such as the focus exerted by the player, or their intrinsic drive to keep playing the game.

Real World Dissociation (RWD) is the describes the participant's engagement with the game or correspondingly, their disengagement from the real world. Examples might include the loss of conscious awareness of the control interface (such as a keyboard and mouse).

Challenge is the degree of difficulty experienced by the player during game play, which includes the player's perception of their own performance.

TABLE C.1: Items and corresponding factors of the IEQ-SF.

Item	Content	Factor	Scoring
1	I felt focused on the game.	Involvement	Normal
2	The game was something that I was experiencing, rather than just doing.	Involvement	Normal
3	I felt motivated when playing the game.	Involvement	Normal
4	I enjoyed playing the game.	Involvement	Normal
5	I felt consciously aware of being in the real world whilst playing.	RWD	Reversed
6	I forgot about my everyday concerns.	RWD	Normal
7	I felt that I was separated from the real-world environment.	RWD	Normal
8	I found myself so involved that I was unaware I was using controls.	RWD	Normal
9	I found the game challenging.	Challenge	Normal
10	I found the game easy.	Challenge	Reversed
11	I felt in suspense about whether or not I would do well in the game.	Challenge	Normal

C.3 Procedure

The IEQ-SF should be applied following a game play session, which may either be an entire experimental period, or a trial block which is part of a greater experimental trial sequence. Because it is intended to be rapidly completed, we strongly recommend that the IEQ-SF be included as part of a practice or training block before the main experimental trials. This is particularly important for experiments where the intent is to apply the IEQ in repeated periods, such the use of a blocking design.

Ideally, questions from the IEQ-SF should be presented to participants in a randomised order and we anticipate few situations in which a non-random order of presentation is justifiable. Due to the short length of the questionnaire, we also recommend taking additional care against careless responses in the form of an attention check. Recent evidence suggests potentially data compromising consequences due to the use of an attention check, but we confer to work by Kung et al. Kung et al., 2018 in the defense of an attention check, and recommend the Prolific model of writing attention check questions such that participants are tested on whether they have paid attention to the question, rather than any overarching instructions.

C.4 Scoring

The IEQ-SF is scored on a per-factor basis. It is encouraged to take the mean rather than the sum of the IEQ scores, as has historically been done with use of the IEQ in previously published research. The reason for this is to standardise scores from different factors, which is particularly important as one factor has fewer items than the others. Therefore, the recommended scoring of the IEQ-SF is to produce 3 mean scores, one for each factor. The overall mean score commonly used in other questionnaires (including the full IEQ) is treated as an optional computation here, and is considered to be entirely secondary to the recommended per-factor scores. For the analysis of IEQ-SF scores, we still recommend the testing of scores per factor rather than a single overall score, due to the possibility of a null-effect in a single factor overpowering remaining factors/items in a result.

The table C.1 includes a column indicating whether a question is reverse scored or not. Like the full IEQ and other similar questionnaires, the IEQ-SF includes some reverse scored items. Here, the researcher should reverse score questions IEQ-SF questions 5 ("I Felt consciously aware of being in the real world whilst playing."), and question 10 ("I found the game easy."), prior to computing any factor-wise average scores.

