Experience-driven MAR games: Personalising Mobile Augmented Reality games using Player Models.

Vivek R. Warriar

PhD thesis

School of Electronic Engineering and Computer Science
Queen Mary University of London

2021

Abstract

We are witnessing an unprecedented growth of Mobile Augmented Reality (MAR) technologies, one of the main research areas being MAR games. While this field is still in its early days, researchers have shown the physical health benefits of playing these type of games. Computational models have been used in traditional (non-AR) digital games to predict player experience (PX). These models give designers insights about PX, and can also be used within games for real-time adaptation or personalised content generation. Following these findings, this thesis investigates the potential of creating models that use movement data and game metrics to predict PX.

An initial pilot study is conducted to evaluate the use of movement data and game metrics to predict players' emotional preferences between different game levels of an exploration-based MAR game. Results indicate that emotional preferences regarding frustration ($\approx$ 93%) and challenge ($\approx$ 93%) can be predicted to a reliable and reasonable degree of accuracy. To determine if these techniques can be applied to serious games for health, an AR exergame is developed for experiments two, three and four of this thesis. The second and third experiments aim to predict key experiential constructs, player competence and immersion, that are important to PX. These experiments further validate the use of movement data and game metrics to model different aspects of PX in MAR games. Results suggest that players' competence ($\approx$ 73%) and sense of mastery ($\approx$ 81%) can be predicted to a reasonable degree of accuracy. For the final experiment, this mastery model is used to create a dynamic difficulty adaptation (DDA) system. The adaptive exergame is then evaluated against a non-adaptive variant of the same game. Results indicate that the adaptive game makes players feel a higher sense of confidence during gameplay and that the adaptation mechanics are more effective for players who do not engage in regular physical activity.

Across the four studies presented, this thesis is the first known research activity that investigates using movement data and game metrics to model PX for DDA in MAR games and makes the following novel contributions: i) movement data and game metrics can be used to predict player's sense of mastery or competence reliably compared to other aspects of PX tested, ii) mastery-based game adaptation makes players feel greater confidence during game-play, and iii) mastery-based game adaptation is more effective for players who do not engage in physical activity. This work also presents a new methodology for PX prediction in MAR games and a novel adaptation engine driven by player mastery. In

summary, this thesis proposes that PX modelling can be successfully applied to MAR games, especially for DDA which results in a highly personalised PX and shows potential as a tool for increasing physical activity.

Acknowledgements

I would like to thank my primary supervisor, Dr Laurissa Tokarchuk, for her guidance and support across my PhD. It has been a pleasure to work with you, and I wish you the best in your future. I would also like to thank Dr John R. Woodward, my secondary supervisor. Thanks for all your valuable input in this research, I wish you the best in your future. I would like to extend my gratitude to Dr Matthew Purver, my independent assessor, for his wise advice and feedback on this thesis. A huge thanks to Prof Paul Cairns at the University of York for his invaluable advice in this research.

Special thanks to my mother Usha, my father Ramesh and my grandmother Ponamma for all your patience and encouragement during these years. All your loving words have kept me strong during challenging times. I am thankful to the others members of my family for their encouragement and kinds words. I am deeply grateful to my partner, Andrea for being a constant source of patience, support and care over these years.

Thanks to the Media & Arts Technology program for giving me the opportunity to do this PhD in such an inspirational and supportive environment full of talented creative people. I would like to thank in particular to Tom Gruion, Jack Ratcliffe, Nick Ballou, Gideon Raeburn, Oliver Withington, Carmen Ugarte, James Weaver, Rémi de Fleurian and Jonathan Pigrem. It has been great meeting and working with you. You are great people and I wish you the best in your careers. Thanks also to the MAT and Queen Mary staff for their help and work that made my journey a bit easier: Dr Nick Bryan-Kinns, Jonathan Winfield, Geetha Bommireddy, Edward Hoskins, Hayley Cork and Ho Kok Huen. Enormous thanks as well to Tim Kay, Keith Bancroft and Tom King for their IT support and help in critical moments. Finally, I also would like to thank my good friends here in England and India for their kindness, friendship and support during my time in London.

This work was funded by the Engineering and Physical Sciences Research Council (EPSRC) and Arts and Humanities Research Council (AHRC) as part of the Doctoral Training Centre in Media and Arts Technology at Queen Mary

University of London (ref: EP/L01632X/1).

Statement of Originality

I, Vivek Warriar, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below. I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material. I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis. I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university. The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature:

Date:

Contents

List of Figures

List of Tables

List of abbreviations

2-AFC	2-Alternative Forced Choice (questionnaire)
4-AFC	4-Alternative Forced Choice (questionnaire)
AR	Augmented Reality
ASD	Autism Spectrum Disorder
BDI	Belief-Desire-Intention (model)
CV	Cross-Validation
DDA	Dynamic Difficulty Adaptation
ECG	Electrocardiography
EEG	Electroencephalograph
GEQ	Game Experience Questionnaire
GPS	Global Positioning System
GSR	Galvanic Skin Response
HMDs	Head Mounted Displays
IMU	Inertial Measurement Unit
LDA	Linear Discriminant Analysis
MAR	Mobile Augmented Reality
MCTS	Mote-Carlo Tree Search
OST	Optical See Through (display)
PBF	Player Behaviour Feature
PCG	Procedural Content Generation
PENS	Player Experience of Need Satisfaction (survey)
PX	Player Experience
PXI	Player Experience Inventory
SDK	Software Development Kit
SFFS	Sequential Floating Forward Selection
SFS	Sequential Forward Selection
SVM	Support Vector Machines
VR	Virtual Reality
VST	Video See Through (display)

Chapter 1

Introduction

With the rapid growth in the games industry and the increasing number of people engaging with digital games, it has become increasingly important to investigate methods to model player experiences in these environments. This approach is referred to as player experience modelling[240]. Player experience (PX) modelling refers to the use of computational models to predict PX based on features computed from the game-play behaviour of players. PX modelling is a popular field of research in traditional digital games. There are two main reasons for this; first, since games are played by many players spread across the world who tend to show a diverse range of behavioural patterns (which is linked to different motivations for their game-play). While it is impossible for game designers to manually conduct user studies with all the various types of players of their game, it is possible to log player behavioural data from their games and use data analysis techniques to gain insights about player's experiences and motivations. Second, robust predictive models of player experience can be used to optimize games to best suit player's ideal experience either through procedural content generation (PCG)[239] or balancing game parameters[122].

Recent advancements in Virtual and Augmented Reality (VR/AR) technology has enabled the games industry to engage players using mixed reality platforms. VR aims to immerse a user into a digital world while AR overlays digital information into the user's real world. Both these technologies have seen a growth in adoption by consumers due to their increasing availability. Game companies have released VR head-mounted displays (HMDs) since the early 1990s; however, the release of the oculus rift in 2010 has set off a new wave in immersive VR HMDs, with companies such as HTC, Google and Samsung releasing their consumer VR HMDs[54].

AR has followed a similar growth over the last decade and is consumed by using HMDs (such as the Microsoft Hololens device) and mobile devices. While

AR HMDs are still not used actively by consumers, mobile AR has increased in popularity and mainstream usage. The release of *PokémonGO*[3] by Niantic labs in 2016 has shown the tremendous potential of AR mobile games, with the total number of downloads crossing 1 billion as of February 2019[228]. The release of mobile AR software development kits (SDKs) such ARcore from Google[1] and ARkit from Apple[87] has made developing for this platform more accessible for developers. The potential of mobile AR has been investigated by research for a number of contexts such as gaming[225, 212, 210], tourism[111, 217], education[231, 229], and physical fitness[109, 9] to name a few.

AR games have been developed for research purposes in both *table-top*[153] and *movement*[207] types of game-play, and are played using either head-mounted devices (HMD) or mobile devices. This research focuses on mobile AR games played within local spaces referred to in this thesis as *local AR games*. These type of games are different to location-based AR games such as *PokémonGO*[3] or the recently released *Harry Potter Wizards Unite*[2] that involve players travelling across several locations to accomplish game objectives. *Local AR games* are a genre of newly emergent games that involve players using the mobile device as a *magic window* into the AR game world and generally involve the player's movement within their local space without the need for travel between locations. Previous research has shown that PCG techniques can be applied to automatically create game levels for AR games[12] however, existing systems have not considered the player's experiences in the generation of these game levels. Existing research has focused on novel applications for mobile AR games such as for entertainment[137], education[115], and multi-player engagement[23]. It is observed that research work into understanding and modelling player experiences in these environments is limited.

This PhD thesis investigates techniques for modelling player experience in mobile AR games. The four experiments described in this thesis shows that PX can be computationally modelled through supervised learning techniques that use mobile sensor data and game metrics. Additionally, these models can be used to adapt the difficulty of these games for an optimal PX. This chapter is structured as follows: section 1.1 describes the motivations of this thesis in the context of existing research work. Section 1.2 describes the aims and approach of this research. Section 1.3 outlines the contributions made by this thesis. Section 1.4 lists the publications that were made across the PhD. Section 1.5 gives details about the impact the COVID-19 pandemic had on this research work. Finally, section 1.6 describes the structure of this thesis.

1.1 Motivations

The development of mobile computing and camera technology for smartphones has enabled rapid advancement in mobile AR games. There exist some examples of state-of-the-art applications of AR mobile games; however, there are a limited number of studies focusing on modelling player experience in this domain. This research aims to bridge this gap in knowledge on approaches to construct models that predict player experience in these environments. This research will complement existing research activities in AR games since understanding player experiences is an important factor for both research and industry.

As described in the section above, body movement is a common factor in these games since players have to move their mobile devices around to view and interact with the AR world. This research explores movement data as an information medium to model player's experiences in these games. Existing research has shown the viability of using people's movement data to predict their emotional states[193, 194] and it also shows that this data is an important medium for game designers to gain insight into player engagement in movement-based games[25, 26]. These activities have been explored in games that use movement-based controllers such as the Nintendo Wii [194, 193] and it is still an open research question to the extent to which similar approaches are aplicable to *local AR games* that involve a *magic window* based interaction. Furthermore, since player movement is an important feature of these games, modelling player experiences based on movement can be a generic tool for predicting player experiences in these game systems.

This approach to building player experience models can be beneficial to the games industry to effectively scale up within mixed reality games since these models can be used to understand player experiences based on player behaviour data without expensive user testing. This potential advantage has motivated the first three experiments of this research. Additionally, these models can optimize the game to best suit a diverse range of players, allowing for personalising these games based on their experience. Which has motivated the final user study of this research that aims to dynamically adjust the difficulty of a *local AR game* based on these models.

Finally, researchers have observed the potential of these games in creating positive health outcomes for their players. This benefit is because movement-based AR games increase a player's physical activity through game-play; these games are also called AR exergames. This potential health benefit has been investigated in *Pokemon Go* players[109, 9] as well as through the creation of AR exergames for research [119, 112]. While these research AR exergames have some game design patterns or parameters in common (e.g., area of the game level and the number of game rewards within the area), it is unclear how different settings of these design patterns or game parameters will impact PX (e.g., What is the impact of large vs smaller game areas on PX?). Thus, the final motivation of this thesis is to evaluate the impact of these commonly used AR exergame parameters on PX. Which would further contribute to the design of effective mobile AR exergames. This motivation is addressed in this thesis's second and third experiments, which uses a mobile AR exergame as a test environment for PX modelling and exergame parameter evaluation.

According to the research motivations presented in this section, this thesis will address the following research questions:

RQ1: *To what extent can player movement and game metric data be used to predict PX in mobile AR games?*

RQ2: *What is the impact of commonly used AR exergame parameters on PX?*

RQ3: *Can these predictive models of PX be used for dynamic difficulty adaptation in mobile games to improve PX?*

These research questions are addressed across the four experiments reported in this thesis. First, **RQ1** is addressed in the first three experiments (chapters 3, 4, 5), which collected PX and corresponding player movement and game metric data to build and evaluate supervised learning models that predict several dimensions of PX in MAR games. Next, **RQ2** is investigated across two studies (chapters 4, 5), using both quantitative and qualitative methods. Finally, **RQ3** is addressed in the final study reported in this thesis (chapter 5), where a player mastery model (built in the previous experiment) is used to develop an adaptive exergame. This adaptive game is evaluated against a non-adaptive version using both qualitative and quantitative approaches.

1.2 Aims and Approach

Since players movement is a common feature of *local AR games*, this research investigates techniques to model players experience in AR mobile games based on player movement data. It is worth noting that this research scope will focus on *local AR games* and not location-based AR games. Since player movement data or source code from existing games is not available, this research involves the development of two *local AR games*. These games have been parameterized, so that game levels for both are generated using a set of designer defined parameters. Changing these parameters will result in a diverse range of emotional responses from players.

The first game is an exploration-based *AR Treasure Hunt* game that is used in the first user study conducted as part of this research (details included in chapter 3 of this thesis). Thus, this first study is an initial proof of the viability of using movement data to model player experience in these games.

The second game is a target acquisition game titled *Running Chickens* where players are rewarded for their physical exertion within local spaces to accomplish game objectives, referred to as exergames. While other research works in AR exergames have used case study approaches to investigate the success of these games, there is limited knowledge on how commonly used game mechanics can impact player experience. Therefore, this game is used for the rest of the experiments conducted in this thesis.

In the second and third experiments, *Running Chickens* is used to investigate to what extent the modelling techniques used in the first user study can be successfully applied to other *local AR games* that are different in both objectives and degree of physicality required for game-play interactions. This is important in further establishing the generalisability of the method used to model and predict player experience from the first study. Additionally, these user studies are used to empirically evaluate the impact of AR exergame parameters on player experience.

Finally, to further investigate the usefulness of these PX models, a final user study is conducted in this research work. In this study, players are used to empirically evaluate a version of *Running Chickens* where the PX models built with data from the third study is used to adjust the difficulty of the game to deliver an optimal experience for the players. This is referred to as *Dynamic Difficulty Adjustment* (DDA) in research literature[249, 122]. The subsection below presents the scope of this research work.

1.2.1 Research Scope

This research will use mobile devices and existing AR frameworks to develop the two AR games. These games use the ARcore SDK[1] in the unity 3d engine[211]. Android mobile devices are used for development and in the user studies reported in this research work. This work is limited to the context of mobile AR games and will not explore HMD based AR experiences. Another important distinction about the games explored in this research is between location-based AR games and local AR games. Location-based AR games involve GPS and have players travelling across several locations to accomplish game objectives, while local AR games are played within the same location and involve the players' movement within their local spaces. Again, this research is limited to local AR games and will not use location-based AR. Finally, this research scope has been limited to single-player games and will not address multi-player AR games.

1.3 Contributions

Following the aims and approach described in the previous section, this work contributes to mobile AR games' advancement. AR mobile games can leverage a number of streams of data such as sensor data and game metrics to model player experience within these games. This work investigates movement data extracted from mobile IMU sensors and game data for real-time player experience prediction and personalisation. This personalisation is essential for exergames that promote physical fitness. This PhD thesis makes the following novel contributions to the existing literature:

1. *Empirical evaluation of the relationship between commonly used AR exergame design patterns on player experience.* This thesis evaluates two commonly used AR game patterns: area of the game level and the number of rewards (e.g. collectables).

    - *Game Area:* The research conducted found that game area positively impacted experiences of tension and challenge. In general, when playing local AR games, players find navigating over large areas challenging, which can lead to a negative experience if they are already overwhelmed or do not generally engage in physical activities.
    - *Number of rewards:* this research work finds that this game mechanic positively impacts experiences of Arousal, Flow, Tension, Challenge, Positive Affect and Immersion. In general, participants found a high number of rewards motivating as it enabled them to get immersed in the game world.

    Finally, both game patterns impact the amount players must navigate around a physical area during gameplay. This research found that as players move around a space, their attention switches between the AR world (through the mobile viewport) and the non-AR world (using their peripheral vision). This attention switching occurs to enable safe movement through the physical space during gameplay. However, this attention switching can break a player's immersion in the game.

2. *A novel methodology for Player Experience Prediction in mobile AR games is presented.* The player experience prediction system measures the player's movement, game parameters (e.g., the settings of different game mechanics) and player performance to infer their player experience in the game.

Although body movement signals have been previously used for emotion recognition, this thesis presents a new methodology that uses this movement data and game data to predict a player's experience in a game.

3. *A novel adaptation engine that uses a player's predicted mastery in a game level to adjust the difficulty of the game experience.* Using the player experience prediction system described above to predict a player's perceived mastery from a game level, this adaptation engine adjusts the difficulty of the game in real-time to provide a personalised game experience in AR exertion games. Furthermore, by using mastery-based adaptation, that game experience increases in difficulty, ensuring that the physical exertion needed from the player is maximised without overwhelming them. This is the first mobile AR exertion game using player mastery prediction for dynamic difficulty adaptation (DDA) to the researcher's knowledge.

4. *The use of mastery-based game adaptation for mobile AR exergames is more beneficial for players who do not usually engage in physical activity.* This finding is likely because in-game rewards more influence this group of people. Thus having the game reduce difficulty (when people feel a low level of mastery in-game) allows them to gain more rewards. However, this is not the case for physically active people as they are motivated by self-improvement. Thus they do not respond as positively to reductions to in-game difficulty.

1.4 Associated Publications

This section lists a number of publications that I have worked on during the PhD program. The following publication is related to the research presented in this thesis:

- Vivek R.Warriar, John R. Woodward, and Laurissa Tokarchuk. *"Modelling player preferences in AR mobile games."* 2019 IEEE Conference on Games (CoG). IEEE, 2019.

  This conference paper contained an adapted version of the experiment presented in chapter 4 of this thesis.

The following is a list of conference publications that are not directly related to the thesis:

- Vivek R.Warriar, Carmen Ugarte and Laurissa Tokarchuk. *"Learning to generate personalised content based on human behaviour."* Human-Like Computing Machine Intelligence Workshop (MI21-HLC), 2019.

This conference paper describes techniques of using player experience models to create personalised game content using quality-diversity algorithms.

- Vivek R.Warriar, Carmen Ugarte, John R. Woodward and Laurissa Tokarchuk. *"Playmapper: Illuminating design spaces of platform games."* 2019 IEEE Conference on Games (CoG). IEEE, 2019.

  This conference paper describes research work that evaluates player behaviour as input to generate Mario game levels using quality-diversity algorithms.

- Nick Ballou, Vivek R. Warriar, and Sebastian Deterding. *"Are You Open? A Content Analysis of Transparency and Openness Guidelines in HCI Journals."* Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021.

  This conference paper describes research work that evaluates the extent to which the Transparency and Openness (TOP) guidelines from the Center of Open Science [1] is followed by HCI journals.

1.5 COVID-19 Impact Statement

The Covid-19 pandemic significantly impacted the research work since it involved human data collection. Due to social distancing measures imposed by the UK government, data collection for the last 2 (out of 4) experiments was affected. The pandemic broke out in the UK while data collection for the third experiment was in progress. Due to health and safety concerns, data collection for experiment 3 was terminated early. This study aimed to collect data from a sample of 40 participants; however, this study was ended after collecting data from 25 participants. The fourth experiment was originally planned as an in-person study; however, the experimental method had to be modified for remote data collection. This resulted in significant delays to the research work due to the changes in research design and obtaining ethics approval for the study.

1.6 Thesis Outline

This thesis is structured in seven chapters, as follows:

Chapter 2 outlines previous research done in the areas of interest of this thesis - Player Modelling in Games; Augmented Reality games; Body movement and engagement in games; Dynamic Difficulty Adaptation in Games. This

[1] Details about the TOP guidelines can be found here: `https://www.cos.io/initiatives/top-guidelines`

chapter also presents an analysis of the gaps identified in the existing literature and presents the approach taken in this thesis.

Chapter 3 presents an initial study that investigates modelling players' emotion preferences (e.g., Did the player find level A of the game more fun than level B?) in an exploration-based mobile AR game. This study also presents the custom game *AR Treasure Hunt*, which was developed for this study.

Chapter 4 reports a follow-up study that further evaluated the modelling pipeline developed in the previous study with two key differences. The first one is that a different game that is more exertion driven is used in the experiment. Secondly, the study aims to model experiential constructs related to player experience research (e.g., immersion or player competence) instead of emotion preferences. The study also included an empirical evaluation of the experiential impact of commonly used AR game patterns. Finally, this study also presented the custom exertion game *Running Chickens*, which was developed for this study.

Chapter 5 reports the last two studies of this thesis, where the player experience prediction pipeline was applied to predicting experiential dimensions related to player motivations and then used to create a real-time dynamic difficulty adaptation engine for the *Running Chickens* game. The first study describes the evaluation of the prediction pipeline and the empirical evaluation of AR game patterns on these for these experiential constructs related to player motivation. The chapter then describes the development of a dynamic difficulty adaptation (DDA) engine based on predicted mastery (which was one of the player motivation related constructs explored). Finally, the second study reported in this chapter presents a user-centric evaluation of this DDA version of the *Running Chickens* game as compared to the non-adaptive version.

Chapter 6 concludes this thesis by summarizing the main findings, contributions and limitations of this research. Finally, future research directions are also proposed.

Chapter 2

Background and Approach

This research aims to apply techniques to model player experiences in mobile AR games based on player movement data and investigates if these models can be used to dynamically adjust the difficulty of a *local AR game*. This chapter will describe relevant research activities from these domains. The literature reviewed in this chapter has been divided into four themes: Augmented Reality games; Player Modelling; Body movement and player engagement in games; and Dynamic Difficulty Adjustment in games.

The first section (section 2.1) will describe current research activities in Augmented Reality games. Section 2.2 will cover research activities in player modelling. Section 2.3 presents research on the role of body movement and player engagement in games. Section 2.4 will describe research on Dynamic Difficulty Adjustment (DDA). Section 2.5. Section 2.6 describes the approach taken in this research. Finally section 2.7 will provide a chapter conclusion.

2.1 Augmented Reality Games

AR is considered complimentary to VR, which immerses a person in a digital world. While AR overlays digital content into the user's real world to enhance their engagement with the world, VR aims to separate a person from the physical world to increase their sense of immersion within a digital world. Milgram's virtuality continuum [143] shows how VR and AR interfaces are related to each other. This continuum is shown in figure 2.1, on the left of the continuum is the real world with no digital information while on the right are VR interfaces (with no elements of the real world). Anything in between these extremes is considered as mixed reality which uses elements from both the real and digital worlds. AR lies closer to the real world on this continuum since these interfaces incorporate digital information into the users' view of the real world. On the other hand,

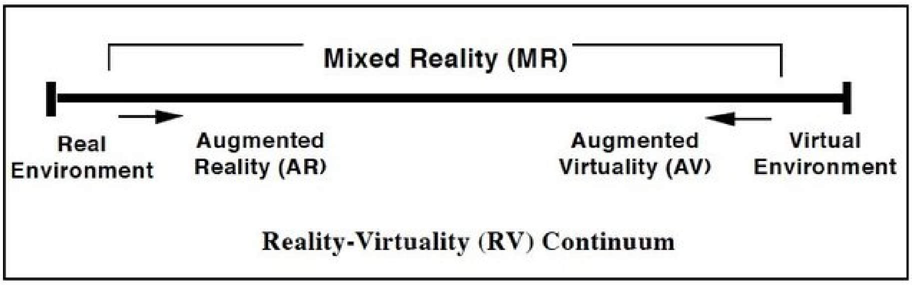


Figure 2.1: Diagram showing Milgram's Virtuality Continum[143]

closer to the VR extreme are augmented interfaces that use information from the real world in a virtual environment. Azuma defines AR systems using the following characteristics [13]: it combines the real and virtual, it is interactive and in real-time, and finally it is registered in 3D.

There are three broad classifications of devices that support AR experiences [212, 225], HMDs, handheld devices and projectors. HMDs can be further divided into two categories: optical see-through (OST) devices that overlay digital content onto the view of the real world and video see-through (VST) devices that render digital content onto a video of the real world which is then presented to the user. Handheld AR refers to using devices such as mobiles and tablets with cameras that overlay digital content in the camera view of the device, this is also referred to as a *mobile augmented reality (MAR)*. Finally, projector-based AR uses a single fixed projector to show digital content onto real objects. This research focuses on MAR game experiences.

AR has been a focus for research in entertainment computing having several applications in AR games [153, 212]. Nilsen et al.[153] describe how AR gaming can leverage game mechanics from both real-world and digital games. The authors show how AR games enhance a player's experience in 4 factors: physical, mental, social and emotional. Table 2.1 shows game mechanics from the real world and digital games. AR games can use mechanics across these factors leveraging advantages of both the real and digital world to best suit the game context and overcome any limitations of the specific mediums.

Research applications in AR games have focused on creating AR versions of more traditional physical and virtual games. There exist several AR counterparts to more traditional physical sports. The game of *Soccer* has been adapted to a number of augmented reality versions[159, 184]. The *AR-soccer*[184] game allowed players the novel interaction of being able to kick and interact with a virtual ball using their feet as they would a traditional physical football. In the user study to evaluate the game, Reimann[184] found that users find physical interactions with virtual content (running and kicking) caused a lot of fun and

	Physical	Mental	Social	Emotional
Real World	Can use player's whole body Real world can provide game environment. Physical artefacts can have game significance.	Players unwilling to resolve complex rules. Supports spatial reasoning, particularly 3D.	Supports natural face to face communication.	Can stimulate players across full range of senses. Limited by practical ability to control environment.
Digital	Physical interaction limited by input devices.	Supports complex game models and rules. Can provide AI opponents and agents.	Mediation limits communication, but can provide other facilities. Allows remote and massively multiplayer games.	Potential for diverse virtual environments and scenarios. Limited to audio and visual stimulation.