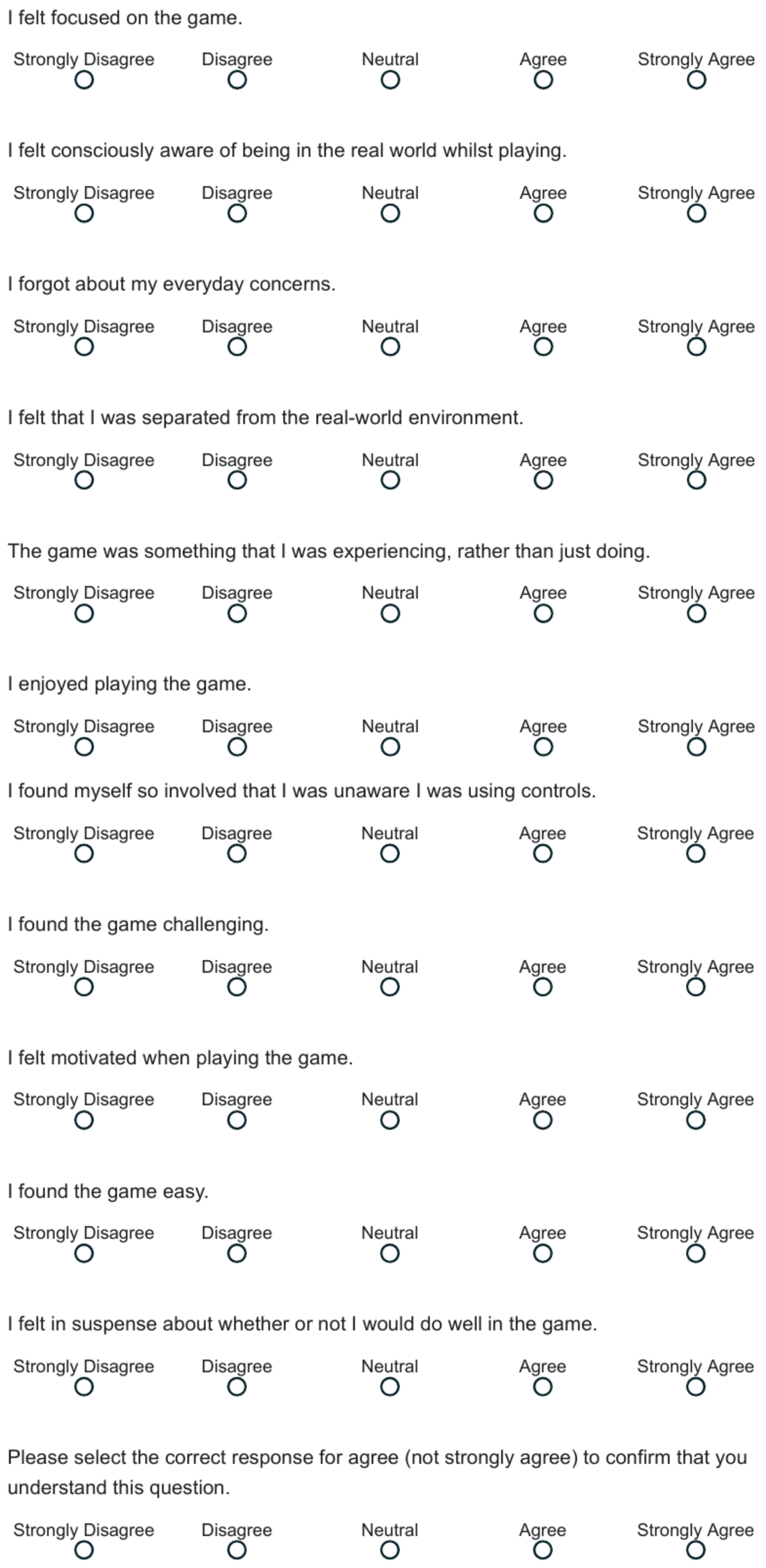
C.5 Participant Instructions

An example excerpt of a participant information sheet including instructions for completing the IEQ-SF following an initial practice trial is provided below.

″At the end of this trial session, you will also be presented with the immersive experience questionnaire so that you are familiarised with the process of filling the short questionnaire. The questionnaire consists of 11 questions for which you must provide answers on a scale of 1 to 5, where 1 is selected if you strongly disagree with the item, and 5 is chosen if you strongly agree. You will be asked to fill this again after [the/each] experimental task, so we recommend that you take your time now to acquaint yourself with the questionnaire.″

C.6 Example from a digital questionnaire platform

We present below images of a digitised form of the IEQ-SF on the online questionnaire platform Qualtrics, which is also sometimes used in a lab setting as part of an experimental session.



I felt focused on the game.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

I felt consciously aware of being in the real world whilst playing.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

I forgot about my everyday concerns.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

I felt that I was separated from the real-world environment.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

The game was something that I was experiencing, rather than just doing.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

I enjoyed playing the game.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

I found myself so involved that I was unaware I was using controls.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

I found the game challenging.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

I felt motivated when playing the game.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

I found the game easy.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

I felt in suspense about whether or not I would do well in the game.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

Please select the correct response for agree (not strongly agree) to confirm that you understand this question.

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
○	○	○	○	○

Appendix D

Chapter 4 Appendices - The Short Form IEQ

D.1 Full Exploratory Factor Analysis Results

	1F-F1	2F-F1	2F-F2	3F-F1	3F-F2	3F-F3	4F-F1	4F-F2	4F-F3	4F-F4	5F-F1	5F-F2	5F-F3	5F-F4	5F-F5
IEQ1	0.72	-0.58	-0.30	-0.50	-0.32	0.17	**-0.56**	**-0.35**	**0.11**	**-0.02**	**-0.23**	**-0.35**	**0.09**	**-0.01**	**0.41**
IEQ2	0.76	-0.73	-0.13	-0.65	-0.16	0.18	-0.72	-0.20	0.12	0.00	-0.28	-0.21	0.08	0.02	0.54
IEQ3	0.62	-0.68	0.02	-0.58	0.01	0.27	-0.63	-0.01	0.22	0.01	-0.13	0.00	0.10	-0.06	0.61
IEQ4	0.54	-0.66	0.11	-0.56	0.11	0.27	-0.66	0.06	0.20	0.07	-0.13	0.07	0.11	-0.01	0.61
IEQ5	0.52	-0.29	-0.42	-0.25	-0.44	0.07	-0.13	-0.37	0.09	-0.26	-0.17	-0.36	0.06	-0.22	0.07
IEQ6	0.38	0.07	-0.80	0.09	-0.79	-0.01	0.03	-0.79	-0.03	-0.03	0.03	-0.80	0.00	0.00	-0.04
IEQ7	0.53	-0.30	-0.43	-0.28	-0.46	0.00	-0.23	-0.42	-0.01	-0.16	-0.24	-0.44	0.02	-0.05	0.08
IEQ8	0.25	0.18	-0.73	0.21	-0.71	0.04	0.08	-0.74	0.01	0.08	0.12	-0.75	0.06	0.06	-0.07
IEQ9	0.14	0.22	-0.59	0.24	-0.56	0.08	0.00	-0.65	0.02	0.28	0.22	-0.65	0.04	0.17	0.04
IEQ10	0.29	-0.14	-0.29	-0.13	-0.29	0.00	-0.26	-0.34	-0.05	0.13	-0.12	-0.39	0.02	0.21	0.11
IEQ11	0.54	-0.48	-0.16	-0.52	-0.22	-0.15	-0.16	-0.08	-0.08	-0.54	-0.59	-0.12	-0.01	-0.22	-0.13
IEQ12	0.54	-0.17	-0.68	-0.15	-0.70	-0.02	-0.06	-0.64	0.00	-0.23	-0.08	-0.63	-0.04	-0.20	0.08
IEQ13	0.66	-0.59	-0.17	-0.57	-0.21	0.01	-0.27	-0.09	0.06	-0.50	-0.57	-0.11	0.10	-0.24	-0.01
IEQ14	0.56	-0.22	-0.62	-0.21	-0.66	-0.08	-0.05	-0.56	-0.04	-0.34	-0.16	-0.55	-0.08	-0.27	0.03
IEQ15	0.48	-0.21	-0.49	-0.19	-0.51	0.00	-0.01	-0.42	0.04	-0.33	-0.12	-0.40	-0.02	-0.29	0.03
IEQ16	0.35	-0.43	0.09	-0.48	0.06	-0.12	-0.33	0.10	-0.11	-0.22	-0.46	0.05	-0.03	0.03	0.05
IEQ17	0.36	-0.38	0.00	-0.13	0.07	0.83	-0.19	0.06	0.80	0.04	-0.06	0.03	0.79	0.03	0.16
IEQ18	0.10	-0.15	0.05	-0.32	0.02	-0.47	**-0.37**	**-0.03**	**-0.51**	**0.08**	**-0.47**	**-0.12**	**-0.31**	**0.39**	**-0.06**
IEQ19	0.73	-0.79	0.00	-0.75	-0.04	0.09	-0.62	-0.01	0.08	-0.23	**-0.46**	**-0.03**	**0.06**	**-0.09**	**0.38**
IEQ20	0.26	-0.21	-0.08	0.04	-0.02	0.83	0.01	-0.01	0.83	0.00	-0.06	-0.04	0.89	0.03	-0.06
IEQ21	0.41	-0.52	0.11	-0.52	0.07	-0.03	-0.41	0.09	-0.04	-0.17	-0.30	0.08	-0.06	-0.07	0.27
IEQ22	0.29	-0.41	0.13	-0.53	0.09	-0.34	**-0.58**	**0.03**	**-0.40**	**0.07**	**-0.17**	**0.05**	**-0.45**	**0.06**	**0.48**
IEQ23	0.56	-0.47	-0.20	-0.46	-0.24	-0.02	-0.17	-0.13	0.02	-0.46	**-0.35**	**-0.11**	**-0.03**	**-0.36**	**0.07**
IEQ24	0.53	-0.58	0.03	-0.55	-0.02	0.04	-0.28	0.09	0.08	-0.43	-0.51	0.07	0.10	-0.19	0.04
IEQ25	0.34	-0.34	-0.03	-0.21	-0.02	0.37	-0.35	-0.07	0.32	0.14	0.28	-0.01	0.12	-0.12	0.63
IEQ26	0.33	-0.25	-0.15	-0.06	-0.12	0.54	-0.01	-0.08	0.56	-0.12	0.15	-0.04	0.40	-0.26	0.23
IEQ27	0.47	-0.31	-0.31	-0.27	-0.33	0.04	0.06	-0.18	0.11	-0.53	-0.09	-0.11	-0.06	-0.62	0.07
IEQ28	0.44	-0.53	0.08	-0.57	0.05	-0.10	-0.36	0.11	-0.08	-0.32	-0.56	0.07	-0.01	-0.04	0.05
IEQ29	0.67	-0.83	0.16	-0.85	0.12	-0.05	-0.71	0.15	-0.06	-0.24	-0.76	0.09	0.06	0.09	0.20
IEQ30	0.48	-0.34	-0.27	-0.31	-0.28	0.07	-0.20	-0.26	0.06	-0.22	-0.21	-0.27	0.06	-0.17	0.09
IEQ31	0.59	-0.66	-0.04	-0.64	-0.07	0.05	-0.52	-0.02	0.02	-0.24	-0.53	-0.08	0.11	-0.04	0.18

Bibliography

Abeele, V. V., Spiel, K., Nacke, L., Johnson, D., & Gerling, K. (2020). Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. *International Journal of Human-Computer Studies*, *135*, 102370. https://doi.org/10.1016/j.ijhcs.2019.102370

Ambinder, M. (2011). Biofeedback in Gameplay: How Valve Measures Physiology to Enhance Gaming Experience, 71.

Ang, C. S., Zaphiris, P., & Mahmood, S. (2007). A model of cognitive loads in massively multiplayer online role playing games. *Interacting with Computers*, *19*(2), 167–179. https://doi.org/10.1016/j.intcom.2006.08.006

Appelhans, B. M., & Luecken, L. J. (2006). Heart rate variability as an index of regulated emotional responding. *Review of General Psychology*, *10*(3), 229–240. https://doi.org/10.1037/1089-2680.10.3.229

Aung, M., Cairns, P., & Cutting, J. (2021a). A Short Form Immersive Experience Questionnaire [Publisher: OSF]. https://doi.org/10.17605/OSF.IO/BF8QN

Aung, M., Cairns, P., & Cutting, J. (2021b). Confirmatory Experiment for The Short Form Immersive Experience Questionnaire [Publisher: OSF]. https://doi.org/10.17605/OSF.IO/3PGV2

Bartle, R. A. From MUDs to MMORPGs: The History of Virtual Worlds (J. Hunsinger, L. Klastrup, & M. Allen, Eds.). en. In: In *International Handbook of Internet Research* (J. Hunsinger, L. Klastrup, & M. Allen, Eds.). Ed. by Hunsinger, J., Klastrup, L., & Allen, M. Dordrecht: Springer Netherlands, 2009, pp. 23–39. ISBN: 978-1-4020-9788-1 978-1-4020-9789-8. https://doi.org/10.1007/978-1-4020-9789-8_2.

Beatty, J. (1976). *Pupillometric measurement of cognitive workload: (506152009-012)* (tech. rep.) [type: dataset]. American Psychological Association. https://doi.org/10.1037/e506152009-012

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin*, *91*(2), 276.