Table 2.1: Table showing game mechanics from real world and digital games [212]

engaging experiences in the game. Similar examples exist in *tennis*[79] and *basketball* [191]. *AR-tennis* presented by Henrysson et al. [79] is a game where two players compete in a match of virtual tennis against each other using individual mobile devices as windows into the world and as a paddle to hit the ball back to the opponent.

A number of research examples have also been developed of games that are variants of traditional digital games, it has been observed that these AR games use existing game mechanics and conventions that have been adapted for an AR paradigm. *X-portal*[207] is a first-person shooter in AR that enhances "full-body engagement and supports new immersive experiences". The game was adapted from an existing war simulation game by mapping the game experience into AR and is competitive in nature. An example of social gaming has been explored in a game titled *Cows vs. Aliens* by Mulloni et al.[147]. The game is a competitive multiplayer team game that exploits mobility and social interaction as core gameplay elements. Bricks [23] is another multi-player mobile AR experience that supports collaborative game-play within a local environment.

A number of AR games exist which incorporate GPS sensor data into the game mechanics. These techniques apply to game experiences that are meant to be played over large areas. *Real Tournament* is an AR first-person shooter experience developed by McCaffery et al.[137]. Reimann and Paelke [185] build an AR outdoor mobile version of their game *Lost Valley* which combines genres from fighting and puzzle-based games. These outdoor mobile games incorporate

GPS data and can be considered as an extension of the area of location-based games [7].

In the domain of commercial games, the recent success of the game *PokémonGO* [3] is worth noting. The app has been downloaded more than a billion as of February 2019[228]. The game uses GPS and IMU sensors from mobile devices to drive game interactions which are based on the popular franchise *Pokémon*. The success of this game has motivated research activities to explore a number of aspects of the game from the potential health benefits of playing these games [109, 9], players' experiences of flow as a consequence of game-play [125], to its positive impact on mental health [138, 102] and motivations for engaging with the game [158]. Another commercial game that has been of interest to this research work is *Zomies! Run* [8] which is an audio-based augmented reality game that research has shown to improve outdoor running experiences for players [203, 230].

The research work conducted as part of this thesis focuses on Augmented Reality Games for physical health benefits which are known as AR exergames. There is a limited amount of research work conducted into AR exergames. *Gioboids* is an AR exergame developed by Lindeman and Lee [119] which is a target acquisition game where players run around the real world capturing digital creatures. Another similar example is *Calory Battle AR* which is a research game presented by Laine and Suk [112], the game involves players travelling around the world diffusing AR bombs. Both these games have been validated using user testing to show that they can facilitate positive game experiences, it is unclear how the game mechanics used in these games can influence player experience. *STAR* is an HMD based AR first-person shooter game developed to increase the physical activity of their players [103]. Research in AR exergames has also investigated the potential of improving the physical fitness of older adults by gamifying the experience of performing exercise at home [226, 154] and to minimise their fall risk[37]. These types of games have also been investigated for applications of stroke rehabilitation [52, 73]. While these research games show the potential of using AR games to improve physical health, it is possible that personalizing these games based on individual needs will further improve the objectives of these serious games. This objective is further investigated in this research work.

There exists a large number of AR frameworks available for the fast prototyping and testing of game concepts. Some of them are ARKit[87], ARCore[1], Vuforia[4] and Wikitude[5]. The frameworks can be integrated with the game engine Unity3D[211] to create mobile games for android and iOS mobile devices. These frameworks use the camera and IMU sensors from mobile devices to track the position and orientation of the device and to detect visual anchors in the environment. Some of these libraries offer environment sensing features such as point and plane detection so that virtual content can be augmented onto the physical environment.

There exist many AR games developed both for research and commercial purposes however, there is an observed gap in the literature on studies that model players experience in these environments. This research will use the ARcore SDK[1] to create two *local AR games* that will serve as test-beds for this research. The aim is to model player's experiences in these environments from their body movement. The next section will present a background on player modelling.

2.2 Player Modelling

The area of research in games referred to as player modelling[243, 202] refers to "the detection, prediction and expression of human player characteristics that are manifested through cognitive, affective and behavioural patterns while playing games." [240]. Yannikakis and Togelius argue that the main aim of player modelling is to understand players' cognitive, affective and behavioural patterns [240].

Models are mathematical or computational representations that capture an underlying function between player behaviour and their emotional response to the game. Modelling human beings has been of interest to the field of human-computer interactions, with player modelling being a subset of this research focusing on digital games. These models are built using machine learning methods, such as supervised learning. Training data is collected from players' interactions with games and labels for these techniques are acquired using an assessment of player experiences or player behaviour. This is usually done through self-reported questionnaires or from annotations by expert observers. These techniques are used to find predictors of the game experience. These predictors can be informative to game designers to adjust aspects of the game or by game adaptation algorithms that adjust the game experience in real-time. It is important to note that player models can be built for both PX and player behaviour prediction or detection. Since this research focuses on PX models, the remainder of this section presents background work on PX modelling. Interested readers are referred to [240] for more information on player behaviour modelling.

Yannakakis and Togelius illustrate a high-level taxonomy for player modelling which is seen in figure 2.2. They observe that irrespective of the application domain (PX vs. player behaviour), player models consist of 3 components: the model's input, the computational model, and the model's output. The model itself is a mapping between the input and the output. This mapping is either manually designed or derived from data, or a mix of the two. The authors present a high-level classification of the approaches to player modelling: model-based (or top-down) and model-free (or bottom-up) approaches[243, 239]. The above definitions are inspired by taxonomies in reinforcement learning in which a world model is available (i.e., model-based) or not (i.e., model-free). Given the two ends of this continuum, hybrid approaches between them can also exist. The gradient red colour of the player model box in figure 2.2 illustrates the continuum between top-down and bottom-up approaches. These classifications are explained in the subsections below (subsections 2.2.1 and 2.2.2). Following this, a taxonomy of inputs for player models is described in subsection 2.2.3 and the various ways a player model output can be represented is presented in subsection 2.2.4.

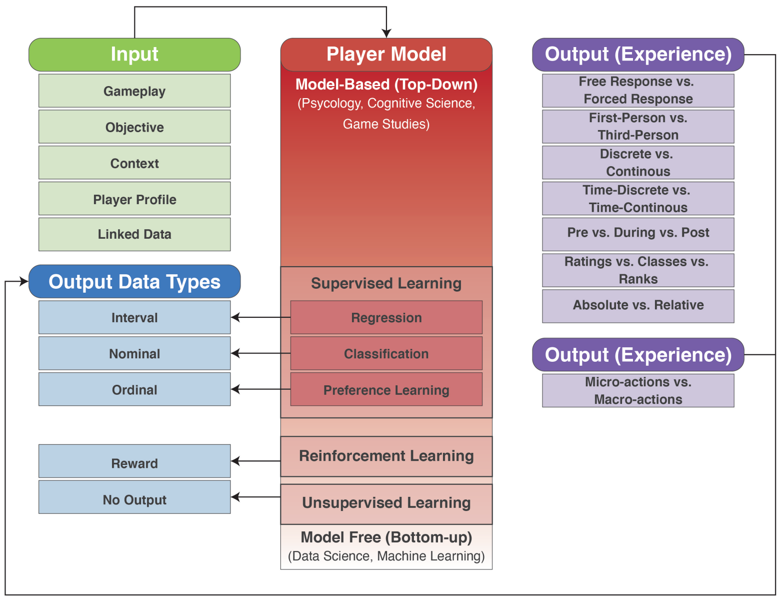


Figure 2.2: The components of player modelling as presented by Yannakakis and Togelius[240]. Model-based and model-free approaches are described in subsections 2.2.1 and 2.2.2 accordingly. The various options for the input of the model are discussed in subsection 2.2.3. The taxonomy for the model's output is discussed in subsection 2.2.4. Finally, the various AI methods (supervised learning, reinforcement learning and unsupervised learning) are used for modelling corresponding output data types.

2.2.1 Model-Based (Top-Down) Approaches

In model-based or top-down[243] approaches a player model is built on a theoretical framework which is proposed by researchers in psychology, humanities and social sciences, to explain phenomena related to human experience or behaviour.

Player models can be developed based on theories of emotion such as the cognitive appraisal theory[63, 197]. Furthermore, player models may rely on representations of human emotions such as the emotional dimensions of arousal and valence[59] that draw from Russell's circumplex model of affect[177] (see figure 2.3). Valence refers to the extent to which an emotion is positive or negative, whereas arousal refers to how intense that emotion is. Following these theoretical models, emotional responses from players are mapped to different player states while playing the game.

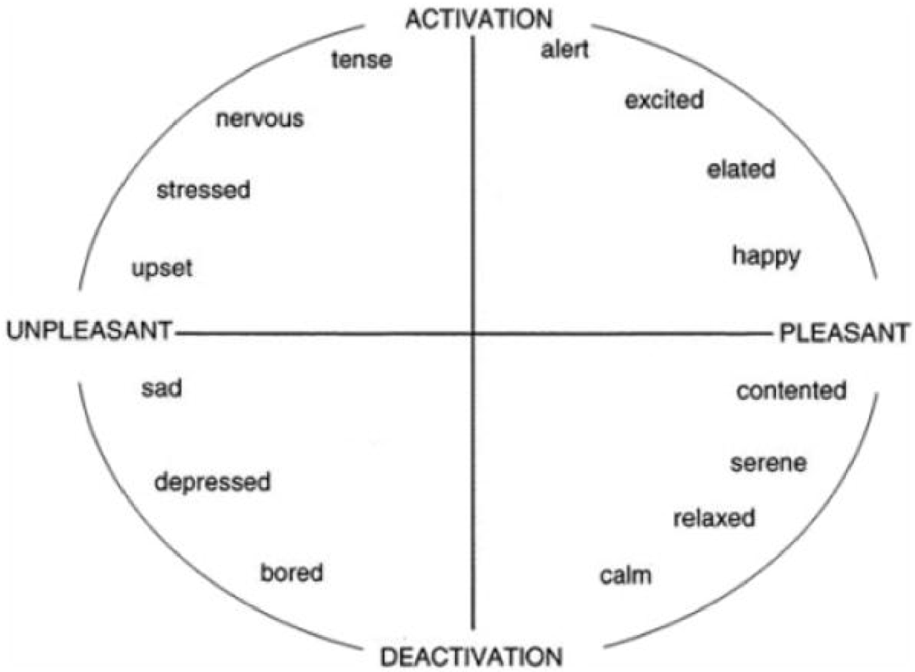


Figure 2.3: A graphical representation of Russell's circumplex model of affect with the horizontal axis representing the valence dimension and the vertical axis representing the arousal or activation dimension.[177]

Yannakakis and Togelius [240] present theories from cognitive-behavioural research which can also serve as the basis for constructing player models. These theoretical frameworks include *the theory of the mind*[179] (modeling aspects of social interactions in games), *usability theory*[150, 89], the *belief-desire-intention (BDI) model*[29, 66], the *cognitive model* by Ortony et al. [157] and Skinner's behavioural theory[201] with is related to reward systems in games.

One of the most relevant concepts from psychology for game studies is the theory of flow by Csikszentmihalyi[44, 45, 46]. This has been a popular construct for modelling player experience. A person in a state of flow during an activity is characterized by high concentration in the present, loss of a sense of self-consciousness, high degree of control over the task, loss of perception of time, and high degree of intrinsic reward from the activity. For a person to enter this flow state it is important to have an appropriate balance between the challenging nature of the task and the skill of the user. Within games, this experience is characterized as a fine balance between boredom (where the task is not challenging for the player) and anxiety (where the task difficulty is too high for the player), also known as the flow channel (see figure 2.4). The flow theory has been adapted to applications of games to understand player experience[208, 209, 149]. Several models have also been developed from studies about how people engage

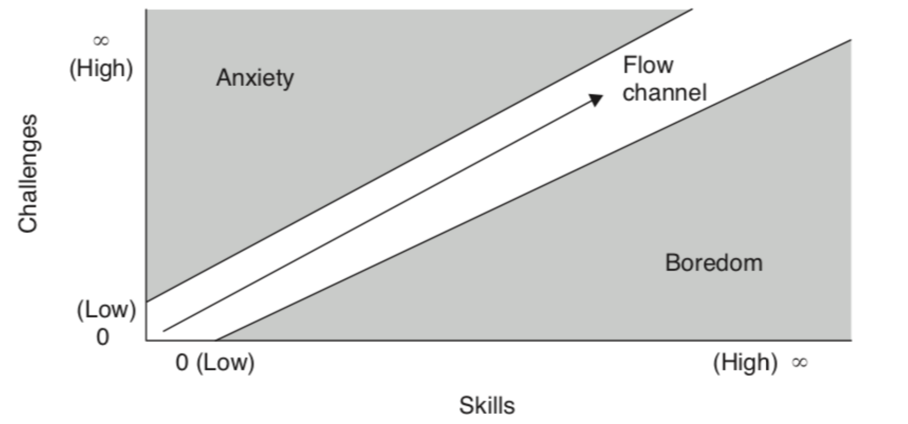


Figure 2.4: Diagram showing the flow channel in games.[195]

with games. A popular model is Malone's dimensions that contribute to fun in games [127] which are challenge, curiosity and fantasy. Challenge refers to the uncertainty of achieving the goals usually due to the difficulty of the task. Curiosity refers to the player's feeling of uncertainty of what will happen next in the game. Finally, fantasy is the ability of the game to show or evoke images of physical objects or social situations not present. Malone's dimensions of fun have been evaluated in player experience research in prey-predator games[234], physical games[235, 241], and racing games[215].

Bartle's[18] classification of player types has been used as a way of modelling players based on their behaviour. This classification identifies four types of players named *killers* (players that focus on winning and are engaged by ranks and leaderboards), *achievers* (players that focus on achieving goals quickly and are engaged by achievements), *socializers* (players that focus on social aspects of games such as developing a network of friends) and *explorers* (players who focus on the exploration of the unknown).

Another popular model for understanding player experience included the theory of fun by Koster[110]. Koster's theory explains the notion of fun with learning in a game: the more you learn the more you tend to play a game. According to this theory, a person stops playing if it is too easy (no learning of new skills) or too hard (learning does not happen either).

Lazzaro's model[114] identifies four factors of fun in games: hard fun (playing to win and challenge one's self), easy fun (playing to explore a new game world or new game experience), serious fun (playing to get better at a skill that matters to the player) and people fun (playing as a need of social engagement with other players).

It is observed that the literature of theoretical approaches to player experience is rich and these models can be used to describe a large number of phenomena related to player experience in games. However, Yannakakis and Togelius express caution when using these theoretical models of player experience in the construction of player models, since these have not been derived from or tested on interactive media such as video games[240].

2.2.2 Model-Free (Bottom-Up) Approaches

Model-free approaches refer to the data-driven construction of an unknown mapping (model) between player input and a player state. Model-free approaches involve the collection of observations which are then analyzed. For this, Player data and labels of player states are collected and used to derive the model (usually using machine learning approaches).

This is traditionally accomplished using techniques in supervised learning such as classification, regression or preference learning [240]. Additionally, the authors note that reinforcement learning can be applied when a reward function, instead, can characterise aspects of player behaviour or experience. Unsupervised learning is applicable when target outputs are not available for predictive purposes but, alternatively, data is used for the analysis of playing behaviour.

2.2.3 Input of a Player Model

Any manifestation of player affect or behavioural pattern could define the input of the model. It is important to note that both Top-down and Bottom-up approaches to player modelling require some form of input(s). The model's input can be of three main types: (1) anything that a player is doing in a game environment gathered from **game-play data** i.e., player input data of any type such as user interface selections, preferences, or game-play actions; (2) **objective data** collected as responses to game stimuli such as physiology, speech and body movements; and (3) the **game context** which comprises of any type of game content viewed, played, and/or created.

Game-Play Data When game-play data is used, the assumption is that a player's actions and preferences are linked directly to their experience. If this is the case, it is possible to infer the player's current experience by analyzing patterns of their game interactions[42, 72]. Any form of data that has been logged from the direct interaction between the player and the game is known as game-play input. These measures of game-play have also been defined as *player metrics*[57].

Objective Data Players can experience a wide spectrum of emotional responses during their game-play. These emotional responses from players in-turn lead to changes in their physiological responses. There exists studies that explore the interplay between physiology and game-play by investigating the impact of different game-play stimuli on a number of physiological signals. Such signals are usually obtained through electrocardiography (ECG) [242], photoplethysmography[242, 216], galvanic skin response (GSR)[129, 82, 81, 83], respiration[216], and electroencephalography (EEG)[152] (among others). In addition to physiology the player's bodily expressions (motion tracking) at different levels of detail can be used to infer affective responses from the game-play stimuli. These include facial expressions[99, 10, 71, 31, 246], muscle activation[41, 51], body movement and posture[99, 11, 219, 53], speech[224, 98, 96, 95, 16], text[161], haptics[156], gestures[84], and eye movement[11, 148]. Although objective measures can be useful in inferring a player's emotional state, in the case of using physiological sensors these are considered as intrusive on the player and can have an impact on their experience[240].

Game Context Data Game context refers to the state of the game during play and excludes any interactions of the player with the game (which has been discussed above). Although game-play data and game context data are closely related to each other. Game context data is considered as a type of game metrics as opposed to player metrics. A few studies have investigated the physiological reactions of players in isolation (without game content). Yannakakis and Togelius argue that player modelling would require information about the current game state. For instance, the model needs to know if the GSR increases because the player died or completed the level. The game context is combined with other input data from the player has been used extensively in the literature for the prediction of different affective and cognitive states relevant to playing experience[41, 134, 167, 199, 198, 187, 183, 133].

2.2.4 Output of a Player Model

The output of the player model is referred to as the player state which is any representation of the player's experience or current emotional, cognitive, or behavioural state. It is important to note that both Top-down and Bottom-up approaches to player modelling will produce some form of output(s). This research work focuses on a player's experience or emotional state. To model the experience of the player, labels or annotation of their experience during game-play is required. These labels ideally need to be as close to the ground truth of experience as possible. The ground truth in affective computing refers to the unknown label or value that best describes an affective state or experience.

There is a distinction made between two methods of labelling ground truth: annotations can either be self-reported or reported by external observers[239]. It can be assumed there is a disparity between the true experience of players and the experience which they self-report or which is perceived by others. Based on this assumption the player's self-reported experience annotations should normally be closer to their inner experience (ground truth) compared to third-person annotation. However, player self-reports of their experience may suffer from self-deception and memory limitations[238]. These limitations have been attributed mainly to the discrepancies between "the experiencing self" and "the remembering self" of a person[97] which is also known as the *memory-experience gap*[144].

In third-person annotation, an expert or an external observer provides the player state which is considered as a more objective method of annotation since it reduces the described biases of self-reporting experiences. In this case, an expert (or a group) may provide particular player state tags while observing a game session. The benefit of third-person annotation is that multiple annotators can be used to better approximate the ground truth of the player experience labels. However, additional third-person observers can be used only in limited scenarios since it is not as cost-effective as first-person annotations. Additionally, it can be argued that first-person annotations are more reliable than third-person annotations if only a single third-person observer is used for annotations. For these reasons, this research will use first-person self-reports to establish the ground truth of player experience in the user studies conducted in this research.

With regards to the format of the labels, there are three different data types to consider: ratings, classes and ranks. The rating-based format represents a player's state with a scalar value or a vector of values. Ratings are generally the dominant practice for quantitatively assessing aspects of a player's experience, opinion or emotion[240]. The vast majority of user studies have adopted rating questionnaires to capture the opinions, preferences and perceived experiences of experiment participants[49]. The most popular rating-based questionnaire follows the principles of a Likert scale[118] in which users are asked to specify their level of agreement with (or disagreement against) a given statement. Other popular rating-based questionnaires for user and player experience annotation include *the Geneva Wheel model*[196], *the Self-Assessment Manikin*[145], *the Positive and Negative Affect Schedule*[227], *the Game Experience Questionnaire*[86], *the Flow State Scale*[92], *the Player Experience of Need Satisfaction* (PENS) survey[189] (which was developed based on *self-determination theory*[48]) and the Player Experience Inventory (PXI)[6] (which was developed based on Means-End Theory[74, 186]). Although it is dominantly used, research has highlighted some problems with establishing ground truth of player experience when using ratings based questionnaires[238]. Firstly it does not account for interpersonal differences between people and their interpretation of the rating scale. Secondly, ratings are inherently ordinal data however during data analysis these are treated as interval data which is fundamentally flawed[240]. Using techniques such as the Affective Slider[22] which was developed for measuring affect in interactive interfaces (seen in figure 2.5) is particularly interesting. This tool has the potential to overcome the limitation of using the interval-based analyses techniques to a ratings-based measure since it provides interval values for affective dimensions.

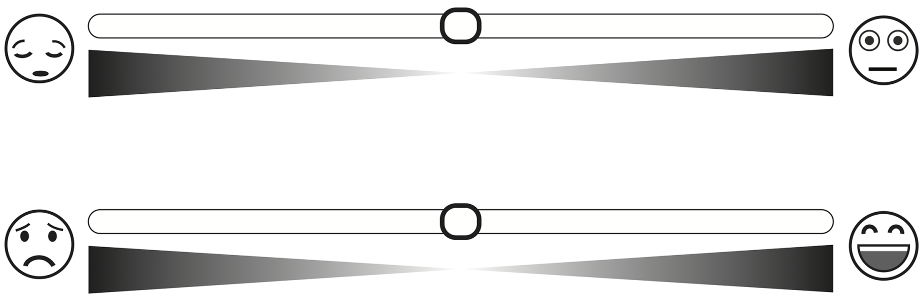


Figure 2.5: Affective Slider proposed by Betella et al.[22], the top slider measures the dimension of arousal and the bottom slider measures the dimension of valence.

The second data type for the annotation of players is the class-based format. Classes allow annotators to select from a finite and non-structured set of options and, thus, a class-based questionnaire provides nominal data among two (binary) or more options. The questionnaire asks subjects to pick a player state from a particular representation which could vary from a simple boolean question to a player state selection form, for instance, the circumplex model of affect[177, 240].

Finally, rank-based questionnaires ask the annotator to rank a preference among options such as two or more sessions of the game[233]. In this case, the annotator compares two options and specifies which one is preferred under a given statement (pairwise preference). With more than two options, the participants are asked to provide a ranking of some or all the options. Examples of rank-based questions include: was this level more challenging than the previous level? Another example of a rank-based questionnaire is the 4-alternative forced choice (4-AFC)[240]. Reporting about subjective experience, preference or emotion via rank-based questionnaires has been growing in popularity in research activities in user modeling[232, 19] and affective computing[216, 244, 135, 237]. This preference for ranking over ratings is due to the reported advantages of ranks minimizing the effects of self-reporting subjectivity biases and findings demonstrating the advantages of ordinal annotation[244, 237] over the rating and class-based data types.

This research uses both rank-based measures through the 4-AFC protocol and interval measures through the affective slider, GEQ and the PXI to measure self-reported experience from players. The 4-AFC protocol is used in initial research work (refer to chapter 3) due to the discussed benefits of rank-based measures however, these measures do not relate to existing techniques of measuring player experience in games that predominantly use ratings based questionnaires. For this reason, this research work introduces approaches of using ratings based questionnaires (which measures continuous data for player experience) to model player experience in AR games (refer to chapters 5 and 6). The next section describes the role of body movement and player engagement in games.

2.3 Body Movement and Engagement in Games

With the advance of body motion-based game technologies such as the Nintendo Wii and Microsoft Kinect, there is a growing interest in the role played by body movement in player experience [146]. Research in embodied interactions suggests that an experience is not predefined by the design of the technology [170, 171, 55]. It is an emergent property of the interplay between the user and the technology where the body is used to mediate this engagement.

This has led to the investigation on the role of body movement in player engagement [26, 90, 141, 88]. Bianchi-Berthouze [26] proposes a taxonomy of body movement that are important to player engagement. This taxonomy consists of (1) Task-Control Body movements (the movements that are afforded by the game interface to achieve game objectives), (2) Task-Facilitating Body Movement (these movements are not required by the game interface, they are consciously or unconsciously performed by the player to facilitate game tasks), (3) Role-Related Body Movements (these are movements performed by the player which indicate a higher sense of presence in the game world), (4) Affective Expression (movements that express the affective state of the player during gameplay), (5) Expression of Social Behaviour (movements that support social interactions between players within a game). In [120] the authors find those game controllers that facilitate a higher degree of body movement engage players more within the game environment. Due to this connection between body movement and player experience, Bianchi-Berthouze [25] argues body movement measures can be used to gain an insight into player experience.

Studies conducted in [164, 151] found that how players used body controllers varied depending on their motivations for playing the game. When the objective was to challenge themselves and win the game, players used these controllers in the most optimal way to maximize their rewards (e.g. minimal and efficient movements while playing a game of tennis on the Nintendo Wii). Whereas, when the motivation was to enjoy the experience rather than winning, players used the controllers more to engage themselves with the role-playing aspects of the game and they derived pleasure from the body movement that the game controller afforded (e.g. using more real-world tennis movement to play the same game) in-spite of this not being as effective a strategy to winning the game. This suggests that body movement can be used to gain insights into player's motivations for playing the game.

Body movement and posture measures to recognize and classify peoples affective states has been explored in the past in some research activities in both game-based and non-game contexts. Early work by Ekman and Friesen [58] has shown that head and body cues can be used by human observers to recognize the dimensions of affect. Paterson et al. [165] mapped speed of head and arm movements to human observers ability to recognize dimensions of affect. Observers viewed and classified acted knocking and drinking motions, and statistical analysis was used to map observers agreement to a 2D affective space. This 2D affective space reflects Russell's circumplex model of valence and arousal[177]. These results also show that the dimension of arousal could be better identified by human observers as compared to valence. Similar observations are made in studies conducted by Kleinsmith et al. [107, 108] and Karg et al.[101]. Kleinsmith et al. examined affective dimensions of whole body posture in an acted scenario [107] which was also followed up in a later study in a non-acted scenario [108]. Their follow-up study [108] which examined non-acted postures used a video game situation and also found a higher degree of agreement with observers for arousal than for valence. Karg et al.[101] investigated acted full-body gait patterns along the dimensions of arousal, valence, and dominance. Similarly, observer agreement was highest for arousal. Clavel et al. [39] further validate these findings in their study. Observers judged the affective state of a virtual agent in face only and posture only conditions and also found that arousal was more recognizable than valence. This indicates that for human observers, arousal can be better identified than valence which has implications for automatic affect recognition systems that use human observers to provide labels for ground truth. This early work shows that the body can be used as an expressive medium to communicate some aspects of emotions to human observers. However, Picard[173] argues that the way in which humans convey emotions or affective messages is affected by factors such as age, gender, culture, and context.

The large number of affect recognition systems that use information from body posture and movement have focused on extracting emotion information from dance sequences [162, 34, 32]. Camurri et al.[33, 34] examined cues and features involved in emotion expression in dance for four affective states (fear, anger, joy and grief). After removing facial information, a set of motion cues was extracted and used to build automatic recognition models. The recognition of fear was the worst, achieving below chance level classification rates.

Several studies investigate affect recognition in non-dance-based scenarios, Pollick et al.[176] conducted a study where they compared affect recognition systems with human observer recognition from body movement in an acted scenario. Actors performed knocking, lifting, and waving actions which were used as stimuli in their study. The results showed that the recognition system was more consistent at recognizing human affect than human observers. Karg et al.[101] examined automatic affect recognition for discrete levels of valence, arousal, and dominance in affective gait patterns. Similar to human observers, recognition rates of the models were best for arousal and dominance, and worst for valence. The results were significantly higher than observer agreement on the same corpus of affective gait patterns reported. The study also used gaits from acted scenarios. Recent work by Sapiński et al. [192] take a similar approach to Karg et al.[101]. They use human gait which is recorded from skeletal information extracted from a Kinect for their predictions on discrete emotional states (happy, sad, surprise, fear, anger, disgust and neutral). Happiness, sadness and fear had high recognition rates while disgust and fear were comparatively difficult to classify. Sanghvi et al.[190] explored affect recognition of body only videos of children playing chess with a robot. They extracted posture and movement features from video and used them to build these models. Their study shows that these features can be useful in detecting the level of engagement during game-play. This study is particularly interesting for this research, however, the game-play scenario used was a static seated situation which is very different from movement-based games.

Kleinsmith and Bianchi-Berthouze have examined automatic recognition of affect from whole-body postures in an acted situation first [27, 106], and conducted a follow-up using a non-acted situation[108]. In their early work[27] on acted postures, they built an automatic recognition model for three discrete categories (angry, happy, sad), and achieved a high average classification rate. As a second step, automatic models were built for recognizing levels of four affective dimensions (valence, arousal, potency, avoidance)[106]. While these models also achieved high classification levels, they were lower than the models for the discrete categories. In later work[108] using non-acted postures and affective states in video games, their models achieved recognition rates lower than their acted studies, but similar to the target rate set by computing the level of agreement between sets of observers.

Bernhardt and Robinson's suggest that although affect can be easily observed in human movement, individual differences in expressing this is noticeable and hypothesize that a classification system must take this into account to improve its accuracy[20]. Gong et al.[70] tested the differences between models with personal biases removed and with personal biases remaining using Pollick et al.'s motion capture database[176]. The automatic recognition rates achieved in both studies were higher with personal biases removed over the rates for the biased motions. Their results were compared with the observers' agreement from Pollick et al.'s study[175] to obtain a baseline on which to validate their models. The results indicated that the recognition models from [20, 70] and the observers' rates from [175] were comparable. Using affective whole body gait patterns, Karg et al.[101] built automatic recognition models to compare differences between inter-individual and person-dependent recognition modes. Similar to previous studies they find that inter-individual recognition accuracies were much lower than the person-dependent recognition accuracies. However, these studies were based on Pollick et al.'s database which used affective actions from actors.

Savva et al.[193, 194] investigated these issues in a non-acted situation. They build an affect recognition system to predict emotional states (High and low-intensity negative emotion, happiness and concentration) of people playing Nintendo Wii tennis. All the emotional states could be predicted with a high degree of accuracy however concentration was the most difficult to predict. Their results further confirmed that person-dependent models perform better than inter-individual models. This shows that it is important to consider inter-personal differences when building affect recognition models based on body movement.

The studies conducted by Savva et al.[193, 194] is of particular relevance to this research since it aims to recognise the affective state of players in a video game context. They analysed player's playing the Nintendo Wii tennis game. Their results showed a high variability of expressions being classified into the same category which was due to the different playing styles. Different players show a varied range of physical movement while playing the game. This has been discussed previously in this section due to differences in motivations behind play (playing to win involves efficient movements while playing for role-playing shows a high range of movement similar to the sport of tennis) [164, 151]. This highlights the importance to investigate movement-based games which are non-acted situations where movement is both task-dependent and is highly varied depending on skills, motivations and engagement during play.

The study conducted by Savva et al.[193, 194] uses full-body motion capture of participants. This is not a feasible approach for players of *local AR games* since these games are played in different environments with mobile devices. Body movement-based affect recognition systems for these environments will have to use the sensors available within the mobile device which are usually the accelerometer and gyroscope. Modern AR SDK such as ARcore and ARKit use these sensors and the camera to return the position and rotation of the device in space. Since these devices are used as a *magic window* for the player in the AR world, the movement of these devices can be analysed as player movement within these spaces. The approach and methodology of how to build affect recognition systems from this movement information is very much an open research question that this research aims to address.

The existing work on affect recognition from movement using mobile sensors is limited. Cui et al.[47] used a smartphone to recognize emotional states of happy, angry and neutral. They primed participants with video stimuli and recorded accelerometer readings of their walking patterns after priming. This work has been continued by Zhang et al.[247] who used a smart bracelet instead of a smartphone. They used personal models for their classification and reported accuracies ranging from 60.0% to 91.3% across all users. Quiroz et al.[182] raised concerns about the validity of the priming process used in the previous studies and conducted a similar investigation focusing on a binary classification of happy vs sad emotions. They investigated two types of stimuli in their study: audio-visual and only audio. They reported classification accuracies of 80%-60% for most users however in some cases classification accuracies were as low as 50% which is close to random chance. From this is it clear the affect recognition from movement using mobile sensors is a challenging area for research. It is also worth noting that these approaches have not been investigated in movement-based mobile game scenarios and it is unclear if affect recognition systems from body movement in these environments are feasible. Since the mobile device used to recognize affect does not just play the role of a passive listener. However, it is the instrument that mediates the game experience as well.

This section has reviewed existing work on engagement and body movement. This research aims to use player movement in *local AR games* to automatically recognize their experience. Additionally, these models will be used online to optimize the difficulty of an AR game based on detected player affect, the next section reviews existing work on dynamic difficulty adjustment in games.

2.4 Dynamic Difficulty Adjustment

Challenge is an important aspect of digital gameplay, if the reason for this challenge is due to a particular game mechanic, it is referred to as functional challenge [40]. The goal of most games is to overcome these challenges to accomplish the objectives of the games. If the difficulty is too high for the player skill, this can lead to frustration, conversely if the difficulty of the game is too low for the player they can easily get bored. Since players can have very diverse skills, it is difficult to design a common difficulty progression for games that fits all players. For this reason research in games has studied Dynamic Difficulty Adjustment (DDA), where the challenge level of a game can be adjusted to fit the skills of the individual player. Denisova and Cairns[49] show that adaptation in games can result in higher levels of immersion in the game. In later studies[50] the authors also show that it is important to consider the player's knowledge about the game adaption, with empirical testing they show that knowledge about game adaption gives players a higher sense of immersion in the game irrespective of the adaptation strategy used in the game. Previous research has proposed several techniques for DDA in games that are based on either player performance or affect-based models. These have been explained in the subsections below.

2.4.1 Performance-based DDA

In performance-based DDA statistical approaches or player modelling techniques can be used to infer the skills of the player, this information is used to adjust the difficulty of the game. Hunicke created the Hamlet system, which uses statistical approaches to determine the appropriate time to intervene in a first-person shooter by giving the player more ammo or a health boost[85]. Zook and Reidl[249] use player modelling techniques to predict a player skill mastery over time. They discuss approaches to using this model to adapt a game's difficulty according to performance curves[248] which is considered as the desired progression of the difficulty of the game (specified by the game designer). In their investigation on the relationship between player immersion and game adaption, Denisova and Cairns[49] use performance-based adaption techniques that adjust the difficulty of the game based on the player's score. Ishihara et al.[91] used a Mote-Carlo Tree Search (MCTS) based approach to dynamically adjust the behaviour of an agent in a 1-on-1 fighting game based on player performance. Results from the user study they conducted indicate that beginners and intermediate level players rated this DDA approach favourably. Another application of MCTS for DDA in a prey-predator game is presented by Hao et al. who use this approach on the game of *Pac-man*[76]. Stephenson and Renz[205] present an adaptation approach to adjusting the level of difficulty of the *Angry Birds* game (a physics-based puzzle game) based on player performance, they evaluated this approach on AI agents that represented models of human players.

In applications of serious games, performance-based DDA is used for physical rehabilitation, Hocine et al.[80] used performance-based DDA to improve training outcomes of stroke patients who require upper limb rehabilitation. Another application of performance-based DDA in serious games is applied to games for learning or cognitive training. Plass et al.[174] explored to what extent performance-based DDA can be used to enhance learning-based games to enhance executive function skills of student learners. Their study found that these skills are improved for older learners (ages 15 and above) with no such benefits observed for younger learners (12 and below).

Another interesting application of performance-based DDA is for game balancing of multi-player games, Baldwin et al.[14] present a framework for DDA in multi-player games that rely on adjusting the abilities of a player's game avatar depend on the performance of their opponents. Moreira et al.[221] show that DDA in Multi-player games has a positive effect on both players, they observed that even when the DDA adjustment is the most noticeable, players report the DDA version of the game as more fun.

A notable limitation of performance-based DDA is that these approaches can be noticed by players, this can potentially result in a negative experience for players. Gerling et al.[67] explored the effect of DDA on a motion-based dance game, they report that if DDA is noticed by player's it has an impact on their self-esteem and they tend to feel cheated because their victory from game-play was not earned based on skill. From their results, it is important to consider the level to which DDA systems can be detected by players during their gameplay. For this reason, Frommel et al.[64] hypothesize that affect-based DDA would be less noticeable for players and thus overcome the limitations of traditional performance-based DDA techniques. Another limitation of performance-based

DDA is the need for a score or metric for player performance, along with a mapping of this performance to the difficulty of the game. Affect-based DDA overcomes these limitations since a model of player affect is used to detect the emotional state of the player and adjust the difficulty of the game depending on the designer's required emotional response.

2.4.2 Affect-based DDA

Affect-based DDA is part of the broader field of affective computing, a term which was coined by Picard[172]. Research in affective computing has shown the benefit of adapting the game experience to enhance player experiences. An example of early work in affect-based DDA is presented by Tijs et al.[213, 214] who propose an affect-based DDA approach to the game of Pac-man where the speed of the enemies is controlled by player affect which is modelled using a number of physiological sensors. Another example of early work is Gilleade and Dix[69] describe a game experience that adapts itself based on player frustration. In later work Gilleade et al.[68] propose three heuristics for emotionally adaptive games: Assist me (refers to the game supporting the player when frustration is detected), Challenge me (which refers to increasing the difficulty of the game when boredom is detected from the player) and Emote me (referring to the game being able to adapt itself to elicit the intended emotional experience from players).

Liu et al.[122] propose a system the automatically detects player anxiety from physiological measures of the player through multiple body-worn physiological sensors. The detected anxiety level was used to adjust the difficulty for participants playing a game. This study is important as it compares performance-based DDA and affect-based DDA in an empirical study. Their results show that the performance of the majority of the participants improved in the affect-based DDA condition, the majority of participants found the affect-based DDA condition more challenging, and participants felt less anxious in the affect-based DDA condition. This study makes a strong case for the use of affect-based DDA to enhance the experience of players. However, a major limitation of their study is the use of a large number of body-worn sensors which is not a practical approach for DDA systems for commercial purposes as they are highly intrusive[240].

In their work on affect-based DDA, Frommel et al.[64] present an approach to emotion-based adaptation in games where players self-reports of boredom and frustration are used to adjust the parameters of the game. Their work does not use a computational model for predicting affect but uses self-reported measures taken from users in-game. To do this the authors use in-game dialogue boxes to allow users to self report their emotional state without interrupting game-play.

However, an argument can be raised that computational models that detect player affect operate in the background of the game while in-game dialogues can still interrupt the game-play experience of players potentially making the effects of the DDA system more noticeable. Their approach was validated in a user study that showed affect-based DDA can have a positive effect on a player's experience.

In recent work on the applications of affect-based DDA in serious games, Bian et al.[24] present a VR driving simulator for individuals with Autism Spectrum Disorder (ASD). Their system uses physiological measures to detect a person's engagement with the environment and adjust the difficulty based on it. They conducted a study that compared this engagement-based version of their system with a performance-based version and found that individuals found the engagement-based condition more enjoyable than the performance-based condition.

This section has described the background of research about DDA in games. It is observed that affect-based DDA has several advantages over performance-based DDA. However, these techniques traditionally rely on body-worn physiological sensors that are intrusive for players[240]. This is potentially another reason for its limited application in commercial games. Furthermore, the previous section has highlighted several techniques to use body movement and posture to detect affective states. However, the use of these body movement-based affect models have not been explored in DDA systems for games. The next section discusses the gaps in the literature presented in this chapter.

2.5 Gaps in Literature

Existing research in mobile AR games shows a gap in the number of empirical studies that model player experience [79, 191, 147]. It has been observed that most of the studies around mobile AR games focus on the creation of novel interactions and game experiences in AR. These studies serve as user validation studies for the various games being tested. Existing research has not focused on the computational modelling of a player's experience for potential use to drive aspects of the game.

Previous research in games has used AI techniques to model a player's experience in digital games such as platformers [167, 166], racing games [216, 215] and prey-predator games [213, 214] among others. These techniques show the potential for using computational models to predict player experience. These models can be used to gain insights into how different aspects of these games can mediate player experience or optimize the game for individual players[239]. However, the potential of these techniques has not been investigated in AR games. These games use embodied interactions where the player movement is an important aspect of such experiences.

Research has shown that body movement is an important aspect of player experience with existing work exploring its relationship to player engagement [26, 90, 141, 88, 120]. However, these studies have been conducted in the context of using full-body game controllers such as the Kinect and Nintendo Wii. These are very different compared to the experience afforded by AR games.

Body movement is an important medium of recognizing affective states by both human observers and automatic affect recognition systems. However, there is a limited number of studies that apply these movement-based affect recognition models in-game contexts. Related to games, these studies have been investigated in contexts where the player is seated playing traditional games[190] or body-based games on the Nintendo Wii [193, 194]. Existing work has not explored the potential of using body movements to automatically recognize various dimensions of PX while playing AR games where the devices act as a magic window into the AR world. Additionally, the study conducted in [190] uses a camera and computer vision techniques to extract body posture information of a seated player and the studies reported in [193, 194] use a motion-capture system to extract full-body movement of the player. For these techniques to be feasible in AR games, affect recognition systems will need to use built-in sensors on the mobile device such as the accelerometer and gyroscope as a measure of player movement. This is a comparatively limited source of movement information as compared to motion-capture systems.

Research has shown that affective states can be predicted from movement data logged by mobile devices in non-game contexts[47, 247, 182] however in these studies, the device acts as a passive listener that makes inferences from a person's walking style after emotional priming. It remains an open question if it is possible to use similar techniques in an AR game context where mobile devices is an active mediator of the experience. This is a more complex problem since in these games player experience and their body movement emerge from the interplay between players and the game world. The use of body movement as a lens to recognize the emotional state of the player would be highly beneficial to AR games for two reasons. First movement data could be a generic tool for modelling player experience across many AR games and secondly, these player models can be used to automatically adapt the game for the individual player to enhance their engagement with the game.

Parallel research in DDA has shown that experience-based models can be used to tune a game experience for players[122, 214, 213, 24]. These techniques usually rely on several physiological sensors which are worn on the body. A limitation of this approach is that body-worn sensors are intrusive and can interfere with player experience[240]. Furthermore, AR games (especially on mobile devices) are played in different environments and involve the player's movement through their space and potentially across different locations. This makes the use of body-worn physiological sensors less feasible for commercial use.

A gap is observed in experience-based DDA research studies that use body movement information to model PX to drive adaption within the game. These approaches would be ideal for AR games since PX prediction can be accomplished by sensors built into the device making the presence of the adaption system less noticeable to players (which the literature in DDA for games indicates is an important aspect of these systems [67]).

This research will aim to build on these gaps in the literature as it will be the first known research activity (to the best of the author's knowledge) that explores the potential of using body movement to model PX in AR games. Specifically, the research will explore key areas of PX in games, namely player preferences, player experience and player motivations. It will also be the first know research activity to investigate to what extent the described PX model can be used to adapt the difficulty of AR games.

2.6 Research Approach

This research will focus on AR games that are played on mobile devices, specifically on how movement data logged from the device while playing the game can be used to model the player's experience. To address this problem domain, this research involves the development of two AR games that are used in user studies. These user studies test the hypothesis that player movement in mobile AR games can be used to model the PX. Each of the games will use a number of parameters to generate the game levels for players. The games will be parameterized in such a way that different parameters will invoke a spectrum of emotional responses from players. Players will be asked to self-report their game experiences from different rounds of these games. The recorded movement data and additional game context features will be analyzed using supervised learning to predict the players' self-reported experience.

The first game developed is an *AR treasure hunt* game which is exploratory in nature. Players will be asked to explore their space collecting treasure items. In this game, players will be motivated by their intrinsic desire to explore their physical surroundings. The game objective does not bias their movement within the level. If a player does not wish to explore the space they can choose to end the game (however their rewards are impacted if they do not explore the space completely). This serves as an ideal testbed to test the research hypothesis since their body movement should reflect their experiential state (keeping movement biases introduced by the nature of the game task to a minimum). Players are asked to self-report their experiences in a rank based manner. This is accomplished by asking players to make several pairwise comparisons of their experiences across two different games levels (e.g. which of the two games was more fun?) using the 4-AFC protocol. Rank-based measures of player experience are used due to the reported advantages of ranking approaches over ratings scales[237, 238]. The PX models tested in this study are based on the discreet emotional states of players. The study models the emotions of boredom, challenge, excitement, frustration and fun which have been emotions of interest in previous player modelling studies[128]. Details of the game and user study conducted have been reported in chapter 3 of this thesis.

To test the generalisability of findings from the first study, another AR game is developed for use in the second user study of this research work. This game is a physical exertion game titled *Running Chickens*. This is a different game environment where players accomplish a target acquisition task that actively biases the body movements of players. Here, the game actively directs the player's body movements within the game space to accomplish game objectives. This game serves as an appropriate test-bed to investigate the generalisability of the PX prediction method proposed in the first user study in a different game task. Additionally, this game is parameterized in such a way that different parameters of the games will result in different degrees of functional challenge[40] for the player. This study builds on the first study by using ratings based measures of player experience. The measures of valence and arousal are be collected at the end of each round of the game using the affective slider[22] and measures for several dimensions of player experience is collected using the Game Experience Questionnaire (GEQ)[86]. This is done to better understand the relationship between player movement in these games and player experience and to investigate the potential of using supervised learning to model these dimensions of player experience. Details for this study are provided in chapter 4.

Finally, the validity of the PX models investigated in this research will be tested for DDA of the *Running Chickens* game developed. The adaptive version of the game will tune the difficulty of the game based on the predicted experiential state of the player. The goal of this study is to investigate to what extent experience-based DDA techniques can be used to enhance the experience of players in an AR game. For this, two conditions of the game will be tested (linear difficulty progression vs. experience-based DDA) with players in a within-participant study design. After each condition qualitative and quantitative measures of player experience are gathered from players. This study validates the potential of using movement-based player modelling techniques to enhance the player experience in AR games. Details of this study are provided in chapter 5.

2.7 Chapter Summary

This chapter has described the previous research activities that are relevant to this thesis. As shown the research in AR games has focused on domain-specific proof of concept game applications that take the approach of adapting game mechanics from the real world or digital games for a mixed reality platform. The research activities conducted around these games focus on validating the specific games for playability and enjoyment with users.

When players engage with mobile AR games they use their device as a magic window into the environment. This interaction involves the player's body movement around their environment. Research in affect recognition has shown that the bodily expressions of posture and movement can be an expressive medium of human emotions. Parallel research in affective computing has shown that models of player's affective states can be used to adjust the difficulty of a game to enhance the player experience and for personalization.

It is possible that mobile AR games can use player movement to recognize their emotional state and dynamically adjust aspects of the game to enhance the player experience of these games. However, research has not explored this hypothesis and it is unclear to what extent player experience models built from body movement can be practically applied to adapt the game experience. This is the main research goal of this research.

This chapter has also discussed the approach taken in this research to build on existing research and contribute to gaps in the literature around player experience modelling in mobile AR games. While related research work has found the game-based, performance and behaviour data can be used to model players experience in traditional (non-AR) games. This research work investigates player modelling in mobile AR games where player movement data is considered as a behaviour measure along with other game metrics to create models that predict player experience using supervised learning techniques. These models are further evaluated for their effectiveness to drive experience-based dynamic difficult adaptation in mobile AR games to personalise the game experience for individual players. The next chapter presents the first user study that has been conducted as part of this research using the *AR treasure hunt* game. Details about the games, methodology of the user study and data analysis techniques applied and results from this study are presented in the next chapter.

Chapter 3

Modelling Player Preferences in an Exploration based AR mobile game.

The previous chapter discussed the gaps in existing literature, the approach taken and the scope of the thesis. The main gap addressed in this research is to explore approaches to modelling player experience in AR games. Although player modelling is a popular topic of research in game AI, previous studies have not explored this problem in AR mobile games.

This chapter reports the results of a study conducted to test the hypothesis that player movement measured from mobile device IMU sensors can be used to model a player's emotional experiences in these games. This is a complex problem since AR games involve the player's movement through a physical space to accomplish tasks within the game.

The study conducted in this chapter shows that in-game player movement can be used to predict a player's emotional preferences regarding variants of an AR treasure hunt game. It is interesting to note that this technique does not require any body-mounted sensors such as ECG, HR or GSR and uses the built-in sensors of the mobile device. This is advantageous for two reasons: first, it is cost-effective and existing mobile devices can be used to model player experiences with no additional sensors. Second, additional body-worn sensors are intrusive and can break the level of immersion players experience as they engage with games [240]. Since previous research has found body movement within movement-based games can have an impact on game engagement and enjoyment of the experience and that body movement can act as an indicator of a person's affective experience [26, 90, 141, 88]. The majority of research in this area has focused on movement game controllers such as the Nintendo Wii [193, 194], mobile Augmented Reality games have not been investigated within this context. In mobile AR games, the device uses movement and computer vision to mediate the game experience. This thesis argues that this movement data can be additionally used to infer a person's game experience which will allow these types of games to be tailored to an individual to further increase game engagement and enjoyment. Other research has begun to explore the potential for mobile devices to detect a person's affective state [182, 47] however, current studies do not explore this within a game or player experience context.

This chapter first presents the aims and motivations of this study in section 3.1. The AR treasure hunt game developed to test the hypothesis is described in section 3.2. Section 3.3 provides information on the experiment protocol used in this study. Section 3.4 provides information about the pilot studies conducted as part of this experiment. Section 3.5 describes the data collected as part of the experiment. Data analysis and results from the study are presented in section 3.6. Section 3.7 discusses the implication of these results and the limitations of this study. Finally, section 3.8 presents a chapter conclusion.

3.1 Aims and Motivations

AR experiences have grown in popularity recently. The most popular medium for AR games are mobile devices, possibly due to the increasing simplicity of building and deploying mobile AR experiences. Popular AR SDKs such as ARKit and ARCore use the camera and inertial sensors to extract the device's position and orientation[121]. This information is then used to overlay digital content in the space around the device, which can be viewed through the device. While this movement data (the device's position and rotation over time) is used to create game experiences, this study explores to what extent it can be used to model a player's experience from the game.