Bian, Y., Yang, C., Gao, F., Li, H., Zhou, S., Li, H., Sun, X., & Meng, X. (2016). A framework for physiological indicators of flow in VR games: Construction and preliminary evaluation. *Personal and Ubiquitous Computing*, *20*(5), 821–832. https://doi.org/10.1007/s00779-016-0953-5

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences*, *12*(3), 291–294. https://doi.org/10.1016/0191-8869(91)90115-R

Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., & Pidruzny, J. N. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, *45*(4), 624–634. https://doi.org/10.1016/j.jesp.2009.02.016

Brown, E., & Cairns, P. A grounded investigation of game immersion. In: *CHI'04 extended abstracts on Human factors in computing systems*. ACM, 2004, 1297–1300.

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. (2007, March). *Handbook of Psychophysiology* [Google-Books-ID: E7hRKwVBXb4C]. Cambridge University Press.

Cairns, P. Can Games Be More Than Fun? (M. Blythe & A. Monk, Eds.) [Series Title: Human–Computer Interaction Series]. en. In: In *Funology 2* (M. Blythe & A. Monk, Eds.). Ed. by Blythe, M., & Monk, A. Series Title: Human–Computer Interaction Series. Cham: Springer International Publishing, 2018, pp. 33–46. ISBN: 978-3-319-68212-9 978-3-319-68213-6. https://doi.org/10.1007/978-3-319-68213-6_3.

Cairns, P. (2019). *Doing better statistics in human-computer interaction*.

Cairns, P., Cox, A., & Nordin, A. I. (2014a). Immersion in digital games: Review of gaming experience research. *Handbook of digital games*, *1*, 767.

Cairns, P., Cox, A. L., Day, M., Martin, H., & Perryman, T. (2013). Who but not where: The effect of social play on immersion in digital games. *International Journal of Human-Computer Studies*, *71*(11), 1069–1077. https://doi.org/10.1016/j.ijhcs.2013.08.015

Cairns, P., Li, J., Wang, W., & Nordin, A. I. The influence of controllers on immersion in mobile games. en. In: ACM Press, 2014, 371–380. ISBN: 978-1-4503-2473-1. https://doi.org/10.1145/2556288.2557345.

Carey, K., Saltz, E., Rosenbloom, J., Micheli, M., Choi, J. O., & Hammer, J. Toward Measuring Empathy in Virtual Reality. en. In: *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play*. Amsterdam The Netherlands: ACM, 2017, October, 551–559. ISBN: 978-1-4503-5111-9. https://doi.org/10.1145/3130859.3131325.

Carreiras, C., Alves, A. P., Lourenço, A., Canento, F., Silva, H., Fred, A., & others. (2015). BioSPPy: Biosignal Processing in Python. https://github.com/PIA-Group/BioSPPy/

Carter, B. T., & Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology*, *155*, 49–62. https://doi.org/10.1016/j.ijpsycho.2020.05.010

Chalmers, R. P. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment [Number: 1]. *Journal of Statistical Software*, *48*(1), 1–29. https://doi.org/10.18637/jss.v048.i06

Chanel, G., Kivikangas, J. M., & Ravaja, N. (2012). Physiological compliance for social gaming analysis: Cooperative versus competitive play. *Interacting with Computers*, *24*(4), 306–316. https://doi.org/10.1016/j.intcom.2012.04.012

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM*, *50*(4), 31. https://doi.org/10.1145/1232743.1232769

Connor, A. M., Greig, T. J., & Kruse, J. (2017). Evaluating the Impact of Procedurally Generated Content on Game Immersion. *The Computer Games Journal*, *6*(4), 209–225. https://doi.org/10.1007/s40869-017-0043-6

Cowan, M. J. (1995). Measurement of Heart Rate Variability [Publisher: SAGE Publications Inc]. *Western Journal of Nursing Research*, *17*(1), 32–48. https://doi.org/10.1177/019394599501700104

Cox, A. L., Cairns, P., Berthouze, N., & Jennett, C. The use of eyetracking for measuring immersion. In: *Workshop on What have eye movements told us so far, and what is next*. 2006, 26–29.

Csikszentmihalyi, M. Flow: The Psychology of Optimal Experience. In: 1990, January.

Cutting, J., Cairns, P., & Kuhn, G. (2020). Nothing else matters: Video games create sustained attentional selection away from task-irrelevant features. *Attention, Perception, & Psychophysics*, *82*(8), 3907–3919. https://doi.org/10.3758/s13414-020-02122-y

Cutting, J. T. (2018). Measuring the experience of playing self-paced games, 230.

Denisova, A. (2016). Adaptive Technologies in Digital Games: The Influence of Perception of Adaptivity on Immersion, 285.

Denisova, A., & Cairns, P. Adaptation in Digital Games: The Effect of Challenge Adjustment on Player Performance and Experience. en. In: ACM Press, 2015, 97–101. ISBN: 978-1-4503-3466-2. https://doi.org/10.1145/2793107.2793141.

Denisova, A., & Cairns, P. The Placebo Effect in Digital Games: Phantom Perception of Adaptive Artificial Intelligence. en. In: ACM Press, 2015, 23–33. ISBN: 978-1-4503-3466-2. https://doi.org/10.1145/2793107.2793109.