This study focuses on AR games that involve players' physical exploration of their local space. Physical AR mobile games are the focus of this study for 2 reasons. First, they follow trends in mobile AR games that are more relatable to past experiences players have had, such as playing *Pokemon GO* (PoGo), *Ingress*[1] or *dARk*[2]. These games use narrative and game design to engage people with their surroundings (albeit in a more complex manner than this study

[1] PoGo and Ingress are both location-based AR games developed by Niantic
[2] dARk is a short story horror experience developed by Combo studio

game). Second, the potential health benefits of physical AR games make a strong case to personalize these environments to promote healthy behaviour among players. Player movement, which is the time-series data of the device's position and rotation during the game is used to model their emotional preferences. In turn, this data is treated as player behaviour in this work.

Following similar studies in player modelling [167, 166, 216] in traditional video games, player behaviour features (PBFs) and game parameters are used to predict a player's emotional preferences. For example, does the player find level A more fun than level B or vice versa? Effective prediction of such preferences will enable procedural content generation (PCG) or game balancing systems to be optimized to a player's ideal emotional preferences. This approach is novel, in that it uses movement data to model players' preferences. Ground truth is established through data collected from self-reported questionnaires. Exploring this domain in the context of subjective preferences is useful for personalizing game experiences in AR. The following research question is investigated in this study:

To what extent can measurements of player movement be used to predict their emotional preferences regarding variants of an exploration-based AR mobile game?

Since data from popular AR mobile games are unavailable, this study included the design and development of an AR Treasure-hunt game. This game is intended to be similar to existing games by incorporating reward systems, exploration of local space, and a narrative that motivates these rewards. This game can serve as a standardized game task in the future for similar studies about player modelling in AR mobile games.

3.2 The AR Treasure Hunt Game

The AR Treasure Hunt game described in this section was designed and developed to explore the relationship between player movement and emotion preferences. The game design decisions were made to create a game that could be parameterized to invoke a wide variety of emotional responses from players. Another crucial aspect of the design was that a player's movement within their local space was not biased by the game, i.e. the game did not direct the player to specific areas within the game level or in a specific direction. Players were allowed free exploration of constrained space. For this initial study, designing a game experience that does not bias a player's movement is important to explore the relationship between the game experience and player movement.

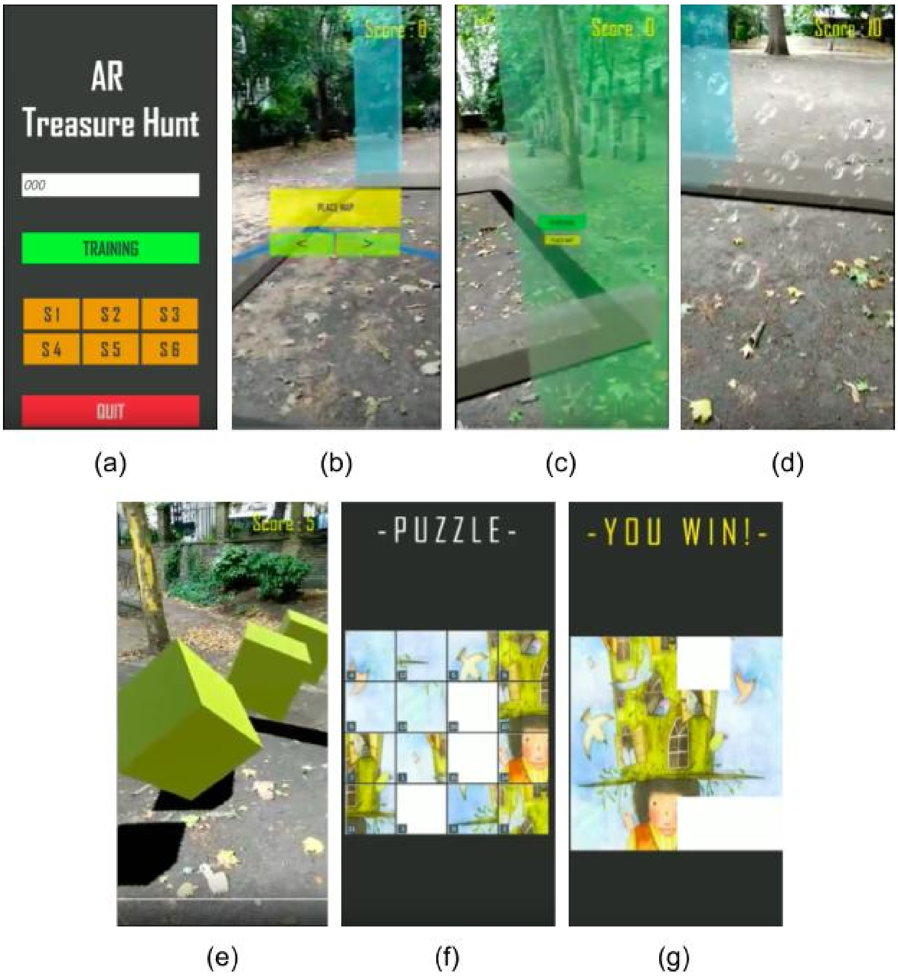


Figure 3.1: Fig. [a-g] show the flow of a single round of the game. [a]: Shows the screen to select an experiment session. [b]: The options for a user to place a game map. [c]: The (green) start button that the player must tap in order to begin the game. [d]: The bubbles in the game indicating treasure in close by. [e]: Treasure that appears which the player collects. [f]: The 2D puzzle presented to the player in the unsolved form. The white squares indicate treasure pieces that were not collected. This screen is presented to the player once the exit area is entered (seen in blue in fig [b] and [d]). [g]: The solved 2D puzzle.

(a) (b)

Figure 3.2: Figure a-b shows a participant playing a single round of the AR treasure hunt game. [a]: Participant playing a round of the game. [b]: View of the AR content when viewed through the mobile device.

The aim of the game is to collect all hidden treasures from a constrained space. This treasure has been randomly distributed within the level area. All these treasure pieces together form a picture. To win, the player will have to put together this picture (like a 2D puzzle). Figure 3.1 shows the screens during a single round of the game. Levels of the game vary across the number of treasure pieces and the size of the game area. This game can be played in parks and other open spaces. Indoor spaces such as a gym can be used as well, the main constraint is that this version of the game requires a space free of obstacles that are at least $35 \times 35m^2$. Figure 3.2 shows a participant playing a round of the AR treasure hunt game.

The game design mechanics has been based on Malone's taxonomy for 'fun' in games [127] which are challenge, fantasy and curiosity. It is important to note that Malone developed this taxonomy in the context of educational games however, these are relevant for other applications of serious games such as applications for physical fitness. Game challenge refers to the system providing players with goals whose attainment is uncertain. These challenges have a direct impact on a person's self-esteem. Successfully overcoming these challenges can have a positive impact on a person's self-esteem and sense of competence which is an important factor of keeping players motivated in the game [181, 188]. It is also important to consider the balance between the player's skill and level of challenge (or difficulty) of the game since an appropriate balance between these two factors is important to put players in a state of flow [208, 209, 149]. Failure in a game challenge will lead to low self-esteem, lowered perceived competence and a decrease in desire to play the game. Additionally, if a player's skill is greater than the level of challenge required by the game, this will lead to boredom with the game and also lower a player's desire to engage with the game. Fantasy is the ability for computer games to invoke images of physical objects or social situations that are not actually present. Fantasy in computer games is important as they satisfy the emotional needs of players. Finally, curiosity refers to a player's motivation to engage with the game independent of any goal-seeking or fantasy fulfilment. The game's environment must be novel and surprising for the player but not completely incomprehensible to maintain their curiosity with the game. The following subsection describes the design decisions that were taken to incorporate Malone's taxonomy for fun[127] in the development of AR Treasure Hunt.

3.2.1 Game Play

The player places the AR level in the world before starting the game (fig 3.1[b-c]). The game begins once the player finalizes this placement. After the level is placed, the player must enter a start area to begin the game (seen as a green column in fig 3.1[c]). Once the player enters this green column they are free to explore the game level to discover and collect the treasure pieces.

While a player is exploring the level, the boundary of the AR level and an exit (to the 2D puzzle) is visible to them (the exit is a blue column seen in fig 3.1[b & d], while parts of this boundary wall are seen in fig 3.1[b-d]). They will not be aware of how many treasure pieces are hidden in the level. This is important since it will maintain a high level of curiosity and fantasy within the game experience [127].

These pieces are invisible to the player by default. If a player is close to a treasure piece they will receive an audio-visual clue. The clues are implemented using a particle system that is designed to look like a bubble emitter with an appropriate sound effect (the bubble are seen in fig 3.1[d]). Bubbles are emitted around the position of the treasure piece. Treasure appears only if the player is close to it (the treasure pieces are seen in fig 3.1[e]). Once it appears the player can collect the item by moving the mobile device into it. Collecting treasure increases the player's score by +1. The player is instructed to explore the level until they believe they have discovered all the hidden treasure and then move to the exit. Only when the player moves into the exit square, they are shown the 2D puzzle (in the shuffled order) and all the treasure pieces that were not collected appear to them as white squares. This is when the player will get confirmation if all pieces were collected or not. This design decision ensures that players are motivated by the exploration and discovery of dynamic content within the game space. Informing the player beforehand of the maximum number of pieces hidden in that level would reduce the sense of exploration and discovery, which is an important aspect of these real-world games. To ensure that the puzzle can be completed even when a low number of treasure pieces are found, numbers have been added to each of the puzzle pieces, these numbers indicate where they must be correctly placed to complete the level. The puzzle along with the white square is seen in fig 3.1[f].

The player wins the round once the puzzle is completed. The game has been designed in such a way that all game levels can be completed. However, the reward for the player varies depending on the amount of treasure they collect, which corresponds to the amount of the picture they get to appreciate at the end of the game round. The 2D puzzle has been added to ensure some amount of challenge which is important in game experiences [127]. The game interactions were designed to be simple to keep the cognitive load from the UI on the players low. The game was developed in Unity, using their experimental AR interface to handle the device and environment tracking for the game. The Unity asset store was used for assets in the game. The mobile device used for development and testing was the Google Pixel 2 XL.

3.3 Experimental Design

The study design was informed by previous studies that model player experience for content creation [167, 166, 216]. This protocol is used to build a data-set of player movement data and corresponding emotional preferences in AR game sessions. The study consisted of a number of sessions of the same format. In each session, participants played 2 rounds of the game with different game parameters in each round (resulting in varying levels that created a spectrum of emotional responses from players). Participants were not given any constraints on how to hold the phone (in portrait or landscape). They could use either hand to hold the phone depending on comfort. 40 participants played the game in portrait mode and 2 participants played the game holding the phone in the landscape orientation. When playing the game participants tended to use their dominant hand to hold the phone (left vs right handed). Since the focus of this study is on modelling a player's emotional preference, pilot studies were conducted to identify appropriate game parameters that could create a diverse range of emotional responses from players. The 2 chosen game parameters were:

- *The Area of the Level (G_A):* 2 sizes of levels are compared. *Large Area (LA)* levels are $\approx 30m \times 30m$ and *Small Area (SA)* levels are $\approx 5m \times 5m$. $G_A \in \{LA, SA\}$

- *Treasure in Level (G_T):* 2 amounts of treasure are compared: Low Treasure (LT) with 9 and High Treasure (HT) with 16 pieces respectively. $G_T \in \{LT, HT\}$

Using 2 game parameters has resulted in $4(2 \times 2)$ levels being compared: (1) $LA \times HT$ (2) $SA \times HT$ (3) $LA \times LT$ (4) $SA \times LT$. Since games are played in pairs, the total number of game pair combinations is 6. This study has focused testing on 3(out of 6) of the game pairs:

- (1) $LA \times HT$ vs (4) $SA \times LT$
- (2) $SA \times HT$ vs (3) $LA \times LT$
- (1) $LA \times HT$ vs (3) $LA \times LT$

The complete comparison space is not explored because real-world optimization for player preferences would rely on similar incomplete data-sets. The current choice of 2 binary variables as game parameters is adequate for the purposes of this exploratory study. It would easily become unfeasible to collect pairwise preferences of the complete comparison space if a more complex set of game parameters is used.

At the end of each game pair, the participants were given a 4-AFC protocol [167, 166]. This is a questionnaire that ranks the 2 games according to different dimensions of emotion preference. The 4-AFC protocol collects the a player's preference data for a particular emotion preference. For example, given 2 games, which game does the participants find more *Fun*. This study focused on *Boredom, Challenge, Excitement, Frustration* and *Fun* as dimensions to measure player preference; since previous research[128] has shown that these states are relevant to digital game-play. Following shows the 4-AFC protocol measuring the dimension of *Fun*:

Please select 1 of the following options

1. Game 1 felt more *Fun* than Game 2
2. Game 2 felt more *Fun* than Game 1
3. Game 1 and Game 2 felt equally *Fun*
4. Neither of the two games felt *Fun*

The same format is used to measure each dimension of preference. This data is used as ground truth for players' preferences between pairs of games. The study began with a briefing for each participant which included a training session on the game and the structure of each session of the study. This was followed by a trial session in the described format; data from this session is discarded. For the trial session, 2 levels were designed with different areas (*LA, SA*) containing 4 treasure pieces each. The trial allowed participants to familiarize themselves with the game and study format. The trial was followed by 3 experiment sessions.

All experiments were conducted during daylight and adequate weather conditions (no signs of rainfall) in a park near Queen Mary University of London's Mile End campus. This is done to minimise the difference in results that may arise due to different locations or poor lighting and weather conditions. The order in which participants experienced each session of the experiment was randomised to minimise ordering effects on the data collected. Participants were anonymised using IDs and were compensated with a hot beverage for their participation in the study. The experiment was conducted using a Google Pixel 2 XL mobile device. The following subsection describes the study procedure used for each participant.

3.3.1 Procedure

All participants provided informed consent before participating in the study. At the beginning of the study, participants filled up a questionnaire about their background and previous experience in MAR games. After which they were given a training session about the game and the questionnaire used in the study. During training, the researcher first demonstrated how the game works to the participant over one level of the game, after which participants played 2 training levels and filled up the 4-AFC protocol after each game pair. During training, the researcher was present with them and they were encouraged to ask any questions about the game or the study procedure.

Once the training was completed, the participants were left alone in the park (while the researcher waited by the entrance of the area), to minimise the effect of the researcher's presence on the data collected. During this time, participants experienced the 3 study sessions of the experiment. In each session, the participant played a game pair and completed the 4-AFC protocol. Participants were asked to take a 2-5 min break between sessions to minimise the effects of physical fatigue from the previous session on the data collected from the next one.

While playing *AR Treasure Hunt*, if the participants experienced any technical issues (the main one being the AR algorithm losing tracking of the environment), they were asked to proceed to the next game and the data from this session was not used in data analysis. At the end of the study, participants were debriefed about the objectives of the research, all the questions were answered and the experiment was concluded. The experiment took 40-60 min for each participant (depending on the length of the breaks they took during the study).

3.4 Pilot Study

Aspects of the AR Treasure Hunt game design and experiment protocol were informed using pilot studies described in this section. The study was designed with two rounds of pilot testing. The first pilot test used a single participant across a period of three days and was used to test three variants of the experiment design. A single version of the experiment was tested on each day giving the participant time to recover from fatigue between tests. The first version had the participant completing 12 sessions of the experiments with pairs of games used in each session. This version of the study took 90 min to conduct. This version was not used since the participant got easily fatigued and found this format frustrating.

Another version that was tested was a short format of the study where participants experienced 4 sessions of the study where each session consisted of a single game as opposed to a pair of games (where the participant was required to provide a rating instead of a ranked preference). This version of the study took 20 min to complete. This method does not provide as much data per participant and also post-session interviews showed that the participant found it was easier to compare two variants of the game (which was the previous study design tested) instead of providing a single subjective rating for each of the game's variants. For these reasons this approach was not used and the next method is used as the final experiment design.

The most satisfactory version was found to be where participants experience four sessions (one training and three experiment sessions) of the game. In this version, each session consisted of a game pair and participants were asked to compare the two games across a number of emotion dimensions using the 4-AFC. This version of the study took 60 min to run (this included a few mins break between sessions) for each participant.

This was the most optimal variant of the tested experiment designs since it collected a larger number of data samples per participant (as compared to the previous version tested), was easier for participants (ranking game pairs was easier for them than providing ratings for each emotion) and reduced the amount of participant fatigue (as compared to the first version of the study design piloted). Therefore, this version was carried forward in a second round of pilot testing with two additional participants. The first round of pilot testing also revealed some minor aesthetic problems in the game. For example, the participant found the shadow cast by the exit point hindered their view of the game level. To fix this, the shadow for this object was not rendered in later versions of the game.

The second round of pilot testing was conducted with two participants. The experiment took an hour for each participant. The second round of pilot testing was conducted as expected. Another major learning from the pilot activities is the need for a demonstration session showing the game-play and task flow of a single session so participants will have a better orientation to the experiment design. This was necessary as all the pilot participants (although being familiar with mobile games), faced a learning curve while using placement controls to place the game map in the physical world. The data from both rounds of pilot testing have been discarded and will not be used in analyses.

3.5 Data Collection

During the study, player behaviour and preference data were collected. As each session consisted of comparing 1 game pair, each subject contributed 3 game pairs of preferences resulting in 126 game pairs (252 individual games). However, due to some software crashes, only 117 game pairs were successfully recorded and used in the data analysis. Software crashes occur when the AR algorithm loses tracking of the devices position and orientation with the real world. This section describes the questionnaires, qualitative data, movement data and game metrics collected during the study as well as the demographic information of the sample of participants in this study.

3.5.1 Participants

Participants were recruited using university mailing lists which included PhD, Masters and Undergraduate students from the Electronic Engineering and Computer Science at Queen Mary University of London. The study sample consisted of 42 volunteers (17 female and 25 male) aged 18-44 (22 were 18-24, 9 were 25-29, 6 were 30-34, 4 were 35-39, 1 was 40-44) took part in this study. Participant's age breakdown has been summarised in table 3.1 When asked about prior experience playing AR games 19 subjects had no prior experience. In the remaining 23 of subjects: 12 reported having only one experience in the past, 10 played a few times before, and 1 participant played AR games regularly. Participants' experience in AR games has been summarised in table 3.2

3.5.2 Emotional Preference Data

The 4-AFC collects preference data between game pairs along dimensions of *Boredom, Challenge, Excitement, Frustration* and *Fun.*

Table 3.1: Summary of participants' ages.

Age range	Number of participants
18-24	22
25-29	9
30-34	6
35-39	4
40-44	1

Table 3.2: Summary of participants' previous experience with AR games.

Previous AR experience	Number of participants
No experience	19
Played only once before in the past	12
Played only a few times in the past	10
Played AR games regularly	1

3.5.3 Player Behaviour Data

Player Behaviour data is measured from player movement in-game sessions. The mobile IMU sensors record the position and rotation of the device during the game. This data is recorded at a frequency of *64 Hz* following guidelines from Preece et al.[178] and the discrete-time signals are stored as a 6-dimensional vector: $\alpha \in \{P_X, P_Y, P_Z, R_X, R_Y, R_Z\}$ for position and rotation. The first 3 elements of α correspond to the position of the mobile device along the x, y and z axis measures in meters. While the last 3 elements of α correspond to the rotation of the device along the 3 axis measured in radians. The sampling frequency used is the highest possible sampling frequency that can be recorded during gameplay using the Pixel 2 XL mobile device. This sampling frequency could vary depending on the hardware used. The player's score (S), which increases as the treasure pieces are collected, is recorded at the same frequency.

3.5.4 Qualitative Data

A structured interview was not conducted in this study however, participants were asked to provide comments about their experience at the end of the study. The goal of these comments was to check the validity of the methodology of this experiment and to make improvements to the method for future investigations.

3.6 Analysis and Results

The resulting dataset from the study was used to explore preference learning approaches to modelling players' preferences. The data is pre-processed and PBFs are extracted. These features along with the game parameters were used to model players' preferences. To better understand the effects of each feature, and to explore to what extent noise from ordering effects have biased the data, statistical analysis has been conducted on the features. This section first describes the method used for data pre-processing and feature extraction followed by the results of the statistical analysis and preference learning evaluations conducted as part of this study.

3.6.1 Data pre-processing

Pre-processing reduces noise and redundancy in the data. This stage is adapted from Li et. al. [116] and is broken into these steps: Data Segmentation, Low Pass Filtering, Coordinate Difference, and Dimensionality Reduction.

Data Segmentation Since the study focuses on in-game behaviour, movement data from when players were interacting with the mobile device before gameplay (e.g., while confirming the placement of the AR level in the physical space) was discarded.

Low Pass Filtering The segmented data may be noisy and contain unwanted high-frequency components. In order to reduce this, a Gaussian filter with coefficients from Li et. al. [116]: $h = \frac{1}{16}[1, 4, 6, 4, 1]$ has been applied. The filter is a 1D convolution of the Gaussian filter and each column of the raw data in α (details in section 4.5.3) given by the following equation:

$$y(n) = \sum_{t=-\infty}^{\infty} x(t)h(n-t) = x(n) * h(n) \tag{3.1}$$

In eq 3.1, x is a column of the vector α (raw data), h is the Gaussian filter and $*$ the convolution operation.

Coordinate Difference Movement qualities such as velocity($\dot{\alpha}$), acceleration ($\ddot{\alpha}$) and jerk[3] ($\dddot{\alpha}$) are extracted for each of the columns of the vector α[20].

$$\dot{\alpha}(t) = x(t) - x(t-1) \tag{3.2}$$

$$\ddot{\alpha}(t) = \dot{\alpha}(t) - \dot{\alpha}(t-1) \tag{3.3}$$

[3] Jerk is the derivative of acceleration

$$\dddot{\alpha}(t) = \ddot{\alpha}(t) - \ddot{\alpha}(t-1) \tag{3.4}$$

This step outputs 3 6-D vectors for velocity ($\dot{\alpha}$), acceleration ($\ddot{\alpha}$) and jerk ($\dddot{\alpha}$). These 3 vectors contain data for velocity, acceleration, jerk, angular velocity, angular acceleration and angular jerk along the x, y and z axis. Analyzing the quality of movement in this way minimizes the impact of inter-participant differences in holding the phone on the preference predictions.

Dimensionality Reduction To reduce the dimensionality of the feature space, the Euclidean norm of the x, y and z axis for velocity, acceleration, jerk, angular velocity, angular acceleration and angular jerk is computed. This output 6-D vector along with the score is the final output of the data pre-processing phase for each game, given by $\beta \in \{V, A, J, RV, RA, RJ, S\}$ at $64Hz$. The units for each of the components of β is provided in table 3.3.

Table 3.3: Units of measurement for components of the β vecotor obtained from data pre-processing

β **component**	Unit of measurement
V	meters per second - m/s
A	meters per second square - m/s^2
J	meters per second cube - m/s^3
RV	radians per second - rad/s
RA	radians per second sqaure - rad/s^2
RJ	radians per second cube - rad/s^3
S	no unit as it is the count of the game score

3.6.2 Feature Extraction

PBFs are extracted from the pre-processed movement data: V, A, J, RV, RA, RJ (first 6 dimensions of the time series vector β), table 3.4 shows the 10 features that are extracted for each dimension resulting in 60 movement features. The S signal (last dimension of β) is used to compute 2 features:

- Completion (C): The fraction of the score at the end of game divided by maximum possible score from game, given by: $C = \frac{Score \quad At \quad End \quad Game}{Max \quad Game \quad Score}$
- Score Rate (SR): The fraction of the score at the end of the game and the time taken to reach it (note: this is different from the total time of the game), given by: $SR = \frac{Score \quad At \quad End \quad Game}{Time \quad To \quad Reach \quad Score}$

The final feature considered is the time of the game in seconds (T). This results in 63 PBFs combined with 2 game parameters, giving us 65 features in

Table 3.4: Extracted movement features

Feature X	Description
X_m	Mean
X_{std}	Standard Deviation
X_{sk}	Skew
X_{kur}	Kurtosis
X_{min}	Minimum value
X_{max}	Maximum value
X_D	Max - Min
X_{tMin}	Time of Minimum value
X_{tMax}	Time of Maximum value
X_{tD}	Time of Max - Time of Min

total which were used in the following analysis.

3.6.3 Statistical Analysis

Statistical analysis was conducted to check for ordering effects in the data and to understand the relationship between features (PBFs, game parameters) and emotional preferences. The Chi-square test is used to check for ordering effects in preference data, which is based on the number of times subjects expressed a preference for the first or the second game in the pair. The Chi-square test is also used to check for statistically significant effects of the 2 game parameters on preferences since these are binary categorical features. The Wilcoxon signed-rank test was used to check for significant effects of the PBFs on preferences as these are continuous features. All tests for significant effects use a p-value $< 1\%$.