Denisova, A., Cairns, P., Guckelsberger, C., & Zendle, D. (2020). Measuring perceived challenge in digital games: Development & validation of the challenge originating from recent gameplay interaction scale (CORGIS). *International Journal of Human-Computer Studies*, *137*, 102383. https://doi.org/10.1016/j.ijhcs.2019.102383

Dirican, A. C., & Göktürk, M. (2011). Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science*, *3*, 1361–1367. https://doi.org/10.1016/j.procs.2011.01.016

Drachen, A., Nacke, L. E., Yannakakis, G., & Pedersen, A. L. Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. In: *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*. ACM, 2010, 49–54.

El-Amrawy, F., & Nounou, M. I. (2015). Are Currently Available Wearable Devices for Activity Tracking and Heart Rate Monitoring Accurate, Precise, and Medically Beneficial? *Healthcare Informatics Research*, *21*(4), 315–320. https://doi.org/10.4258/hir.2015.21.4.315

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*(7), 591–601. https://doi.org/10.1037/0003-066X.63.7.591

ESA. (2017). 2017 Essential Facts About the Computer and Video Game Industry. Retrieved June 25, 2018, from http://www.theesa.com/article/2017-essential-facts-computer-video-game-industry/

Fekr, A. R., Radecka, K., & Zilic, Z. Tidal volume variability and respiration rate estimation using a wearable accelerometer sensor. In: *2014 4th International Conference on Wireless Mobile Communication and Healthcare - Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*. 2014, November, 1–6. https://doi.org/10.1109/MOBIHEALTH.2014.7015894.

Fridman, L., Reimer, B., Mehler, B., & Freeman, W. T. Cognitive Load Estimation in the Wild. en. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. Montreal QC, Canada: ACM Press, 2018, 1–9. ISBN: 978-1-4503-5620-6. https://doi.org/10.1145/3173574.3174226.

GDC. (2011, July). GDC '08: Are casual games the future? - News at GameSpot. Retrieved July 27, 2018, from https://web.archive.org/web/20110711072428/http://uk.gamespot.com/news/6186207.html?tag=result%3Btitle%3B0

Gillinov, S., Etiwy, M., Wang, R., Blackburn, G., Phelan, D., Gillinov, A. M., Houghtaling, P., Javadikasgari, H., & Desai, M. Y. (2017). Variable Accuracy of Wearable Heart Rate Monitors during Aerobic Exercise. *Medicine and science in sports and exercise*, *49*(8), 1697–1703. https://doi.org/10.1249/MSS.0000000000001284

Gomes, P., Margaritoff, P., & Silva, H. pyHRV: Development and evaluation of an open-source python toolbox for heart rate variability (HRV). In: *Proc. Int'l Conf. on Electrical, Electronic and Computing Engineering (IcETRAN)*. 2019, 822–828.

Gow, J., Cairns, P., Colton, S., Miller, P., & Baumgarten, R. Capturing Player Experience with Post-Game Commentaries. en. In: Global Science & Technology Forum (GSTF), 2010, April. https://doi.org/10.5176/978-981-08-5480-5_085.

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations [_eprint: https://onlinelibrary.wiley.com/ 8986.1996.tb01071.x]. *Psychophysiology*, *33*(4), 457–461. https://doi.org/10.1111/j.1469-8986.1996.tb01071.x

Greitzer, F. L., Kuchar, O. A., & Huston, K. (2007). Cognitive science implications for enhancing training effectiveness in a serious gaming context. *Journal on Educational Resources in Computing*, *7*(3), 2–es. https://doi.org/10.1145/1281320.1281322

Grodal, T. (2000). Video games and the pleasures of control.

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later [Publisher: SAGE Publications Inc]. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. en. In: *Advances in Psychology*. Vol. 52. Elsevier, 1988, pp. 139–183. ISBN: 978-0-444-70388-0. https://doi.org/10.1016/S0166-4115(08)62386-9.

Haverkamp, N., & Beauducel, A. (2017). Violation of the Sphericity Assumption and Its Effect on Type-I Error Rates in Repeated Measures ANOVA and Multi-Level Linear Models (MLM). *Frontiers in Psychology*, *8*. Retrieved March 27, 2023, from https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01841

Heathers, J. A. J. Methodological Improvements in Heart Rate Variability, 154.

Hudson, M., & Cairns, P. (2016). The effects of winning and losing on social presence in team-based digital games. *Computers in Human Behavior*, *60*, 1–12. https://doi.org/10.1016/j.chb.2016.02.001

Iacovides, I., Cox, A., Kennedy, R., Cairns, P., & Jennett, C. Removing the HUD: The Impact of Non-Diegetic Game Elements and Expertise on Player Involvement. en. In: ACM Press, 2015, 13–22. ISBN: 978-1-4503-3466-2. https://doi.org/10.1145/2793107.2793120.

IJsselsteijn, W. A., Ridder, H. d., Freeman, J., & Avons, S. E. Presence: Concept, determinants, and measurement. In: *Human Vision and Electronic Imaging V*. *3959*. International Society for Optics; Photonics, 2000, June, 520–530. https://doi.org/10.1117/12.387188.

Ivarsson, M., Anderson, M., Åkerstedt, T., & Lindblad, F. (2013). The Effect of Violent and Nonviolent Video Games on Heart Rate Variability, Sleep, and Emotions in Adolescents With Different Violent Gaming Habits. *Psychosomatic Medicine*, *75*(4), 390–396. https://doi.org/10.1097/PSY.0b013e3182906a4c

Jainta, S., & Baccino, T. (2010). Analyzing the pupil response due to increased cognitive demand: An independent component analysis study. *International Journal of Psychophysiology*, *77*(1), 1–7. https://doi.org/10.1016/j.ijpsycho.2010.03.008

Jennett, C., Cox, A. L., & Cairns, P. Investigating computer game immersion and the component real world dissociation. en. In: ACM Press, 2009, 3407. ISBN: 978-1-60558-247-4. https://doi.org/10.1145/1520340.1520494.