This study followed the method to compute correlation coefficients from [236], given by $c(z) = \sum_{i=1}^{N_s}\{z_i/N_s\}$ where N_s is the number of pairs where subjects expressed clear preferences for one of the two games (picking the first 2 options of the 4-AFC), and $z_i = 1$ when the subject preferred the game with the larger value of the examined feature, and $z_i = -1$ when the subject chooses the other game. From the 117 game pairs that were analyzed N_s is 59, 105, 89, 81, 106 for *Boredom, Challenge, Excitement, Frustration* and *Fun* respectively. Variance in N_s shows that subjects find it difficult to express a clear emotional preference between game variants. This is especially observed in the N_s for boredom (59) which indicates that participants have a clear preference for boredom only 50.42% of the time.

Results from order testing (to check if the ordering of the game has created noise in the preferences) and the correlation analysis of statistically significant features are described below for each dimension of emotional preference tested.

Boredom Participants had a preference of *Boredom* 50.42% of the time. Order testing showed a significant ($p = 0.001$) effect: subjects tended to find the second game more boring. Table 3.5 shows the significant PBFs and game parameters.

Table 3.5: Statistically significant (p-value < 1%) correlation coefficients for *Boredom*.

Feature	**Symbol**	$c(z)$
Controllable Level Features		
Area of level	G_A	0.339
Treasure in level	G_T	-0.322
Player Behaviour Features		
Acceleration Mean	A_m	0.390
Acceleration Standard Deviation	A_{std}	0.390
Maximum Acceleration	A_{max}	0.458
Max-Min Acceleration	A_D	0.458
Jerk Mean	J_m	0.424
Jerk Standard Deviation	J_{std}	0.458
Maximum Jerk	J_{max}	0.424
Max-Min Jerk	J_D	0.424
Velocity Mean	V_m	0.390
Time	T	0.390

Challenge Participants had a preference 89.74% of the time. Order testing was not significant. Statistical testing of the features showed that area of the level was the only game parameter that had a significant effect (details provided in table 3.6) while 39 PBFs showed a significant effect. Only the top ten correlation coefficients are reported in table 3.6.

Excitement Participants had a preference 76.06% of the time. Order testing was not significant. Among all the features, only *Treasure in Level* G_T (a game parameter) had a significant effect with $c(z) = 0.339$.

Frustration Participants had a preference 69.23% of the time. Order testing was not significant. Feature tests showed that area of the level was the only game parameter that had a significant effect (details in table 3.7) while 38 PBFs showed a significant effect. Only the top ten correlation coefficients for *Frustration* are reported in table 3.7.

Table 3.6: Statistically significant CGFs and Top ten statistically significant (p-value $< 1\%$) PBFs correlation coefficients for *Challenge*.

Feature	**Symbol**	$c(z)$
Controllable Level Features		
Area of level	G_A	0.686
Player Behaviour Features		
Time of Max Acceleration	A_{tMax}	0.467
Time of Max Velocity	V_{tMax}	0.476
Time of Max Ang. Acceleration	RA_{tMax}	0.504
Time of Max Ang. Jerk	RJ_{tMax}	0.523
Maximum Ang. Velocity	RV_{max}	0.467
Max-Min Ang. Velocity	RV_D	0.467
Time of Min Ang. Velocity	RV_{tMin}	0.467
Time of Max Ang. Velocity	RV_{tMax}	0.504
Score Rate	SR	-0.714
Time	T	0.771

Table 3.7: Statistically significant CGFs and top ten statistically significant (p-value $< 1\%$) PBFs correlation coefficients for *Frustration*.

Feature	**Symbol**	$c(z)$
Controllable Level Features		
Area of level	G_A	0.691
Player Behaviour Features		
Maximum Acceleration	A_{max}	0.481
Max-Min Acceleration	A_D	0.481
Time of Max Jerk	J_{tMax}	0.530
Time of Max Velocity	V_{tMax}	0.444
Time of Max Ang. Acceleration	RA_{tMax}	0.481
Time of Max Ang. Jerk	RJ_{tMax}	0.555
Minimum Ang. Velocity	RV_{min}	-0.481
Time of Max Ang. Velocity	RV_{tMax}	0.456
Score Rate	SR	-0.679
Time	T	0.802

Fun Participants had a preference 90.60% of the time. Order testing showed a significant ($p = 0.002$) effect, subjects tended to find the first game more fun. Among the game parameters, *Area of level (G_A)* had a significant effect with $c(z) = -0.198$. Among the PBFs *Completion (C)* had a significant effect with $c(z) = 0.245$.

3.6.4 Preference Learning

Preference learning techniques are applied to explore to what extent PBFs and game parameters can be used to predict players' preferences. Here two approaches to modelling player preferences are investigated.

In the first approach, the large margin algorithm [60] was used on the dataset of 65 features and corresponding preference labels. This approach was originally developed for a driving route recommendation system [60]. This technique has been previously applied in similar studies of modelling PX in super Mario bros [167, 166] and racing games [216]. While this technique has not been applied to MAR games which require processing movement data from sensors for player behaviour features, it has been applied to model traditional digital games and is generalise-able across different genres of digital games. To increase the model accuracy of the large margin algorithm approach, feature selection techniques were applied. Feature selection is used to reduce the dimensionality of the feature space (which is discussed further below).

This problem of reducing the dimensionality of the feature space has motivated the use of another approach to preference learning that extracts player types from the PBFs. In this second approach, the PBFs are clustered using unsupervised learning techniques to extract player types from this behaviour data. This approach has been used in previous research to successfully extracted different types of players based on behavioural metrics [56, 57, 160]. The impact of the game parameters on each player type is first analyzed using similar statistical tests and correlation analysis as described above. This is done to understand how the emotional preference of each player type differs from the other. These player types are also used as input for the preference learning models. Details on each of the two approaches are described below.

Large Margin Algorithm Approach

This method aims to model features of interest through a linear combination of a weighted vector that maps preferences to features. This is given by $P(F) = FW^T$, where $P(F)$ is the subjects preference, F is the extracted feature vector and W is the weights to be optimized. Since the objective is to predict pairwise preferences, a function $P(F_x)$ is required where $P()$ is the preference function and F_x is a set of features that have been computed from a game X. $P(F_A) > P(F_B)$ if a subject has a preference for Game A over Game B. Here F_A and F_B are PBFs and game parameters extracted from each of the games A and B respectively. This inequality can be expressed through a linear combination $F_AW^T > F_BW^T$, which is further rewritten as $(F_A - F_B)W^T > 0$ or $F_DW^T > 0$, where F_D is the feature difference vector for the preference. The problem is thus reformulated as a linear classification of estimating W, where the input features are the feature difference between the original feature space of the 2 game pairs being compared.

Previous research explores this as a binary classification problem where $F_DW^T \in \{0, 1\}$: 0 is assigned to an instance where the subject has a preference for Game A over Game B and 1 for the opposite preference. This is accomplished by either filtering the data to remove instances with no preferences[60] or by forcing choice onto subjects via the 2-AFC protocol[216]. However, for this exploratory study, it would be beneficial to investigate both binary (via data filtering) and ternary classifications were $F_DW^T \in \{0, 1, 2\}$: 2 is the class assigned to instances where the subject had no preference (options 3 and 4 in the 4-AFC protocol). This approach is investigated since it matches the format of the data collected without filtering.

Since the objective is to optimize the weight vector W, which can be used to linearly combine the feature space F_D to predict preferences, the performance of two linear classifiers: logistic regression and linear discriminant analysis (LDA); and one non-linear classifier: support vector machines (SVM) for this classification problem is investigated. LDA has been used in a related study [216]. However, it is unclear from previous work if other models can outperform this approach. This study evaluates model performance using the sample accuracy and standard deviation from 10-fold cross-validation (CV).

Feature Selection Previous studies observe that model performance improves through feature selection techniques [167, 166, 236]. While there are a large number of approaches, this study uses sequential forward selection (SFS) and sequential floating forward selection (SFFS) in this study as they are often used in similar work. In [167], SFS and SFFS outperformed other techniques tested. SFS is a bottom-up search algorithm that tries to find the best performing feature set. It starts with the best performing single feature and adds new features from the remaining set such that model performance of the new set generates the best possible overall performance over other potential features for addition. SFFS is similar to SFS except that when a forward step is performed, the algorithm also checks if a feature from the existing set can be excluded to improve overall model performance.

The feature difference is calculated from each game pair and used in preference learning techniques. The extracted features are used to predict preferences. A summary of findings is shown in Table 3.8 which provides the accuracy and standard deviation from 10-fold CV of the Logistic Regression, LDA and SVM classifiers, to predict the various emotional dimensions of preferences in both binary and ternary classification scenarios. Corresponding accuracies of the best performing feature subset from the feature selection techniques (SFS, SFFS) along with the number of features in the subset have also been provided in Table 3.8. It has been a common observation across all dimensions and types of classifiers that the base (all 65 features) performance without feature selection performs poorly with very high standard deviation (as high as $\pm 28.55\%$ for the ternary *Fun* LDA classifier). However, all base classifiers perform higher than random chance (50% for binary and 33.34% for ternary).

Boredom The best binary classifier was SVM with a subset of 14 features found with SFFS - the performance was $86.33 \pm 10.2\%$. The best ternary classifier was SVM with a subset of 21 features found with SFS - the performance was $60.29 \pm 13.8\%$.

Challenge The best binary classifier was SVM with a subset of 5 features. Both SFS and SFFS found the same feature set - the performance was $93.36 \pm 4.4\%$. The best ternary classifier was SVM as well with the same subset of 5 features (found with both SFS and SFFS) - the performance was $84.85 \pm 5.5\%$.

Excitement The best binary classifier was LDA with a subset of 15 features found with SFFS - the performance was $75.61 \pm 12.5\%$. The best ternary classifier was SVM with a subset of 14 features found with SFFS - the performance was $56.92 \pm 10.7\%$.

Frustration The best binary classifier was SVM with a subset of 24 features found with SFFS - the performance was $93.89 \pm 6.1\%$. The best ternary classifier was SVM as well with a subset of 11 features found with SFFS - the performance was $75.43 \pm 8.3\%$.

Fun The best binary classifier was SVM with a subset of 16 features found with SFFS - the performance was $79.21 \pm 9.7\%$. The best ternary classifier was SVM with a subset of 10 features found with SFFS - the performance was $69.42 \pm 7.0\%$.

Table 3.8: Summary of results from the preference learning techniques of the large margin approach including feature selection for both binary and ternary scenarios. For each classifier, the number of features used, the sample accuracy and standard deviation from 10-fold CV are shown. The best performing binary and ternary classifier for each emotion has been highlighted.

			Log. Reg.			**LDA**			**SVM**		
			$F_{\#}$	Acc	$\pm SD$	$F_{\#}$	Acc	$\pm SD$	$F_{\#}$	Acc	$\pm SD$
Boredom	Binary	All	65	64.33%	15.8%	65	50.67%	28.6%	65	78.00%	12.9%
	$N_s = 59$	SFS	28	79.67%	9.9%	27	80.00%	12.5%	15	84.67%	9.1%
		SFFS	29	79.33%	15.3%	40	84.67%	11.8%	14	86.33%	10.2%
	Ternary	All	65	45.71%	17.0%	65	37.58%	12.9%	65	47.64%	8.0%
	$N_s = 117$	SFS	3	57.37%	14.7%	6	58.21%	13.0%	21	60.29%	13.8%
		SFFS	36	57.45%	16.3%	6	58.21%	13.0%	32	58.69%	10.9%
Challenge	Binary	All	65	81.00%	10.3%	65	74.54%	13.1%	65	81.10%	10.0%
	$N_s = 105$	SFS	15	91.54%	5.0%	4	91.45%	5.0%	5	93.36%	4.4%
		SFFS	3	91.45%	5.0%	4	91.45%	5.0%	5	93.36%	4.4%
	Ternary	All	65	67.70%	12.7%	65	62.80%	14.1%	65	72.86%	9.1%
	$N_s = 117$	SFS	2	82.20%	5.2%	12	82.26%	6.0%	5	84.85%	5.5%
		SFFS	2	82.20%	5.2%	12	82.26%	6.0%	5	84.85%	5.5%
Excitement	Binary	All	65	55.44%	15.8%	65	57.81%	20.5%	65	56.39%	8.2%
	$N_s = 89$	SFS	12	73.64%	13.6%	9	72.39%	13.7%	8	68.53%	9.6%
		SFFS	9	73.64%	14.5%	15	75.61%	12.5%	12	71.89%	10.5%
	Ternary	All	65	44.05%	12.9%	65	36.41%	11.7%	65	43.01%	6.9%
	$N_s = 117$	SFS	31	54.99%	14.1%	20	54.88%	11.4%	2	55.17%	14.7%
		SFFS	15	53.05%	13.3%	11	55.78%	11.2%	14	56.92%	10.7%
Frustration	Binary	All	65	85.38%	8.8%	65	68.29%	14.1%	65	82.60%	8.4%
	$N_s = 81$	SFS	1	90.41%	8.9%	1	90.41%	8.9%	4	92.78%	5.9%
		SFFS	1	90.41%	8.9%	34	92.92%	9.5%	24	93.89%	6.1%
	Ternary	All	65	50.57%	14.4%	65	45.57%	20.4%	65	59.64%	14.6%
	$N_s = 117$	SFS	11	71.83%	10.6%	16	72.65%	11.6%	16	72.92%	8.3%
		SFFS	13	73.57%	9.8%	20	74.05%	10.0%	11	75.43%	8.3%
Fun	Binary	All	65	65.62%	13.9%	65	58.47%	11.9%	65	60.37%	5.2%
	$N_s = 106$	SFS	42	73.68%	8.9%	35	74.33%	6.5%	16	74.68%	6.8%
		SFFS	32	76.26%	8.6%	34	77.59%	11.4%	16	79.21%	9.7%
	Ternary	All	65	59.56%	9.2%	65	47.53%	11.7%	65	54.77%	5.0%
	$N_s = 117$	SFS	4	65.75%	7.8%	1	65.67%	8.1%	9	67.68%	6.8%
		SFFS	32	65.94%	8.4%	18	68.99%	9.0%	10	69.42%	7.0%

Player Types Approach

This approach aims to reduce the dimensionality of the feature space by clustering player behaviour to smaller clusters. Previous studies have clustered player behaviour to extract player types from this data [56, 57, 160]. These player types can either be used as features for predictive modelling (as is the case in this study) or to gain insight into the behavioural patterns exhibited by players in a game.

In this study, the k-means clustering technique is used to extract player types from the data as it has been used previously to infer player types from behavioural data [160]. The optimal number of clusters is selected using the elbow method, i.e. plotting the within-cluster sum of squares and choosing the number of clusters from the graph [105].

Two types of input features for the k-means model are tested:

1. *Large Margin Features* whose feature vector has a dimensionality of 63.
2. *Game Pair Features* whose feature vector has a dimensionality of 126 (63 $\times$ 2).

In the first approach, the feature difference vector (which is calculated using the large margin algorithm described earlier in the section) is used as input. The dimension of the input feature vector of the large margin approach is 63. Second, the extracted features from each game pair are used without calculating the feature difference vector. This leads to an input feature vector of 126 dimensions (63×2).

Using behavioural features to infer player types using clustering in an accepted method (similar to the later approach), it is unclear to what extent player types can be inferred from large margin features (the former approach). Using large margin features reduces the dimensionality by half (only 63 features as opposed to 63×2), it is unclear to what extent this is a valid approach to extract player types from behavioural data. Since these features are descriptors of differences in behaviour between two games as opposed to descriptors of behaviour within a game.

Since this study is motivated by modelling player's emotional experiences, statistical tests are conducted within in each cluster to test the effects of each of the two game parameters on the emotional preference of each player type. These tests were conducted on the clusters that emerged from the two types of input feature vectors tested. It is interesting to note that the two types of input feature vectors result in a similar number of clusters and a similar number of data points within each cluster.

Finally, the player types extracted from the feature difference vector (using the large margin algorithm) along with the 2 game parameters is used as input for the three classifiers to predict the emotional preference. As described earlier in this section, preference learning is approached as a classification problem and both binary and ternary classification models are tested in this analysis.

The optimal number of clusters for both input feature vectors was found to be three. Figure 3.3 shows the plot of the within-cluster sum of squares (WCSS) in each case the elbow is observed at three clusters[105]. The chi-square test was used to check for significant effects of the game parameters on player preference within each cluster in both cases. Correlation coefficients of statistically significant (p-value $< 1\%$) game parameters and corresponding emotional preference are reported below.

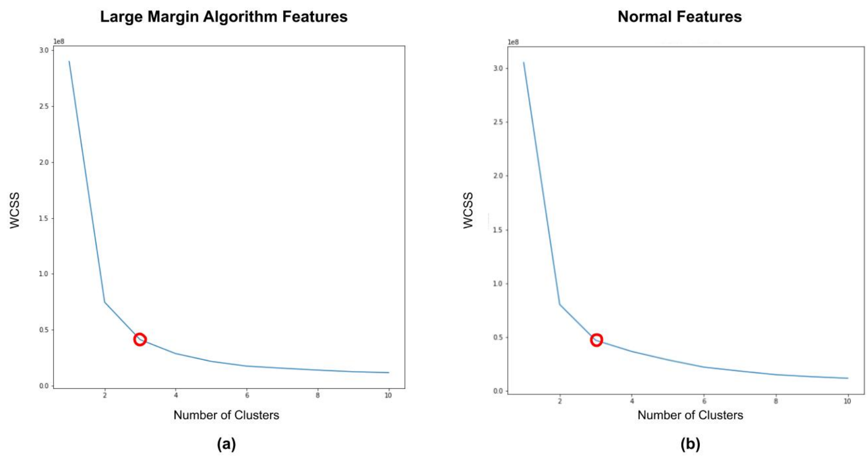


Figure 3.3: Figure a-b shows the Elbow method across both the tested input feature vectors. The within-cluster sum of square (WCSS) is plotted along the y-axis and the number of clusters is plotted along the x-axis. The point of the elbow in both graphs is highlighted in red. [a]: Shows the plot for the large margin algorithm features vector. [b]: Shows the plot for the normal feature vector.

Large Margin Algorithm features The three clusters contained: 87, 20 and 10 data points or game pairs respectively. Table 3.9 shows a summary of the statistical analysis conducted to understand the impact of game parameters on emotion preference within each player type (or cluster). The results show that game area impacts the emotion preference in a similar way for all three clusters. Larger game areas resulted in more frustration and higher challenge for all three clusters. Additionally, a larger game area resulted in increased boredom for players in cluster 1.

Table 3.9: Statistical analysis of game parameters on emotion preferences for Large Margin Algorithm based player types

Cluster	No. of Data Points (game pairs)	Game Parameter	Description of statistical effect
Cluster 1	87	Area of level (G_A)	Participants felt more **Frustrated** when area was larger, $c(z) = 0.69$
		Area of level (G_A)	Participants felt more **Bored** when area was larger, $c(z) = 0.34$
		Area of level (G_A)	Participants felt more **Challenged** when area was larger, $c(z) = 0.68$
Cluster 2	20	Area of level (G_A)	Participants felt more **Frustrated** when area was larger, $c(z) = 0.69$
		Area of level (G_A)	Participants felt more **Challenged** when area was larger, $c(z) = 0.68$
Cluster 3	10	Area of level (G_A)	Participants felt more **Frustrated** when area was larger, $c(z) = 0.69$
		Area of level (G_A)	Participants felt more **Challenged** when area was larger, $c(z) = 0.68$

Game Pair Features The three clusters contained: 90, 19 and 8 data points or game pairs respectively. Table 3.10 shows a summary of the statistical analysis conducted to understand the impact of game parameters on emotion preference within each player type (or cluster). The results show that game area impacts emotion preference in a similar way for the first two clusters. Larger game areas resulted in more frustration and higher challenge for both. Additionally, a larger game area resulted in increased boredom for players in cluster 1. Finally, a high amount of game treasure resulted in decreased boredom for players in clusters 1. There was no significant impact of game parameters on emotion preference for players in the third cluster.

Table 3.10: Statistical analysis of game parameters on emotion preferences for Game Pair based player types

Cluster	No. of Data Points (game pairs)	Game Parameter	Description of statistical effect
Cluster 1	90	Area of level (G_A)	Participants felt more **Frustrated** when area was larger, $c(z) = 0.69$
		Area of level (G_A)	Participants felt more **Bored** when area was larger, $c(z) = 0.34$
		Area of level (G_A)	Participants felt more **Challenged** when area was larger, $c(z) = 0.68$
		Treasure in Level (G_T)	Participants felt less **Bored** when treasure in level was low, $c(z) = -0.32$
Cluster 2	19	Area of level (G_A)	Participants felt more **Frustrated** when area was larger, $c(z) = 0.69$
		Area of level (G_A)	Participants felt more **Challenged** when area was larger, $c(z) = 0.68$
Cluster 3	8	No significant effects	

It is observed that the two types of inputs feature vectors for clustering has resulted in similar clusters emerging. Players within each cluster do not show different emotional preferences from each other. The player types extracted from clustering the large margin algorithm features along with the 2 game parameters is used as input to predict player preferences. The feature difference from the two games pairs is computed using the large margin algorithm resulting in 63

PBFs and 2 game parameters. The PBFs are clustered using the k-means algorithm into 3 clusters. This cluster output is a categorical ternary variable which was then one-hot encoded and one of the columns was dropped to avoid the dummy variable trap resulting in 2 PBFs. These 2 PBFs along with the 2 game parameters is used as input for the classification models. Table 3.11 shows the summary of the sample accuracy and standard deviation from 10-fold CV of the logistic regression, LDA and SVM classifiers in both the binary and ternary classification scenarios. The best performing binary and ternary classifiers are reported below for each of the dimensions of emotional preference.

Boredom The best binary classifier was Logistic Regression, the performance was $82.67 \pm 13.9\%$. The best performing ternary classifier was the SVM, the performance was $61.04 \pm 14.2\%$.

Challenge The best binary classifier was the LDA, the performance was $83.90 \pm 4.0\%$. The best performing ternary classifiers were both the logistic regression and LDA, with the performance was $71.26 \pm 6.2\%$.

Excitement The best binary classifier was the SVM, the performance was $63.05 \pm 16.7\%$. The best performing ternary classifiers was also the SVM, the performance was $45.34 \pm 13.8\%$.

Frustration The best binary classifier was Logistic Regression, the performance was $85.23 \pm 7.2\%$. The best performing ternary classifier was also logistic regression, the performance was $62.63 \pm 6.1\%$.

Fun The best binary classifier was Logistic Regression, the performance was $66.01 \pm 10.1\%$. The best performing ternary classifier was also logistic regression, the performance was $59.92 \pm 9.5\%$.

3.6.5 Feature Recommendations

Since a large feature space is explored in the large margin approach, an analysis is conducted to identify sets of movement features that are important in modelling a player's emotional preference. The recommendations are based on a grounded analysis of features in terms of statistical effects and the likelihood of each being selected in the best performing feature subset. This list of feature combinations can be used as a starting point in similar work to predict player preferences.

Two common features are observed in predicting all the emotion dimensions, given by: $\Theta \in \{A_{min}, RV_{min}\}$. A number of feature sets were also

Table 3.11: Summary of results from the preference learning techniques of the player types approach including both binary and ternary scenarios. For each classifier, the sample accuracy and standard deviation from 10-fold CV are shown. The best performing binary and ternary classifier for each emotion has been highlighted.

		Log. Reg.		**LDA**		**SVM**	
		Acc	$\pm SD$	Acc	$\pm SD$	Acc	$\pm SD$
Boredom	Binary $N_s = 59$	82.67%	13.9%	81.00%	14.7%	76.00%	14.0%
	Ternary $N_s = 117$	54.30%	13.8%	54.30%	13.8%	61.04%	14.2%
Challenge	Binary $N_s = 105$	83.00%	5.3%	83.90%	4.0%	76.27%	4.2%
	Ternary $N_s = 117$	71.26%	6.2%	71.26%	6.2%	68.59%	5.9%
Excitement	Binary $N_s = 89$	58.47%	13.7%	57.36%	11.5%	63.05%	16.7%
	Ternary $N_s = 117$	41.17%	9.2%	41.17%	9.2%	45.34%	13.8%
Frustration	Binary $N_s = 81$	85.23%	7.2%	85.23%	7.2%	80.37%	5.6%
	Ternary $N_s = 117$	62.63%	6.1%	61.72%	5.1%	61.02%	7.5%
Fun	Binary $N_s = 106$	66.01%	10.1%	64.19%	10.3%	62.19%	12.6%
	Ternary $N_s = 117$	59.92%	9.5%	58.25%	9.7%	56.43%	11.6%

found that could predict 4(/5) dimensions. A set of 4 features can be used to predict *Boredom, Challenge, Excitement and Frustration* (not *Fun*), given by: $\Lambda \in \{A_m, A_{std}, A_{max}, V_m\}$. A single feature can be used to predict *Boredom, Challenge, Frustration, Fun* (not *Excitement*), given by: $\Pi \in \{G_A\}$. Similarly, feature sets emerge that can predict 3(/5) dimensions. A single feature was found to be able to predict *Challenge, Frustration, Fun*, given by $\Phi \in \{C\}$. A single feature can be used to predict *Boredom, Challenge, Frustration*, give by: $\Psi \in \{J_m\}$. Common feature sets also emerge that can predict pairs of emotions. A large set of features could predict *Challenge and Frustration*, given by: $\Omega \in \{T, A_D, J_{std}, RA_{tMin}, RJ_{tMin}, RV_{max}\}$. Another pair of features could predict *Boredom, Excitement*, given by: $\Delta \in \{G_T, RJ_{tD}\}$. Feature recommendations for each dimension of preference are given by the composition of the sets of features presented above, illustrated in fig 3.4. For instance, *Challenge* is a set of 15 features, given by: $\mathbb{CH} \in \{\Theta, \Lambda, \Pi, \Psi, \Phi, \Omega\}$. Table 3.12 presents summary of the feature sets presented in this section as well as the dimensions of emotion preference that the feature set corresponds to. Showing the recommendations as compositions of other feature sets allows us to appreciate important relationships across emotions. For instance, a total overlap in the features that predict *Challenge* and *Frustration* is observed (implications are discussed in section 3.7).