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies*, *66*(9), 641–661. https://doi.org/10.1016/j.ijhcs.2008.04.004

Jennett, C. I. (2010). *Is game immersion just another form of selective attention? An empirical investigation of real world dissociation in computer game immersion* (PhD Thesis). UCL (University College London).

Jerčič, P., Sennersten, C., & Lindley, C. The effect of cognitive load on physiological arousal in a decision-making serious game [ISSN: 2474-0489]. In: *2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*. ISSN: 2474-0489. 2017, September, 153–156. https://doi.org/10.1109/VS-GAMES.2017.8056587.

Kassner, M., Patera, W., & Bulling, A. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. UbiComp '14 Adjunct. New York, NY, USA: ACM, 2014, 1151–1160. ISBN: 978-1-4503-3047-3. https://doi.org/10.1145/2638728.2641695.

Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., & Ravaja, N. (2011). A review of the use of psychophysiological methods in game research. *Journal of Gaming & Virtual Worlds*, *3*(3), 181–199. https://doi.org/10.1386/jgvw.3.3.181_1

Kline, P. (2000). *A psychometrics primer* [OCLC: 44724310]. Free Association Books.

Kline, P. (2014, January). *The New Psychometrics: Science, Psychology and Measurement* [Google-Books-ID: 9BisAgAAQBAJ]. Routledge.

Konstan, J. A., & Riedl, J. (2012). Recommender systems: From algorithms to user experience. *User Modeling and User-Adapted Interaction*, *22*(1), 101–123. https://doi.org/10.1007/s11257-011-9112-x

Kumari, S., Power, C., & Cairns, P. Investigating Uncertainty in Digital Games and its Impact on Player Immersion. en. In: ACM Press, 2017, 503–509. ISBN: 978-1-4503-5111-9. https://doi.org/10.1145/3130859.3131311.

Kung, F. Y. H., Kwok, N., & Brown, D. J. (2018). Are Attention Check Questions a Threat to Scale Validity? [_eprint: https://iaap-journals.onlinelibrary.wiley.com *Applied Psychology*, *67*(2), 264–283. https://doi.org/https://doi.org/10.1111/apps.12108

Lang, P. J. (1995). Studies of Motivation and Attention. *American Psychologist*.

Li, Y., & Huang, D. (2009). Pupil Size and Iris Thickness Difference Between Asians and Caucasians Measured by Optical Coherence Tomography. *Investigative Ophthalmology & Visual Science*, *50*(13), 5785.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, *4*(1), 6–14. https://doi.org/10.1016/S1364-6613(99)01418-7

Matias Kivikangas, J., Nacke, L., & Ravaja, N. (2011). Developing a triangulation system for digital game events, observational video, and psychophysiological data to study emotional responses to a virtual character. *Entertainment Computing*, *2*(1), 11–16. https://doi.org/10.1016/j.entcom.2011.03.006

McCall, C., Hildebrandt, L. K., Bornemann, B., & Singer, T. (2015). Physiophenomenology in retrospect: Memory reliably reflects physiological arousal during a prior threatening experience. *Consciousness and Cognition*, *38*, 60–70. https://doi.org/10.1016/j.concog.2015.09.011

McMahan, R. P., Ragan, E. D., Leal, A., Beaton, R. J., & Bowman, D. A. (2011). Considerations for the use of commercial video games in controlled experiments. *Entertainment Computing*, *2*(1), 3–9. https://doi.org/10.1016/j.entcom.2011.03.002

Mendiburu, F. D., & Simon, R. (2015, September). *Agricolae - Ten years of an open source statistical tool for experiments in breeding, agriculture and biology* (tech. rep. e1748) [ISSN: 2167-9843]. PeerJ Inc. https://doi.org/10.7287/peerj.preprints.1404v1

Michailidis, L., Balaguer-Ballester, E., & He, X. (2018). Flow and Immersion in Video Games: The Aftermath of a Conceptual Challenge. *Frontiers in Psychology*, *9*. Retrieved May 23, 2022, from https://www.frontiersin.org/article/10.3389/fpsyg.2018.01682

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review*, *5*(2), 119–124. https://doi.org/10.1177/1754073912468165

Mäyrä, F., & Ermi, L. (2005). Fundamental components of the gameplay experience : Analysing immersion, 28.

Nacke, L., & Lindley, C. (2008, November). *Flow and immersion in first-person shooters: Measuring the player's gameplay experience* [Pages: 88]. https://doi.org/10.1145/1496984.1496998

Nacke, L. E., Bateman, C., & Mandryk, R. L. (2014). BrainHex: A neurobiological gamer typology survey. *Entertainment Computing*, *5*(1), 55–62. https://doi.org/10.1016/j.entcom.2013.06.002

Nacke, L. E., Grimshaw, M. N., & Lindley, C. A. (2010). More than a feeling: Measurement of sonic user experience and psychophysiology in a first-person shooter game. *Interacting with Computers*, *22*(5), 336–343. https://doi.org/10.1016/j.intcom.2010.04.005

Nordin, A. I., Cairns, P., & Hudson, M. (2014). The Effect Of Surroundings On Gaming Experience, 8.

Nordin, A. (2014). *Immersion And Players' Time Perception in Digital Games* (PhD Thesis). University of York.