Table 3.12: Summary of feature sets and corresponding dimension of emotion preference

Set Name	Feature Components	Related Emotion Preference Dimensions
Θ	Minimum Acceleration (A_{min}) Minimum Angular Velocity (RV_{min})	Fun, Excitement, Boredom, Challenge and Frustration
Λ	Mean Acceleration (A_m) Standard Deviation of Acceleration (A_{std}) Maximum Acceleration (A_{max}) Mean Velocity (V_m)	Excitement, Boredom, Challenge and Frustration
Π	Game Area (G_A)	Fun, Boredom, Challenge and Frustration
Φ	Game Completion (C)	Fun, Challenge and Frustration
Ψ	Mean Jitter (J_m)	Boredom, Challenge and Frustration
Ω	Game Time (T) Difference between max and min Acceleration (A_D) Standard Deviation of Jitter (J_{std}) Time of minimum Angular Acceleration (RA_{tMin}) Time of minimum Angular Jitter (RJ_{tMin}) Maximum Angular Velocity (RV_{max})	Challenge and Frustration
Δ	Game Treasure (G_T) Difference between time of max and min angular jitter (RJ_{tD})	Boredom and Excitement

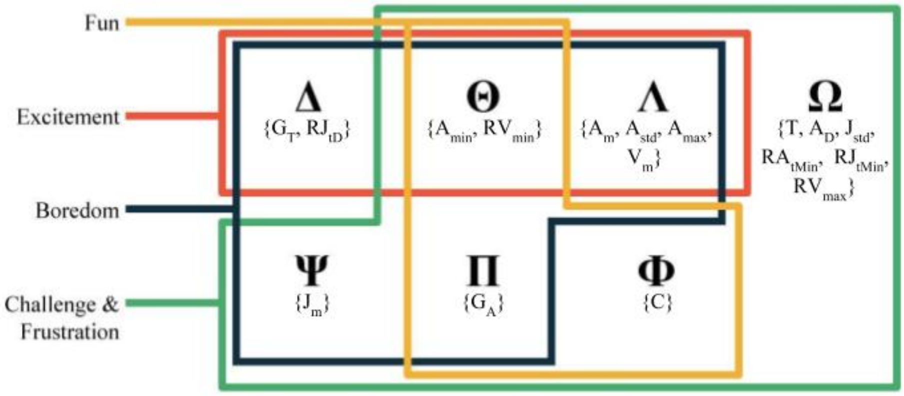


Figure 3.4: The figure shows how recommended feature sets for the various dimensions of emotional preference can be expressed as compositions of important feature sets.

3.7 Discussion

The results indicate that combinations of game parameters and PBFs can be used to accurately predict dimensions of emotional preferences. The performance of three models for classification have been compared in this study and it is shown that SVM classifiers (unexplored in previous research for this problem) were the best performing classifiers for both accuracy and stability, indicated by higher accuracy and lower standard deviation of 10-fold CV accuracies. Although all classifiers perform better than random chance, it is observed that binary classifiers outperform ternary classifiers in both accuracy and stability. This is not surprising since the binary classifiers attempt to model a simpler problem.

It is observed that the player types approach to preference learning did not perform as well as the large margin approach. A possible explanation for this is the limited dataset of 117 game pairs. When clustering was applied to this dataset, three clusters emerged. The game pairs were not evenly distributed among the three clusters. Most game pairs ($\approx 76.9\%$) were assigned to the first cluster in both variants of the clustering approach tested. If a larger dataset with game pairs that are equally distributed between the emergent clusters is used, this approach could result in better performance in predicting a player's emotion preference. It is worth noting that the accuracy and stability of classifiers from the player types approach are higher than the base performance (using all the features) in the large margin algorithm approach. However, once feature selection is applied to the latter, the accuracy and stability of all classifiers across all emotional dimensions are higher than in the player types approach.

In the large margin approach, *Boredom, Excitement* and *Fun* are difficult to model. This is observed in statistical analysis as well as the low accuracy and stability of the classification models tested. Predicting *Challenge* and *Frustration* show higher accuracy and stability. Due to the small data set, it is possible that classifiers for *Boredom, Excitement* and *Fun* show over-fitting due to a high standard deviation of 10-fold CV accuracies. The performance of *Challenge* and *Frustration* shows more acceptable stability with both binary and ternary classifiers showing a variance of $\approx 5\%$. This variance will be further reduced if a larger data-set is used. These results are similar to results from other studies that model players' preferences [167]. Although the authors investigate Super Mario in their work, they find that *Fun* and *Boredom* were the most difficult to predict, while *Challenge* and *Frustration* were the best performing classifiers.

It is possible that these observations are due to the underlying relationship between game activities and the specific emotional dimension. Some theories[17] consider emotions as being constructed from more fundamental properties called Valence and Arousal: Valence is the amount of goodness or badness in experiences, while Arousal is the psychological state of being awake. *Fun, Excitement,* and *Boredom* are more resonant with the emotional dimensions of valence. *Frustration* and *Challenge* are resonant with the emotional dimension of arousal. *Frustration* is a construct of negative valence and high arousal, while *Challenge* is strongly linked to player performance and high arousal states. As this approach uses movement data, this information medium could be more useful to detect variable arousal rather than variable valence based emotional states. Although it is useful to use valence and arousal to interpret these results, the current approach of asking players about preferences across easily understandable emotions has obvious advantages as it is more intuitive for people to compare 2 experiences based on *Fun* or *Frustration* rather than valence and arousal. Studies about emotions[17] also tell us that people are different in their ability 'to represent their experiences as categorically distinct events' and this ability is influenced by context and language abilities. This means that peoples interpretations of the different emotion dimensions used in this study are subjective and are influenced by their cultural and linguistic abilities. This is observed in this study by the variable proportions of clear preference across the emotions tested. In order to understand the relationship between player movement and the emotional dimensions of valence and arousal, a revised study design is required.

The current game requires considerably high amounts of walking; these results are applicable to similar AR games that require movement in local space. The techniques proposed in this chapter have considerable potential to create content that is optimized for an ideal balance of *Challenge* and *Frustration* (i.e. the best balance of most challenging and least frustrating). Currently, a simple game design space has been explored and it is possible that *Frustration* and *Challenge* is being predicted by the same underlying feature correlations. In this game, most participants appear to prefer a large amount of Treasure (more fun, exciting and less boring) and do not prefer walking a large amount (more challenging and frustrating). In this case, it would be impossible to find a game experience that is both challenging but not frustrating. It would be interesting to see how this approach scales in more complex game design spaces. This can easily be achieved by using a game parameter set of higher complexity, for example, $G_A \in \{XS, S, M, L, XL\}$. This set contains values for the game area parameter that allows for a more diverse range of levels moving from small to large.

This study serves as a starting point in a better understanding of how player behaviour in AR environments can be used to model their preferences. An aspect of this problem that is unresolved is guidelines of following a binary or ternary approach to classification. Binary classification performs better and is more stable. However, this is unsurprising since it models a simpler problem. This advantage over ternary classification seems preferable and could be accomplished by forcing a binary choice onto participants referred to as the 2-AFC[216]. A critique of this approach is that it seems a naive way of achieving better performance and could prove detrimental to optimization for true player experience.

This work is built on studies of detecting emotions from movement and has been applied to predict emotional preferences in MAR games, a complex problem that this work has begun probing. Future research in addressing the discussed gap in establishing ground truth improved feature extraction, and different models to address this problem (other non-linear classifiers or models for time-series data) would increase our understanding. The subsection below discusses some limitations of this study.

3.7.1 Limitations

One of the main limitations of the current study is the limited amount of data. Since large training sets are usually required to train machine learning models, the accuracies of the preference learning models observed in this study are potentially limited due to the size of the dataset.

As mentioned above, it is possible that the high performance of the *challenge* and *frustration* classifiers are due to the same underlying feature correlations. One possible reason for this is the relatively simple game design space (four level variants) being explored in this study. If this limitation persists when a more complex space is explored, it would be impossible to create content that is challenging but not frustrating for the player. This limitation must be further investigated in future work by checking the hypothesis that preference learning models based on movement data cannot differentiate between *frustration* and *challenge* and that modelling preference across the dimensions of valence and arousal can better differentiate between frustration and challenge. Additionally, this limitation is further validated by participant comments at the end of the study. Most participants tended to feel that a frustrating experience for them meant that the game level was challenging. This conceptual overlap between *challenge* and *frustration* shows the need to explore alternative measures of player experience.

Another limitation was observed from a number of post-experiment interviews with participants. Participants felt that the ease with which they completed the 2D puzzle, as well as the selection of the puzzle image (which was selected from a database of 100 images at random), biased their reported emotional preferences. Since images were randomly selected, the study did not control for the emotional response created by a specific puzzle image. Figure 3.5 shows examples of the image database that were used for the puzzle mechanic. Since the 2D puzzle is at the end of the game round after the AR exploration, PBFs or game parameters from this part of the game have not been used. This would have added noise to the preference data. This limitation will be addressed in the study described in the next chapter by using a new game.

Another limitation is that it is difficult to generalize the finding of this study to all genres of AR games. The AR Treasure-hunt described here is a valid testbed for this pilot study, however, future work exploring this relationship in other AR games will need to be conducted. The results observed in this study while not generalisable to all mobile AR experiences are still applicable to exploration-based AR experiences such as *dARk* which is a story-driven horror game.

Post-experiment interviews with participants showed the current game experience got predictable by the end of the study (there were always either 9 or 16 treasure pieces in each level). This is due to the study design that explores a simple game design space that utilizes only 2 settings for game treasure a high setting (16 pieces) and a low one (9 pieces). This predictability is a limitation as it reduces the curiosity and fantasy aspects of the experiences which are both important dimensions in Malone's theory of fun in games [127]. Due to this, it is possible that precipitants approached the final session of the study less like a game and more like an experiment task which would've further biased the preference data.

Finally, due to the exploratory nature of this study, only a short unstructured interview was conducted as part of debriefing the participant. The interview did yield useful findings on the validity of the game and the subjective mea-

Figure 3.5: The figure shows a selection of 9 out the 100 images used for the puzzle mechanic for the AR Treasure Hunt game.

sures used. However, this approach to gathering qualitative data is insufficient in understanding how different the game parameters could affect a player's experience across the different levels. Future work will address this limitation by conducting player interviews with participants.

3.8 Chapter Summary

The study reported in this chapter has analyzed player movement in an exploration-based AR mobile game. In order to do this, an AR Treasure hunt game was designed and developed. This game was used in a user study where participants self-reported their emotional preference about variants of the game. Statistical analysis and predictive modelling showed that a combination of game context information and player movement can be used to accurately predict a player's preferences regarding *Challenge* and *Frustration* with a high degree of accuracy while all other emotions tested performed above random chance (*Fun*, *Excitement* and *Boredom*). Based on these observed results, it was proposed that these techniques are better suited to measure variable arousal states rather than differences in valence. A number of limitations of the existing study were also discussed regarding the interactions used in the game as well as the approach of analysing game experience as emotional preferences.

This has motivated the study proposed in the next chapter. The next study builds on the findings from this study by using a different game that focuses more on physical exertion as opposed to open-ended exploration. The next study will also incorporate the collection of self-reported valence and arousal data in order to understand its relationship with player movement. Additionally, most traditional questionnaires to measuring PX in games will also be used. Using a new game in the next user study will overcome a number of limitations observed from this study such as predictability of the game task. This would also show to what extent the results reported in this study can be generalized to AR mobile games.

Chapter 4

Modelling Game Experience in an AR Exergame

The previous chapter presented the first study of this research where supervised learning techniques were used to predict a player's preference across a number of emotion dimensions. The study reported in this chapter aims to overcome two main limitations observed in the previous study specifically the game and the questionnaire used.

First, the game used both movement-based interactions in AR and touchscreen-based interactions both of which contributed to the overall experience of the game. Since the focus of this research is on movement-based AR interactions, the puzzle-based touch screen task added undesirable noise to the data collected.

Second, it was discussed that measuring player experience as individual dimensions of emotions lead to conceptual overlaps between *challenge* and *frustration.* This was seen in comments from participants and the feature selection algorithms converging to similar feature sets while trying to predict each.

Finally, participants found the game predictable after the first 2 rounds (4 game levels) which lead to boredom in the game. For this reason, this study focuses on applications of mobile AR exergames. Exergames traditionally combine physical exercise and digital games. Academic investigation into exergames has shown its benefits for physical health, cognitive and emotional well-being [204, 9]. AR has be been investigated as a novel platform for these games [119, 112]. However, there is a notable gap in studies that investigate how different game parameters can affect PX in these AR exergames.

To address this gap, this study first introduces a new game titled *Running*

Chickens which has been designed based on existing examples of AR exergames [119, 112]. This is different from the *AR treasure hunt* game in the previous study since game mechanics are used to motivate players to increase their physical exertion during gameplay. These game mechanics refer to introducing a time constraint for each level (which adds a sense of time pressure for players) and having game collectables evade players (which adds a sense of challenge that can only be overcome by increasing the amount of physical exertion by running after the object to collect it).

This chapter first presents the aims and motivations for this study in section 4.1. The AR exertion game: *Running Chickens* developed to investigate player modelling is described in section 4.2. Section 4.3 describes the experiment protocol used in this study. The study design has been informed using pilot studies which have been described in section 4.4. Section 4.5 provides details about the data collected during the study. The data analysis method and results from the analysis are presented in section 4.6. Section 4.7 discusses the implications of these results and the limitations of the study. Finally, section 4.8 presents a chapter conclusion.

4.1 Aims and Motivations

To further investigate player modelling in AR games, this study uses a different game in order to overcome limitations of the previous game task namely the influence of non-AR game interactions on player experience. While the previous game and study serve as a useful exploratory study, the game itself cannot generalise to other AR mobile games. Since AR mobile games is a broad genre, the remaining studies conducted in this research will focus on AR mobile exergames.

The literature on exergames has shown benefits to both youth and older adults. It is considered a potential solution to childhood obesity[112]; providing physical, social and cognitive benefits when incorporated in physical education curriculum[204]. Gamifying physical activity for older adults has shown to have higher motivation, enjoyment, and engagement as compared to regular physical activity[100].

There exists a large number of commercial exergames that use physical sensors to translate body movement into digital gameplay interactions from game systems just as the *Switch* and *Wii* from *Nintendo*, the *Kinect* platform from *Microsoft* which has several rhythm-based and fitness games. These games are played indoors usually on a home game console. Mobile exergames usually use GPS to increase the player's physical activity by walking to different locations, using GPS[119, 112, 9]. A popular mobile AR exergame is *Run Zombies*, which is an audio-based game for runners. While not traditionally an exergame, another popular AR mobile game is *Pokemon GO* from *Niantic labs* has been shown to have positive health benefits due to its mass adoptions and location-based game design[9].

AR mobile exergames have also been developed by researchers to explore the potential benefits of these types of games. It is a generally accepted premise that movement-based games have a positive health impact by increasing the physical activity of the player. There is a notable gap in studies that evaluate the impact of exergame parameters on PX.

The main aim of this study is to explore the impact of different game parameters of an AR exergame on a player's experience. Additionally, exploring techniques to predict PX within these games have the potential of creating fitness applications that can be optimised for ideal PX. Which should in turn increase a player's engagement with them leading to increased health benefits. While the previous chapter explored player preferences across a number of emotion dimensions, this study uses standardised rating-based measures of PX namely the Affective Slider [22] and the GEQ [86] as subjective measures of PX. These measures have been developed for use in HCI based on theories of PX and are commonly used to evaluate the PX facilitated by a game system.

The two research questions that are explored in this study are:

- What is the impact of AR exergame parameters on PX?
- To what extent can movement-based, performance and game features be used to predict PX?

In order to explore these research questions, an AR Exergame is developed (details about the game is provided in the next section) based on existing examples of such games proposed by other researchers. The game is intended to use player movement within a local space to accomplish game objectives. Different game parameters are used to influence the amount of player movement within the space. For instance, each level changes the size of the physical area which the player must traverse through as well as the amount of physical exertion (walking vs jogging) that needs to be employed in order to accomplish game objectives. The game design has been inspired from existing examples of such games from research namely *GioBoids*[119] and *Calory Battle AR*[112].

In *Gioboids* players are directed to different locations to capture flocks of creatures. While in *Calory Battle AR* players must travel to different areas and diffuse bombs. These games have two game parameters in common: 1) the size of the game space in which the different game elements are populated 2) the number of game elements within the space. While both games involve an evaluation study of the games' enjoyment, empirical work that evaluates the effects of different game parameter settings (eg: large vs small spaces) does not exist. While both game parameters control the extent to which players move within the game, it is difficult to understand how different game settings will impact PX. This would be beneficial to game designers of similar games in creating the progression system of these games since the goal is to keep players engaged while gradually increasing their physical activity. The AR exergame developed for this study uses similar game parameters to drive player movement within the game. The next section describes the design and development of *Running Chickens* (the AR exergame designed as part of this research).

4.2 AR Exergame: *Running Chickens*

Running Chickens is the title of the AR Exergame designed and developed for the purposes of the research. The objective of Running Chickens is to catch digital chickens within a physical space. In each game, players are given a set amount of time and are presented with a flock of digital chickens within their local space. Similar to the previously mentioned examples, the size of the game level and the number of chickens in each level are controllable game parameters that are varied between levels.

The player must catch as many chickens as possible before the time runs out. The player can capture a chicken by colliding the mobile device with the chicken. Unlike the game used in the previous chapter (*AR Treasure Hunt*), the targets are visible. The player can see the chickens and must move towards them.

A novel game parameter introduced in this game is that chickens are programmed to evade the player. This evasion mechanic is set up to facilitate different challenge levels with the game. Chickens can show high evasion in difficult levels of the game and low evasion in easy levels of the game. This evasion mechanic is implemented by applying an acceleration to the chicken along a vector moving away from the player (which is given by the negative unit vector of the difference between the creature position and the player position). This mechanic is triggered when the player is within 0.5m of the creature.

In order to ensure that the chickens do not constantly evade players (making the game unplayable), the evasion mechanic is implemented with properties of stamina and cool down time. This stamina property allows the chicken to evade the player for a fixed amount of time after which it must rest for the cooldown time. In this version of the game, both stamina and cool down times are set to 3 seconds each. This means that once triggered to evade a player, a chicken can do so for 3 seconds after which it must rest for 3 seconds before it can evade the player again. This design decision prevents the evasion mechanic from being constantly triggered, giving the player a window of opportunity to capture the chickens when they are resting (in cooldown time). Figure 4.1 shows an illustration of this evasion mechanic for each chicken.

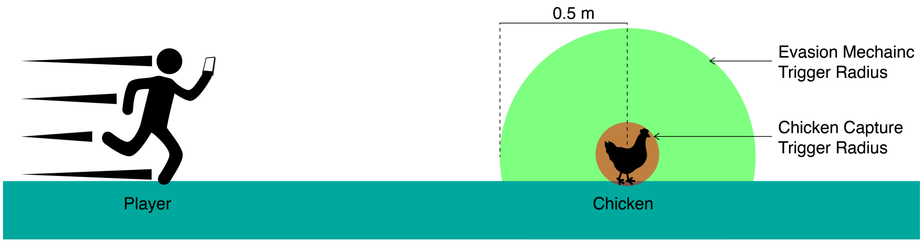


Figure 4.1: The figure illustrates how the evasion mechanic is triggered. If the player moves the mobile device into the radius of the evasion mechanic, it will be triggered causing the chicken to run away from them. If the player moves the mobile into the capture radius, the chicken will be captured adding a point to their game score.

The evasion mechanic is parameterized through the magnitude of the acceleration that is applied to the chicken when the mechanic is trigger. A high acceleration will result in the creature moving away from the player at a very high speed (making it difficult to capture) while a low acceleration will move the creature away from the player at a slow speed (making the creature easy to capture). In the version of the game used in this research, the evasion mechanic is set to two simple states where either the chicken runs away from the player or it does not move. The numeric value of this mechanic was tuned with pilot testing. The goal was to make the evasion of the chickens challenging on the player but still keeping it playable.

Each level of the game is played within a fixed square area. Similar to the previous game, players will be able to see the walls of the game level and the flocks of chickens will be bounded within this space. The area of the game is also parameterized to two discreet states: small levels of 15×15m and large levels of 30×30m. Finally, each game level will contain flocks of chickens in varying numbers. The amount of chickens is the final parameter that has been selected for this game. The flocks range from small (5 chickens) to large (20 chickens).

The game has been set up in such a way that the three parameters mentioned above will be used to generate each level of the game. At the beginning of the game, participants will scan the world and place the game level within the world. Once the level is placed, they can start the game. In each level, the player will be given a minute in each level to capture all chickens (or as many as they can) before the time runs out. A scoring system is set up in the game in such a way that a player gets a point for each chicken that is captured. The round ends when the player runs out of time or has captured all the chickens in the level. If the player captures all chickens before their time runs out, a point will be added for each second that is remaining in the time limit in the form of a time bonus. This further encourages the player to collect chickens as soon as possible. Figure 4.2 shows the screenshots of the main game and provides details about the UI components used in the main game screens. Figure 4.3 shows the screen flow of a single level. Similar to the previous game, it has been designed to be played in outdoor parks. Figure 4.4 shows a player testing the Running Chickens game.

The game has been developed in the Unity game engine [211] using the ARcore SDK [1]. The chicken model used in the game was downloaded from the unity asset store (fig 4.5 shows the chicken model), sound effects for the game was downloaded from `freesound.org`.

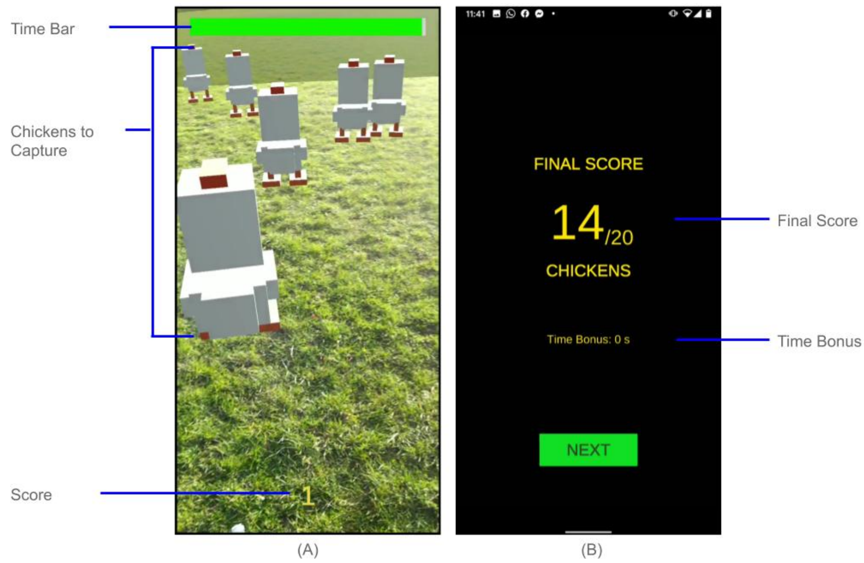


Figure 4.2: The figure [a-b] shows the game screens of running chickens. [a]: shows an in-game screen with the chickens that must be captured along with UI elements showing the time bar in each level as well as the player score. The Time bar indicates the amount of time left to collect chickens. The score indicates the points collected by the player (1 for each chicken collected). [b]: shows the screen at the end of each level, with the final score (number of chickens collected) as well as the additional time bonus received by the player.