Norman, K. L. (2013). GEQ (Game Engagement/Experience Questionnaire): A Review of Two Papers. *Interacting with Computers*, *25*(4), 278–283. https://doi.org/10.1093/iwc/iwt009

Nourbakhsh, N., Chen, F., Wang, Y., & Calvo, R. A. (2017). Detecting Users' Cognitive Load by Galvanic Skin Response with Affective Interference. *ACM Transactions on Interactive Intelligent Systems*, *7*(3), 1–20. https://doi.org/10.1145/2960413

Nourbakhsh, N., Wang, Y., & Chen, F. GSR and Blink Features for Cognitive Load Classification (D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler, Eds.). en. In: In *Human-Computer Interaction – INTERACT 2013* (D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler, Eds.). Ed. by Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., & Winckler, M. Vol. 8117. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 159–166. ISBN: 978-3-642-40482-5 978-3-642-40483-2. https://doi.org/10.1007/978-3-642-40483-2_11.

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. en. In: ACM Press, 2012, 420–423. ISBN: 978-1-4503-1438-1. https://doi.org/10.1145/2414536.2414602.

Orchard, L. N., & Stern, J. A. (1991). Blinks as an index of cognitive activity during reading. *Integrative Physiological and Behavioral Science*, *26*(2), 108–116. https://doi.org/10.1007/BF02691032

Palinko, O., & Kun, A. L. Exploring the effects of visual cognitive load and illumination on pupil diameter in driving simulators. en. In: ACM Press, 2012, 413. ISBN: 978-1-4503-1221-9. https://doi.org/10.1145/2168556.2168650.

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

Peavler, W. S. Pupil Size, Information Overload, and Performance Differences. *Psychophysiology*, *11*(5), 559–566. https://doi.org/10.1111/j.1469-8986.1974.tb01114.x

Perrett, F. (2018). The Challenge with Uncertainty and Immersion: Investigating the Relationship Immersion and Uncertainty, 59.

Plass, J. L., & Kalyuga, S. (2019). Four Ways of Considering Emotion in Cognitive Load Theory. *Educational Psychology Review*, *31*(2), 339–359. https://doi.org/10.1007/s10648-019-09473-5

Plass, J. L., & Kaplan, U. Emotional Design in Digital Media for Learning. en. In: *Emotions, Technology, Design, and Learning*. Elsevier, 2016, pp. 131–161. ISBN: 978-0-12-801856-9. https://doi.org/10.1016/B978-0-12-801856-9.00007-4.

Poock, G. K. (1973). Information Processing vs Pupil Diameter. *Perceptual and Motor Skills*, *37*(3), 1000–1002. https://doi.org/10.1177/003151257303700363

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–734. https://doi.org/10.1017/S0954579405050340

Power, C., Cairns, P., Denisova, A., Papaioannou, T., & Gultrom, R. (2018). Lost at the Edge of Uncertainty: Measuring Player Experience in Digital Games. *International Journal of Human-Computer Interaction*, *35*. https://doi.org/10.1080/10447318.2018.1507161

Power, C., Denisova, A., Papaioannou, T., & Cairns, P. Measuring Uncertainty in Games: Design and Preliminary Validation. In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing*

Systems. CHI EA '17. New York, NY, USA: Association for Computing Machinery, 2017, May, 2839–2845. ISBN: 978-1-4503-4656-6. https://doi.org/10.1145/3027063.3053215.

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A motivational model of video game engagement. *Review of General Psychology*, *14*(2), 154–166. https://doi.org/10.1037/a0019440

Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A.-S., McNamara, J. O., & Williams, S. M. (2001). Types of Eye Movements and Their Functions [Publisher: Sinauer Associates]. *Neuroscience. 2nd edition*. Retrieved April 15, 2023, from https://www.ncbi.nlm.nih.gov/books/NBK10991/

Quintana, D. S., & Heathers, J. A. J. (2014). Considerations in the assessment of heart rate variability in biobehavioral research. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00805

Ramsey, P. H. (1980). Exact Type 1 Error Rates for Robustness of Student's t Test with Unequal Variances [Publisher: American Educational Research Association]. *Journal of Educational Statistics*, *5*(4), 337–349. https://doi.org/10.3102/10769986005004337

Ravaja, N., & Saari, T. (2004). Emotional Response Patterns and Sense of Presence during Video Games: Potential Criterion Variables for Game Design, 9.

Ravaja, N., Saari, T., Laarni, J., Kallinen, K., Salminen, M., Holopainen, J., & Järvinen, A. (2005). The Psychophysiology of Video Gaming: Phasic Emotional Responses to Game Events, 14.

Ravaja, N., Saari, T., Salminen, M., Laarni, J., & Kallinen, K. (2006). Phasic Emotional Reactions to Video Game Events: A Psychophysiological Investigation. *Media Psychology*, *8*(4), 343–367. https://doi.org/10.1207/s1532785xmep0804_2

Renshaw, T., Stevens, R., & Denton, P. D. (2009). Towards understanding engagement in games: An eye-tracking study. *On the Horizon*, *17*(4), 408–420. https://doi.org/10.1108/10748120910998425

Revelle, W. (2020, December). Psych: Procedures for Psychological, Psychometric, and Personality Research. Retrieved March 10, 2021, from https://CRAN.R-project.org/package=psych

Revelle, W., & Zinbarg, R. E. (2009). Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The Motivational Pull of Video Games: A Self-Determination Theory Approach. *Motivation and Emotion*, *30*(4), 344–360. https://doi.org/10.1007/s11031-006-9051-8

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the t test to departures from population normality [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *111*, 352–360. https://doi.org/10.1037/0033-2909.111.2.352

Schachter, S. The Interaction of Cognitive and Physiological Determinants of Emotional State11Much of the research described in this paper was supported by Grant MH 05203 from the National Institute of Mental Health, United States Public Health Service, and by Grant G 23758 from the National Science Foundation. (L. Berkowitz, Ed.). en. In: In *Advances in Experimental Social Psychology* (L. Berkowitz, Ed.). Ed. by Berkowitz, L. Vol. 1. Academic Press, 1964, January, pp. 49–80. https://doi.org/10.1016/S0065-2601(08)60048-9.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state [Place: US Publisher: American Psychological Association]. *Psychological Review*, *69*, 379–399. https://doi.org/10.1037/h0046234

Schwalm, M., Keinath, A., & Zimmer, H. D. (2008). Pupillometry as a method for measuring mental workload within a simulated driving task, 14.

Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology*, *45*(5), 679–687. https://doi.org/10.1111/j.1469-8986.2008.00681.x

Sirois, S., & Brisson, J. (2014). Pupillometry. *WIREs Cognitive Science*, *5*(6), 679–692. https://doi.org/10.1002/wcs.1323

Slater, M. (2003). A note on presence terminology. *Presence connect*.

Slater, M., & Wilbur, S. (1997). A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, *6*(6), 603–616. https://doi.org/10.1162/pres.1997.6.6.603

Stern, J. A., & Skelly, J. J. (1984). The Eye Blink and Workload Considerations [Publisher: SAGE Publications]. *Proceedings of the Human Factors Society Annual Meeting*, *28*(11), 942–944. https://doi.org/10.1177/154193128402801101

Stern, J. A., Walrath, L. C., & Goldstein, R. (1984). The Endogenous Eyeblink. *Psychophysiology*, *21*(1), 22–33. https://doi.org/10.1111/j.1469-8986.1984.tb02312.x

Strauch, C., Barthelmaes, M., Altgassen, E., & Huckauf, A. Pupil Dilation Fulfills the Requirements for Dynamic Difficulty Adjustment in Gaming on the Example of Pong. In: *ACM Symposium on Eye Tracking Research and Applications*. ETRA '20 Adjunct. New York, NY, USA: Association for Computing Machinery, 2020, June, 1–9. ISBN: 978-1-4503-7135-3. https://doi.org/10.1145/3379157.3388934.

Sweetser, P., Johnson, D., Ozdowska, A., & Wyeth, P. GameFlow heuristics for designing and evaluating real-time strategy games. en. In: ACM Press, 2012, 1–10. ISBN: 978-1-4503-1410-7. https://doi.org/10.1145/2336727.2336728.

Sweetser, P., Johnson, D., & Wyeth, P. Revisiting the GameFlow Model with Detailed Heuristics, 8.

Sweetser, P., & Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. *Computers in Entertainment*, *3*(3), 3. https://doi.org/10.1145/1077246.1077253

Sweller, J. (2011). Cognitive Load Theory - ScienceDirect. Retrieved June 26, 2018, from https://www.sciencedirect.com/science/article/pii/B9780123876911000028

Thayer, J. F., & Lane, R. D. (2009). Claude Bernard and the heart–brain connection: Further elaboration of a model of neurovisceral integration. *Neuroscience & Biobehavioral Reviews*, *33*(2), 81–88. https://doi.org/10.1016/j.neubiorev.2008.08.004

Thomas, M. L. (2011). The Value of Item Response Theory in Clinical Assessment: A Review. *Assessment*, *18*(3), 291–307. https://doi.org/10.1177/1073191110374797

Thompson, M., Nordin, A. I., & Cairns, P. (2012). Effect of Touch-Screen Size on Game Immersion, 6.

Tsagris, M., & Pandis, N. (2021). Normality test: Is it really necessary? *American Journal of Orthodontics and Dentofacial Orthopedics*, *159*(4), 548–549. https://doi.org/10.1016/j.ajodo.2021.01.003

Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software*, *3*(31), 1026. https://doi.org/10.21105/joss.01026

van Dooren, M., de Vries, J. G.-J., & Janssen, J. H. (2012). Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & Behavior*, *106*(2), 298–304. https://doi.org/10.1016/j.physbeh.2012.01.020

Vorapongsathorn, T., Taejaroenkul, S., & Viwatwongkasem, C. (2004). A comparison of type I error and power of Bartlett's test, Levene's test and Cochran's test under violation of assumptions. *26*(4).

Westland, J. C. (2011). Electrodermal Response in Gaming. *Journal of Computer Networks and Communications*, *2011*, 1–14. https://doi.org/10.1155/2011/610645

Witmer, B. G., & Singer, M. J. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, *7*(3), 225–240. https://doi.org/10.1162/105474698565686

Yannakakis, G. N., Martinez, H. P., & Garbarino, M. Psychophysiology in Games (K. Karpouzis & G. N. Yannakakis, Eds.) [Series Title: Socio-Affective Computing]. en. In: In *Emotion in Games* (K. Karpouzis & G. N. Yannakakis, Eds.). Ed. by Karpouzis, K., & Yannakakis, G. N. Vol. 4. Series Title: Socio-Affective Computing. Cham: Springer International Publishing, 2016, pp. 119–137. ISBN: 978-3-319-41314-3 978-3-319-41316-7. https://doi.org/10.1007/978-3-319-41316-7_7.

Zimmerman, D. W. (2000). Statistical significance levels of nonparametric tests biased by heterogeneous variances of treatment groups. *The Journal of General Psychology*, *127*(4), 354–364. https://doi.org/10.1080/00221300009598589

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *57*(1), 173–181. https://doi.org/10.1348/000711004849222