4.3 Experiment Design

This section describes the experiment design for the user study described in this chapter. The study follows a within-participant design where a participant faces all conditions of the study. Participants were not given any constraints on how

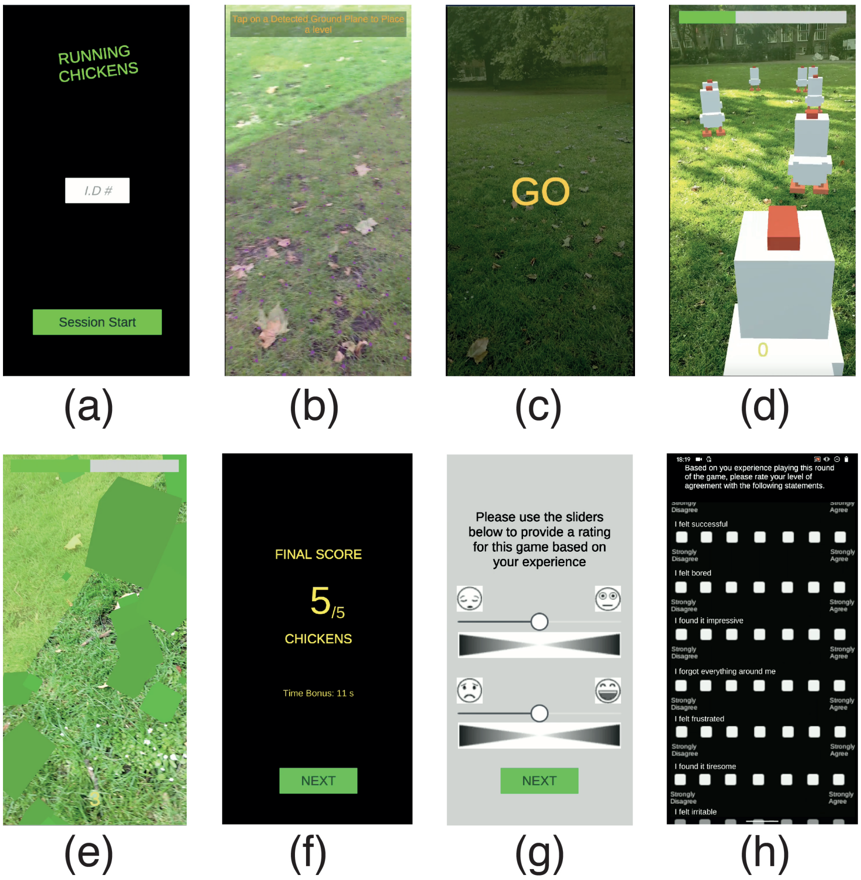


Figure 4.3: The figures [a-h] shows the game screens of running chickens for a single game level. [a]: when the game application is launched players (in the user study) are required to provide an ID that is provided to them. [b] Once a level begins the player will need to place the game level in the physical world. The text at the top of the screen reads, "Tap on the detected ground plane to place a level". The ground plane is seen in purple along the ground. [c] Once the user places the level, they are presented with a count down ("3.. 2.. 1.. GO!") to the beginning of the level. [d] Once the level begins the player sees chickens populate the game level (which they are required to collect). [e] Shows particle effects that are triggered when a player captures a chicken. [f] Shows the game over screen at the end of the level. [g] Shows a questionnaire that the player must complete at the end of each level. This screen shows the affective slider [22] [h] Shows the in-game GEQ being presented to the user which is another questionnaire the player must complete at the end of the level.

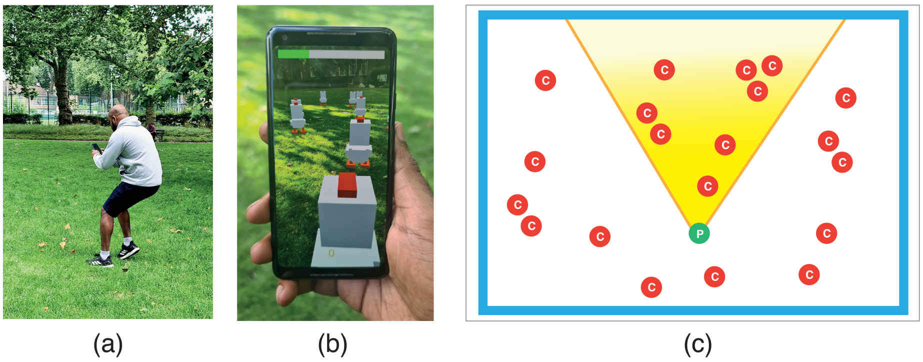


Figure 4.4: The figures [a-c] shows a player using the Running Chickens game. [a] Shows the player capturing a chicken. [b] Shows the view of the mobile device from the player's hand. [c] Shows an illustration of the game world as seen from the player's hand. The player is shown using the green circle while chickens are shown using red circles. The yellow area shows the device viewport and the boundary of the game level is shown using the blue line.

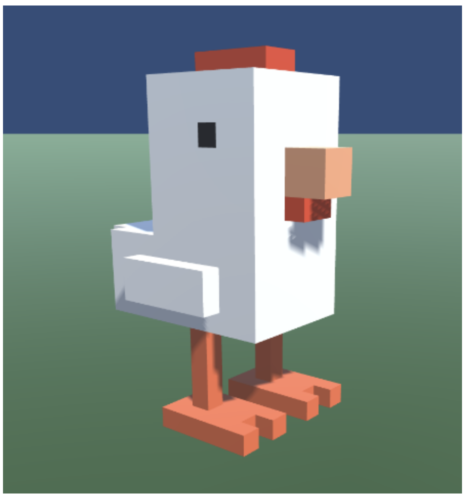


Figure 4.5: The figure chicken model downloaded from the unity asset store that is used in the Running Chickens game.

to hold that phone (in landscape vs portrait orientation), however, the screen design was made to play in portrait orientation.

Here, a condition refers to a specific game level. Three controllable game parameters are used to generate each level of the game:

- *The Area of the Level (G_{Area}):* 2 sizes of levels are compared. *Large Area (Lrg)* levels are $30m \times 30m$ and *Small Area (Sml)* levels are $15m \times 15m$.
- *Number of Chickens (G_{No}):* 2 amounts of chicken in each level are compared: *Low Number* with 5 and *High Number* with 20 pieces respectively.
- *Chicken Movement (G_{Mov}):* 2 setting of the chicken evasion mechanic are selected for evaluation: *Non-Evasive* condition where the chickens do not evade players, *Evasive* condition where the chickens evade players. The speed of chickens and the length of the cooldown time of this mechanic in the evasive condition has been selected through pilot testing which is described in the next section. This is done in order to ensure that the chickens can be caught by players.

Using 3 game parameters to generate levels has resulted in 8 ($2 \times 2 \times 2$) levels of the game. In this study, participants play all 8 levels of the game. In order to minimise ordering effects from biasing the data, the order in which the participants play each level of the game was randomised. At the end of each level of the game, participants were asked to fill in player experience measures relating to that specific level. This study used two measures:

- The Affective slider [22]: This is a tool that has been developed by Betella et al. for measuring affect in interactive interfaces. It measures 2 dimensions of affect: Valence and Arousal using separate sliders for each.
- The Game Experience Questionnaire (GEQ) [86]: This questionnaire was developed by Ijsselsteijn et al. to measure player experience in games across a number of factors. The GEQ comprises the following sub-scales: Competence, Immersion, Flow, Tension, Challenge, Positive Affect and Negative Affect. For this study, the in-game version of the GEQ was used. The in-game GEQ is a shorter version that has been designed for repeated measures studies to prevent questionnaire fatigue.

The study consisted of 9 sessions (1 training + 8 experiment), each experiment session took approximately 2 min to complete while the training session took up to 10 min to complete. The training session involved a demonstration by the researcher of the game and interactions. Which was followed by the participant practising the game for 2 levels. For training, a simple level is used with 3 chickens that do not evade the player in small a 5m $\times$ 5m area (data from the training session is not used in analysis). In each experiment session, participants played a level of *Running Chickens* and completed the Affective Slider and the in-game GEQ.

All experiments were conducted during daylight and adequate weather conditions (no signs of rainfall) in a park near Queen Mary University of London's Mile End campus. This is done to minimise the difference in results that may arise due to different locations or poor lighting and weather conditions. The order in which each participant experienced each session of the experiment was randomised to minimise ordering effects on the data collected. Participants were anonymised using IDs and were compensated £10.00 for their participation in the study. The experiment was conducted using a Google Pixel 2 XL mobile device. The following subsection describes the study procedure used for each participant.

4.3.1 Procedure

All participants provided informed consent before participating in the study. At the beginning of the study, participants filled up a questionnaire about their background and previous experience in MAR games. After which they were given a training session about the game and the questionnaires used in the study. During training, the researcher first demonstrated how the game works to the participant over one level of the game, after which participants played 2 training levels and filled up the Affective Slider and the in-game GEQ after each level. During training, the researcher was present with them and they were encouraged to ask any questions about the game or the study procedure.

Once the training was completed, the participants were left alone in the park (while the researcher waited by the entrance of the area), to minimise the effect of the researcher's presence on the data collected. During this time, participants experienced the 8 study sessions of the experiment. In each session, the participant played one of the game levels and filled up the questionnaires. Participants were asked to take a 2-5 min break between sessions to minimise the effects of physical fatigue from the previous session on the data collected from the next one.

While playing *Running Chickens*, if the participants experienced any technical issues (the main one being the AR algorithm losing tracking of the environment), they were asked to proceed to the next session and the data from this session was not used in data analysis. At the end of the study, participants were debriefed about the objectives of the research, all the questions were answered and the experiment was concluded. The experiment took 40-60 min for each participant (depending on the length of the breaks they took during the study).

4.4 Pilot Study

Aspects of the game design and experiment protocol described previously in the chapter were informed using pilot studies. The study was designed with two rounds of pilot testing. The first pilot test was conducted to select optimal game parameters specifically the speed with which the chickens evade players. 4 participants took part in the first round of pilot studies, each participant took part in 3 sessions where they tested different game parameter settings. Each session was conducted on separate days to prevent participant fatigue. The same participants experienced all study sessions as it is important to get a comparative evaluation of the different game parameter settings used in each game.

In the first session, participants played 5 levels of the game where the area and number of chickens were fixed (to the large area and the high number of chickens). In each level, the speed of chickens was varied from slow to fast. This is implemented by varying the acceleration with which the chicken moves away from the player when the evasion mechanic is triggered. The different accelerations tested in this pilot was $a \in \{0, 1.5, 3, 4.5, 6\}$ measured in m/s^2. At the end of the session, participants were asked how their experience was different depending on the speed of the chickens. Most participants preferred a moderate speed for the chickens (either 3 or 4.5 m/s^2), since if the chickens moved too slowly the evasion mechanic was ineffective. While if the evasion speed was too high, participants were easily discouraged as it made capturing the chickens very difficult.

In the second session, participants played all 8 levels of the game, where the speed and number of chickens, area of the level were varied across the high and low settings described in section 4.3. For this pilot, the high condition of the evasion mechanic was set to 4.5 m/s^2. The goal of this session was to test all the game parameter variations within an experiment study. At the end of the second session, the speed of chickens was lowered as participants got easily tired across all 8 rounds of the study. This was observed by participants showing clear signs of fatigue towards the end of the session and was confirmed by their feedback at the end of the session. In the third session, participants tested all 8 levels with the lower chicken speed (3 m/s^2). This was found to be optimal game parameter settings out of all the parameters tested.

The second round of pilot testing was conducted with two participants. This test included the full experiment protocol along with the PX questionnaires at the end of each level. The experiment took an hour for each participant.

The second round of pilot testing worked as expected and only minor aesthetic changes have been noticed at this stage. The 2 main aesthetic changes were the font size of the score text of the game screen had to be increased and the sound effect for when a chicken was decreased in volume. Another main learning from this pilot was that participants found the Affective Slider a bit confusing to use in the context of their experience with the game. This limitation was overcome with added focus on the PX measures during the training session with participants in the main study. The data from both rounds of pilot testing have been discarded and will not be used in analyses.

4.5 Data Collection

During the study, PX, player movement data (measured through the mobile sensors) and game metrics were collected. As each game level was a study condition, each participant contributed 8 games to the data set. This resulted in a total of 320 games played across the study. However, due to some technical crashes, only 298 games were recorded and used in data analysis.

The data set collected is used to analyse the impact of game parameters on PX. Additionally, it is also used to train and evaluate supervised learning classification algorithms to predict PX from the mobile sensor data and game metrics. This section describes the questionnaires, qualitative data, movement data and game metrics collected during the study as well as the demographic information of the sample of participants in this study.

4.5.1 Participants

Participants were recruited using university mailing lists which included PhD, Masters and Undergraduate students from the Electronic Engineering and Computer Science at Queen Mary University of London. The study sample consisted of 40 volunteers (13 female and 27 male) aged 18-44 (24 participants were aged 18-24, 13 were 25-29, 3 were above 30) took part in this study (summarized in table 4.1). When asked about prior experience playing AR games 19 of the subjects had no prior experience. In the remaining 21 subjects: 6 reported having only one experience in the past, 14 played a few times before, and 1 participant played AR games regularly (summarised in table 4.2). All participants provided informed consent before taking part in this study.

4.5.2 Questionnaire Data

The Affective Slider[22] and in-game GEQ[86] are used to collect PX measures of each game. The affective slider uses 2 sliders to provide floating point scores

Table 4.1: Summary of participants' ages.

Age range	Number of participants
18-24	24
25-29	13
Above 30	3

Table 4.2: Summary of participants' previous experience with AR games.

Previous AR experience	Number of participants
No experience	19
Played only once before in the past	6
Played only a few times in the past	14
Played AR games regularly	1

for Valence and Arousal that range from 0-1.0. While the in-game version of the GEQ consists of a 14 item questionnaire where participants use a 7 point Likert scale to provide their level of agreement with the different statements in the questionnaire. The items of the questionnaire are aggregated to provide scores for the following 7 subcomponents of the questionnaire: Competence, Sensory and Imaginative Immersion, Flow, Tension, Challenge, Negative Affect, and Positive Affect. Additionally, qualitative data related to PX was also collected from the interviews conducted with each participant.

4.5.3 Qualitative Data

At the end of the study, a semi-structured interview was conducted with players to get qualitative insights into their experience with the game. The interview mainly focused on their preferences between the different level parameters, any specific strategies players used, main experiential highlights from their gameplay and any design suggestions for future iterations of the game.

4.5.4 Player Behaviour Data

Player Behaviour data has been measured from player movement in-game sessions. The mobile IMU sensors record the position and rotation of the device during the game. This data is recorded at a frequency of *64 Hz* and the discrete-time signals are stored as a 7-dimensional vector: $\alpha \in \{P_X, P_Y, P_Z, R_X, R_Y, R_Z, R_W\}$ for position (in meters) and rotation (in quaternions). Quaternions are a four-dimensional vector (X, Y, Z and W), which is a number system that describes that rotation in a three-dimensional space that prevents gimbal lock, which is a phenomenon that occurs when measuring rotation with Euler angles. Gimbal lock occurs when one of the rotation axes realigns with the other axis and eventually causes loss of one degree of freedom [169].

The player's score (S), which increases as chickens are captured (one point for every chicken captured), is recorded at the same frequency. The time remaining (T) at the end of the round if the player captures all the chickens in the level is also recorded as a measure of performance.

4.6 Analysis and Results

The resulting data set from the experiment was used to understand the impact of different game parameters on PX and to explore the potential of supervised learning to predict dimensions of PX based on player movement and game-based data. The following subsections describe the quantitative, qualitative and supervised learning evaluations conducted as part of this study.

4.6.1 Analysis of Game Parameters influence on Player Experience

The relationship between game parameters on PX is analysed using the three-way ANOVA with a p-value of 0.05. Residual analysis was performed to test for the assumptions of the three-way ANOVA. Normality within groups was assessed using QQplots. In some cases the normality assumption was not met however, as indicated by Blanca et. al [28] the ANOVA test is still a valid option for analysis. It is worth noting the Blanca et. al. proved this hypothesis using a Monte Carlo simulation study. The assumptions of homogeneity of variances was assessed by Levene's test. A posthoc Tukey test was also conducted if significant main or interaction effects were observed from the ANOVA test. The results of each of the PX dimensions are reported below:

Valence: Residuals were not normally distributed ($p < 0.05$) and there was homogeneity of variances ($p > 0.05$). Inspection of the QQplot by experiment conditions showed most groups fall approximately along the reference line shown in figure 4.6, indicating that data within groups show minor deviations from normality. There was a significant main effect from the number of chickens on the Valence score, $F(1, 290) = 4.17$, $p = 0.041$ with an effect size $\eta^2_{partial} = 0.014$. A post hoc Tukey test showed that when the number of chickens was low the valence was higher and differed significantly ($p = 0.041$). This effect is observed in figure 4.7, which shows the plot of the valence scores across the study conditions.

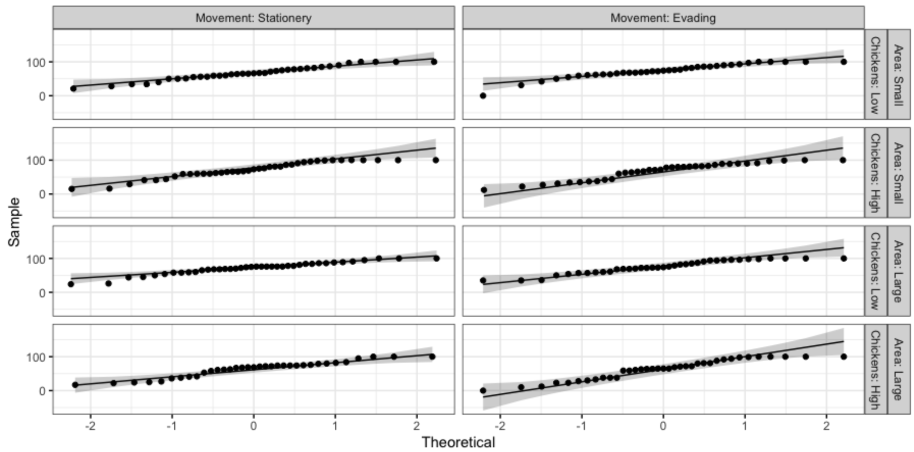


Figure 4.6: The figure shows the **QQplots for the Valence scores** across the study conditions which show minor deviations from normality across the different conditions.

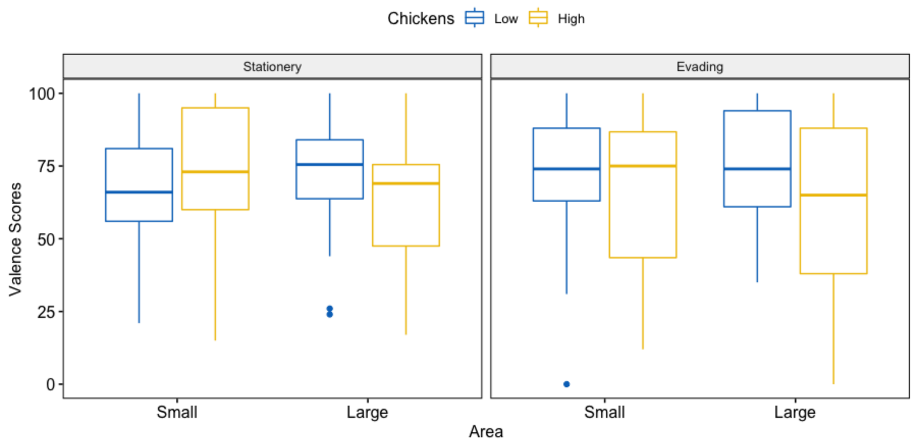


Figure 4.7: The figure shows the **boxplot plot for the Valence scores** across the study conditions. The image shows that when number of chickens is low, valence scores are higher - this effect is clearly observed in the larger game areas.

Arousal: Residuals were not normally distributed (p $<$ 0.05) and there was homogeneity of variances (p $>$ 0.05). Inspection of the QQplot by experiment conditions showed most groups fall approximately along the reference line shown in figure 4.8, indicating that data within groups show minor deviations from normality. There was a significant main effect from the evasion mechanic, F(1, 290) = 5.98, p = 0.015 with an effect size $\eta^2_{partial} = 0.020$. A significant 2-way interaction effect was also observed between the number of chickens and the evasion mechanic, F(1, 290) = 4.61, p = 0.033 with an effect size $\eta^2_{partial} = 0.016$. Posthoc testing showed that when chickens evaded players, arousal scores were significantly higher (p = 0.015). Additionally, when the number of chickens was low, the evasive chickens resulted in significantly higher arousal (p = 0.07). This effect is observed in figure 4.9, which shows the plot of the arousal scores across the study conditions.

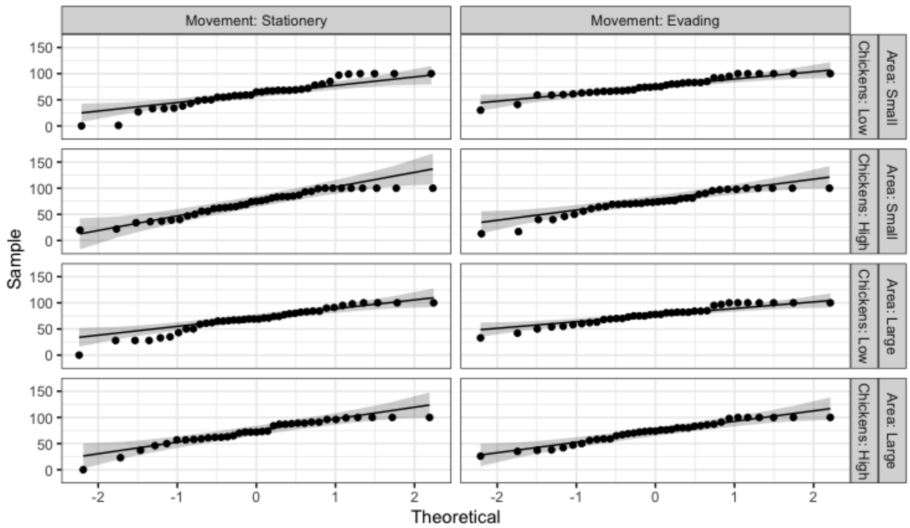


Figure 4.8: The figure shows the **QQplots for the Arousal scores** across the study conditions which show minor deviations from normality across the different conditions.

Competence: A moderate reliability was found for competence scores, Cronbach's $\alpha = 0.67$. Residuals were not normally distributed (p $<$ 0.05) and there was homogeneity of variances (p $>$ 0.05). Inspection of the QQplot by experiment conditions showed most groups fall approximately along the reference line shown in figure 4.10, indicating that data within groups show minor deviations from normality. There was a significant main effect from the number of chickens, F(1, 290) = 10.23, p = 0.001 with an effect size $\eta^2_{partial} = 0.034$. A significant 2-way interaction effect was also observed between the number of

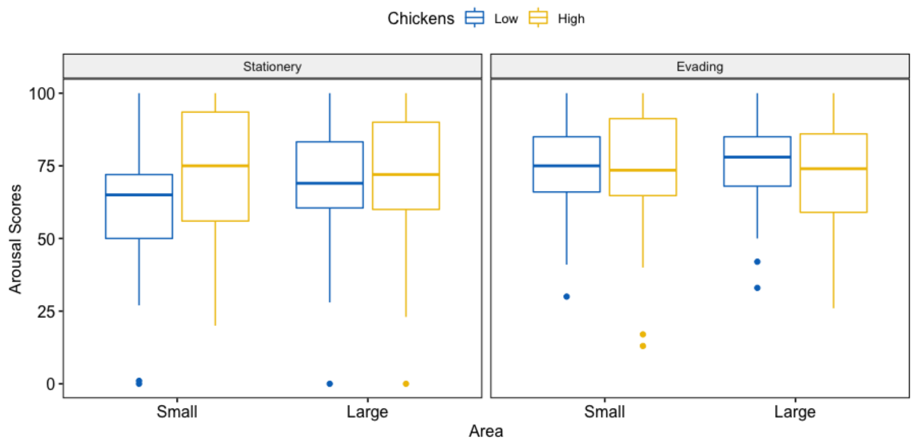


Figure 4.9: The figure shows the **boxplot plot for the Arousal scores** across the study conditions. The image shows that when chickens evade players, arousal scores are higher. Additionally, when the number of chickens are low, evading chickens results in higher arousal scores.

chickens and the evasion mechanic, F(1, 290) = 4.61, $p < 0.001$ with an effect size $\eta^2_{partial} = 0.066$. Posthoc testing showed that when the number of chickens is low, competence scores were significantly higher as compared to when the number of chickens is high ($p = 0.01$). When the number of chickens is low and the chickens are stationary, players reported significantly higher competence scores as compared to levels with a high number of chickens that evade the player ($p = 0.005$). Keeping the number of chickens high and the chickens stationary resulted in significantly higher competence scores as compared to levels with a high number of chickens that evade players ($p < 0.01$). This effect is observed in figure 4.11, which shows the plot of the competence scores across the study conditions.

Sensory and Imaginative Immersion: A moderate reliability was found for immersion scores, Cronbach's $\alpha = 0.68$. Residuals were not normally distributed ($p < 0.05$) and there was homogeneity of variances ($p > 0.05$). Inspection of the QQplot by experiment conditions showed most groups fall approximately along the reference line shown in figure 4.12, indicating that data within groups show minor deviations from normality. There was a significant main effect from the evasion mechanic, F(1, 290) = 12.78, $p = 0.002$ with an effect size $\eta^2_{partial} = 0.031$. Posthoc testing showed that when chickens evaded players, immersion scores were significantly higher ($p = 0.02$). This effect is observed in figure 4.13, which shows the plot of the immersion scores across the study conditions.

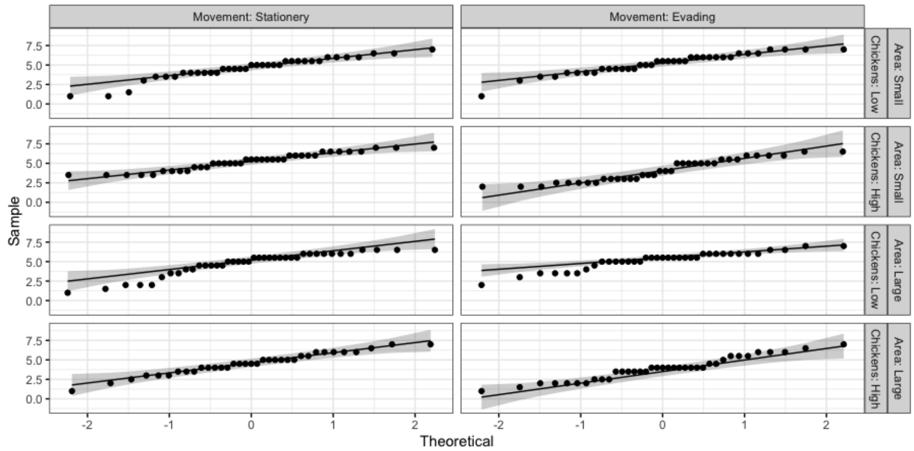


Figure 4.10: The figure shows the **QQplots for the Competence scores** across the study conditions which show minor deviations from normality across the different conditions.

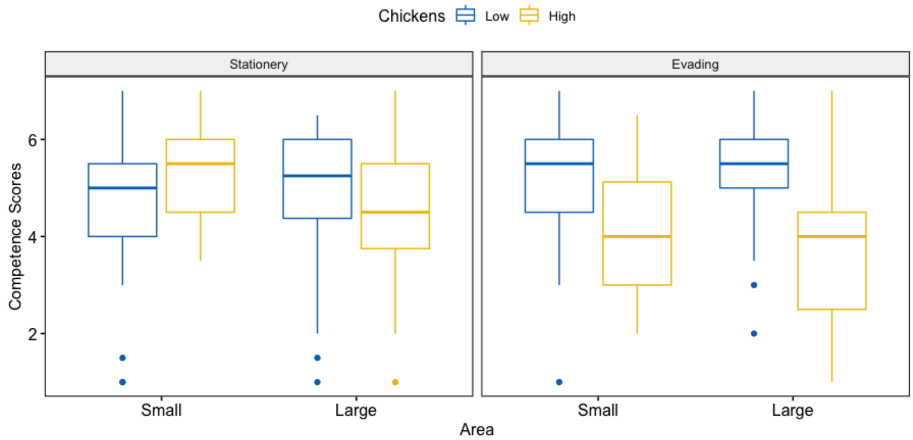


Figure 4.11: The figure shows the **boxplot plot for the Competence scores** across the study conditions. The image shows that when the number of chickens is low competence scores are higher. Additionally, when the number of chickens is low and the chickens are stationary, players reported significantly higher competence scores as compared to levels with a high number of chickens that evade the player. Finally, when the number of chickens high and the chickens stationary resulted in significantly higher competence scores as compared to levels with a high number of chickens that evade players

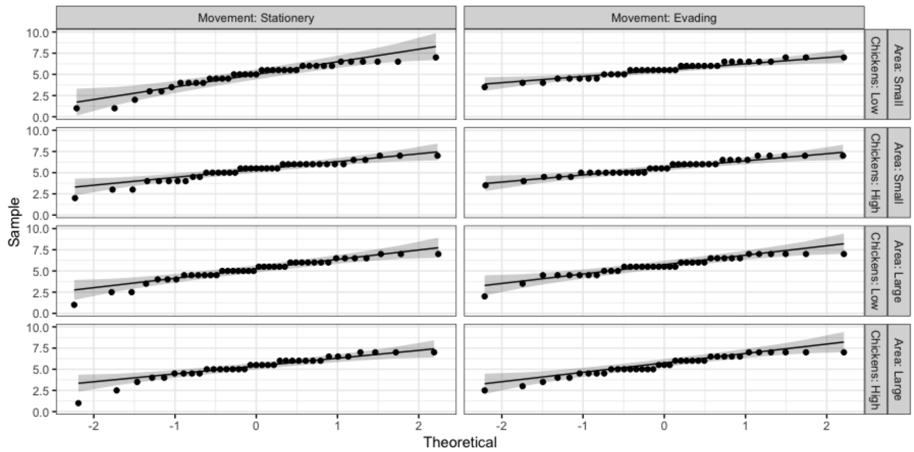


Figure 4.12: The figure shows the **QQplots for the Immersion scores** across the study conditions which show minor deviations from normality across the different conditions.

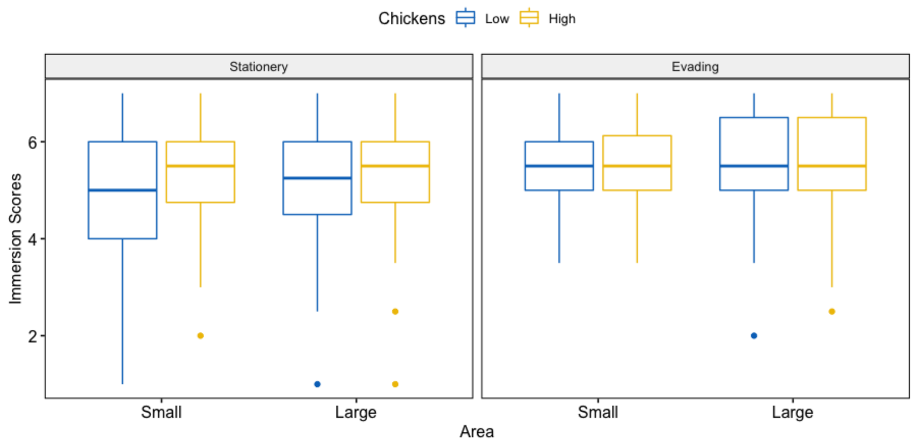


Figure 4.13: The figure shows the **boxplot plot for the Immersion scores** across the study conditions. The image shows that when chickens evade players, immersion scores are higher.

Flow: A moderate reliability was found for flow scores, Cronbach's $\alpha = 0.69$. Residuals were not normally distributed ($p < 0.05$) and there was homogeneity of variances ($p > 0.05$). Inspection of the QQplot by experiment conditions showed most groups fall approximately along the reference line shown in figure 4.14, indicating that data within groups show minor deviations from normality. There was a significant main effect from the number of chickens, $F(1, 290) = 4.40$, $p = 0.036$ with an effect size $\eta^2_{partial} = 0.014$. There was also a significant main effect from the evasion mechanic, $F(1, 290) = 12.81$, $p = 0.018$ with an effect size $\eta^2_{partial} = 0.018$. Posthoc testing showed that when the number of chickens was high, flow scores were significantly higher ($p = 0.036$). Additionally, when chickens evade players, flow scores were significantly higher ($p = 0.018$). This effect is observed in figure 4.15, which shows the plot of the flow scores across the study conditions.

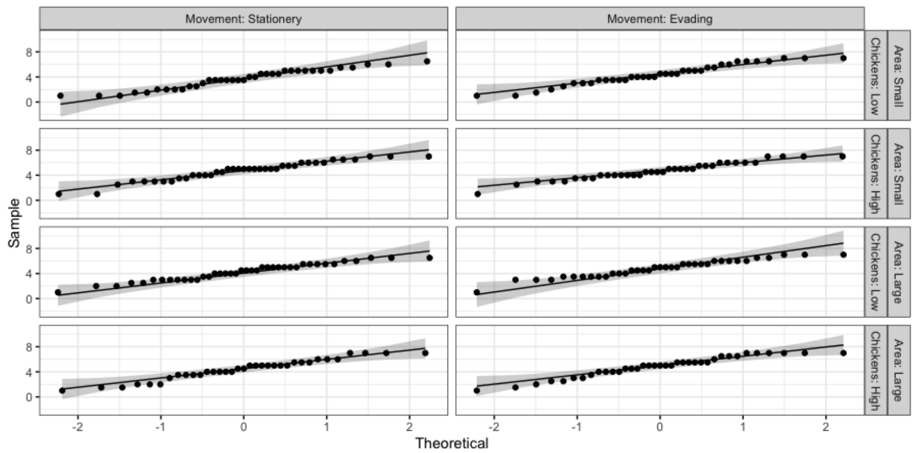


Figure 4.14: The figure shows the **QQplots for the Flow scores** across the study conditions which show minor deviations from normality across the different conditions.

Tension: A high reliability was found for tension scores, Cronbach's $\alpha = 0.81$. Residuals were not normally distributed ($p < 0.05$) and there was homogeneity of variances ($p > 0.05$). Inspection of the QQplot by experiment conditions showed most groups fall approximately along the reference line shown in figure 4.16, indicating that data within groups show minor deviations from normality. There was a significant main effect from the number of chickens, $F(1, 290) = 10.61$, $p = 0.001$ with an effect size $\eta^2_{partial} = 0.034$. There was also a significant main effect from the evasion mechanic, $F(1, 290) = 4.26$, $p = 0.039$ with an effect size $\eta^2_{partial} = 0.014$. A significant 2-way interaction effect was also observed between the game area and the number of chickens, $F(1, 290) = 5.92$, $p =$

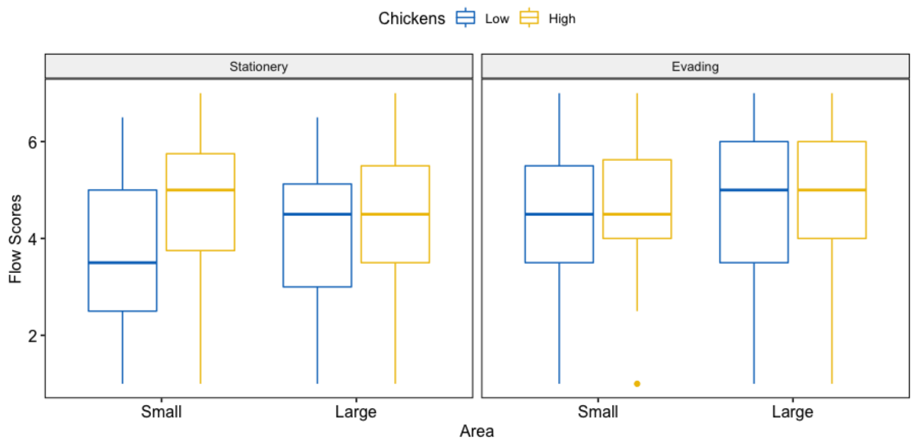


Figure 4.15: The figure shows the **boxplot plot for the Flow scores** across the study conditions. The image shows that when number of chickens is high flow scores are higher. Additionally, when chickens evade players, flow scores were significantly higher.

0.015 with an effect size $\eta^2_{partial} = 0.019$. Posthoc testing showed that when the number of chickens was high, tension scores were significantly higher (p = 0.001). Additionally, tension scores are significantly higher when chickens evade the player as compared to when they are stationary (p = 0.039). A large area and a high number of chickens result in significantly higher tension scores as compared to levels with a small area and a low number of chickens (p = 0.008). Additionally, a large area and a high number of chickens result in significantly higher tension scores as compared to a large area and a low number of chickens ($p < 0.001$). This effect is observed in figure 4.17, which shows the plot of the tension scores across the study conditions.

Challenge: A moderate reliability was found for challenge scores, Cronbach's $\alpha = 0.69$. Residuals were not normally distributed ($p < 0.05$) and there was homogeneity of variances ($p > 0.05$). Inspection of the QQplot by experiment conditions showed most groups fall approximately along the reference line shown in figure 4.18, indicating that data within groups show minor deviations from normality. There was a significant main effect from the game area, $F(1, 290) = 8.58$, p = 0.003 with an effect size $\eta^2_{partial} = 0.022$. There was also a significant main effect from the number of chickens, $F(1, 290) = 54.76$, $p < 0.001$ with an effect size $\eta^2_{partial} = 0.142$. There was also a significant main effect from the evasion mechanic, $F(1, 290) = 27.31$, $p < 0.001$ with an effect size $\eta^2_{partial} = 0.071$. Posthoc testing showed that when the game area was large, challenge scores were significantly higher (p = 0.003). Additionally, challenge scores were

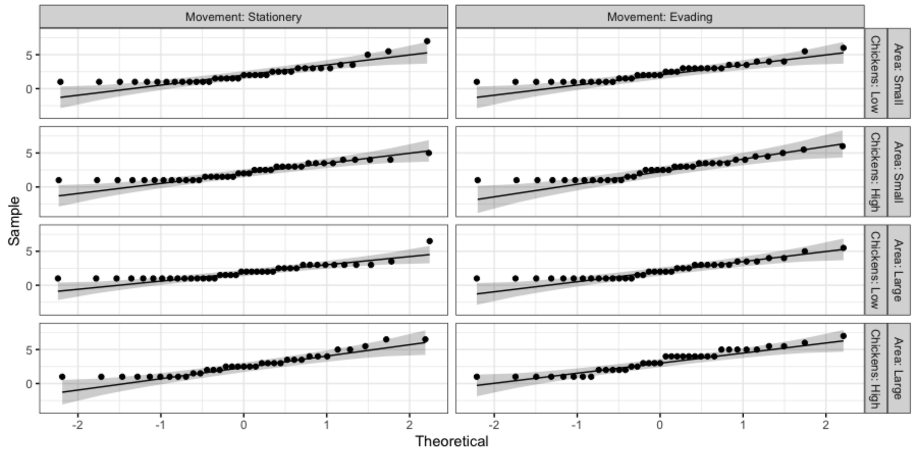


Figure 4.16: The figure shows the **QQplots for the Tension scores** across the study conditions which show minor deviations from normality across the different conditions.

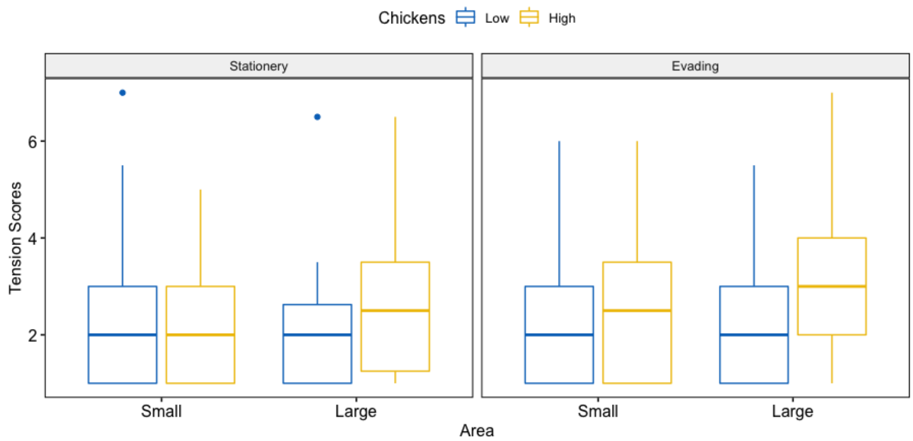


Figure 4.17: The figure shows the **boxplot plot for the Tension scores** across the study conditions. The image shows that when the number of chickens is high tension scores are higher. Additionally, when chickens evade players, tension scores were higher. Finally, a large area and a high number of chickens result in significantly higher tension scores as compared to a large area and a low number of chickens.

significantly higher when the number of chickens was high as compared to when it was low ($p < 0.001$). Finally, challenge scores were significantly higher when chickens evade players as compared to when they were stationary ($p < 0.001$). This effect is observed in figure 4.19, which shows the plot of the challenge scores across the study conditions.

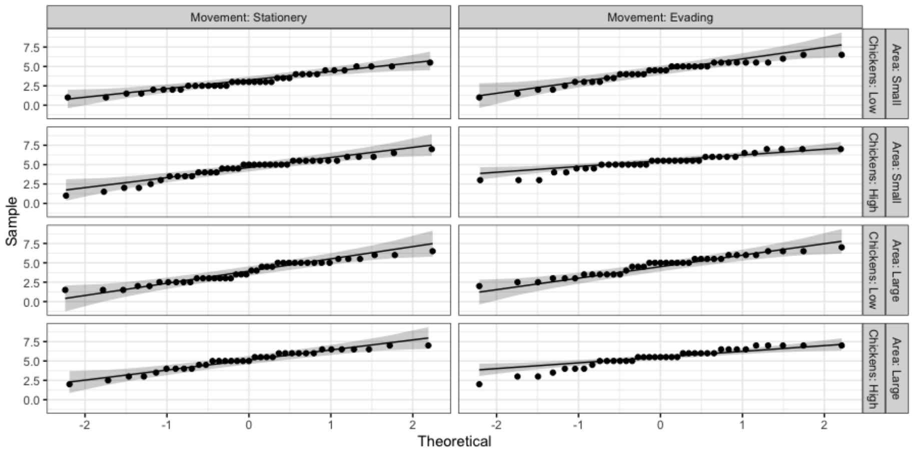


Figure 4.18: The figure shows the **QQplots for the Challenge scores** across the study conditions which show minor deviations from normality across the different conditions.

Positive Affect: A good reliability was found for positive affect scores, Cronbach's $\alpha = 0.78$. Residuals were not normally distributed ($p < 0.05$) and there was homogeneity of variances ($p > 0.05$). Inspection of the QQplot by experiment conditions showed most groups fall approximately along the reference line shown in figure 4.20, indicating that data within groups show minor deviations from normality. There was a significant 2-way interaction effect observed between the game area and the number of chickens, $F(1, 290) = 4.91$, $p = 0.027$ with an effect size $\eta^2_{partial} = 0.016$. A significant 2-way interaction effect was also observed between the number of chickens and the evasion mechanic, $F(1, 290) = 5.63$, $p = 0.018$ with an effect size $\eta^2_{partial} = 0.018$. Posthoc testing showed that when the game area is large, more chickens lead to significantly lower positive affect scores as compared to levels of the same area with few chickens ($p = 0.046$). Additionally, in levels where chickens evade players, more chickens lead to significantly lower scores as compared to similar levels with fewer chickens ($p = 0.041$). This effect is observed in figure 4.21, which shows the plot of the positive affect scores across the study conditions.

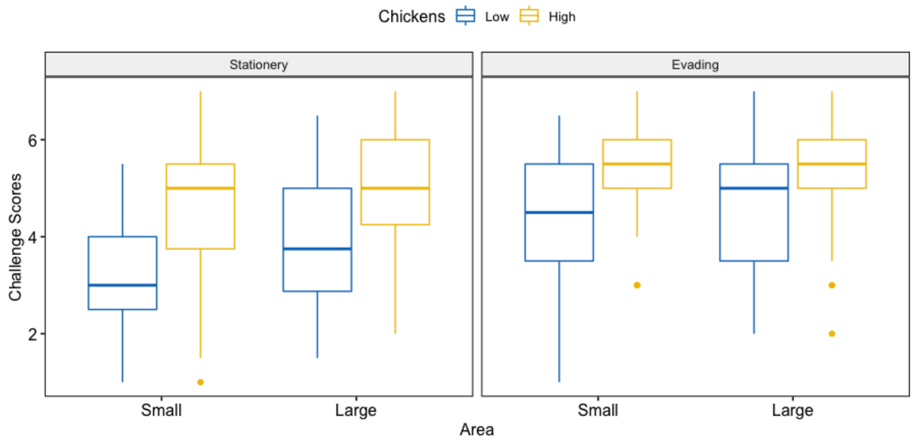


Figure 4.19: The figure shows the **boxplot plot for the Tension scores** across the study conditions. The image shows that when the game area is large, challenge scores were higher. Additionally, when the number of chickens is high, challenge scores were higher. Finally, chickens evade players, challenge scores are higher.

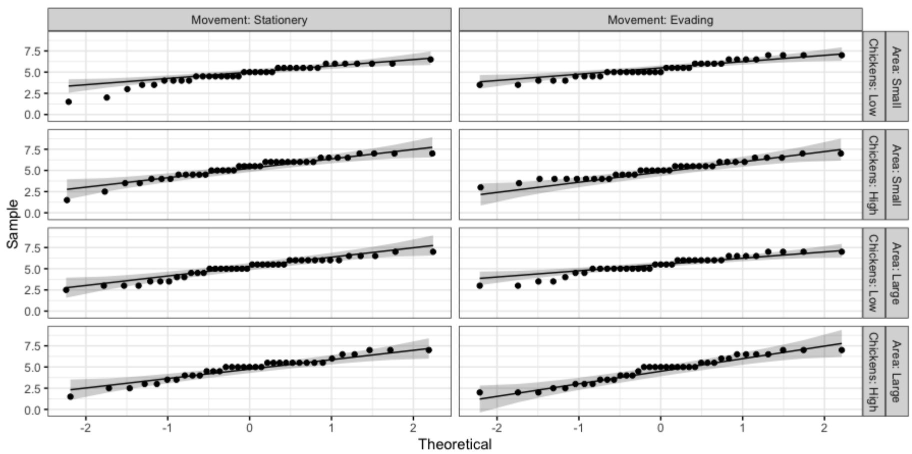


Figure 4.20: The figure shows the **QQplots for the Positive Affect scores** across the study conditions which show minor deviations from normality across the different conditions.

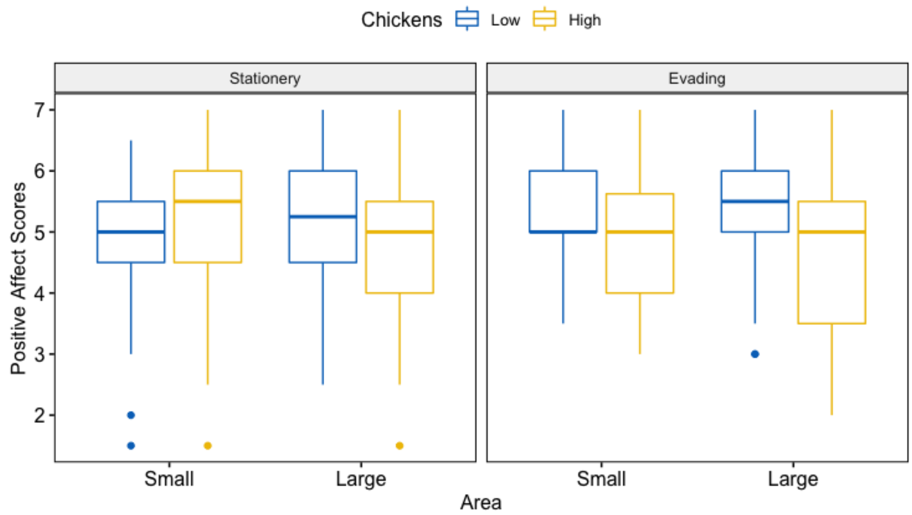


Figure 4.21: The figure shows the **boxplot plot for the Positive Affect scores** across the study conditions. The image shows that when the game area is large, a high number of chickens lead to lower positive affect scores as compared to levels with the same area and a low number of chickens. Additionally, in levels where chickens evade players, more chickens lead to significantly lower scores as compared to similar levels with fewer chickens.

Negative Affect: A poor reliability was found for negative affect scores, Cronbach's $\alpha = 0.49$. Residuals were not normally distributed ($p < 0.05$) and there was homogeneity of variances ($p > 0.05$). Inspection of the QQplot by experiment conditions showed most groups fall approximately along the reference line shown in figure 4.22, indicating that data within groups show minor deviations from normality. No significant main or interaction effects were observed from the game parameters on negative affect scores. The boxplot of Negative score across is shown in fig 4.22 which shows that scores are similar across all conditions.

This subsection reported the results of statistical analysis of the impact of game parameters on PX. Table 4.3 summarises the impact of PX on each of the game parameters used in the game.

4.6.2 Grounded Analysis of Semi-Structured Interviews

This subsection presents the main themes that emerged from the grounded theory analysis of the interview data. The semi-structured interviews were analysed and categorized according to the grounded theory analysis described by Strauss as the constant comparative method of data analysis[206]. The inter-

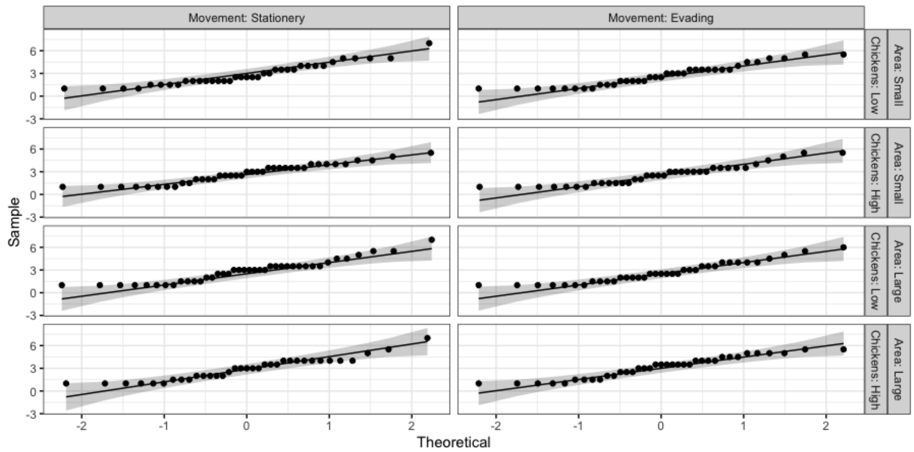


Figure 4.22: The figure shows the **QQplots for the Negative Affect scores** across the study conditions which show minor deviations from normality across the different conditions.

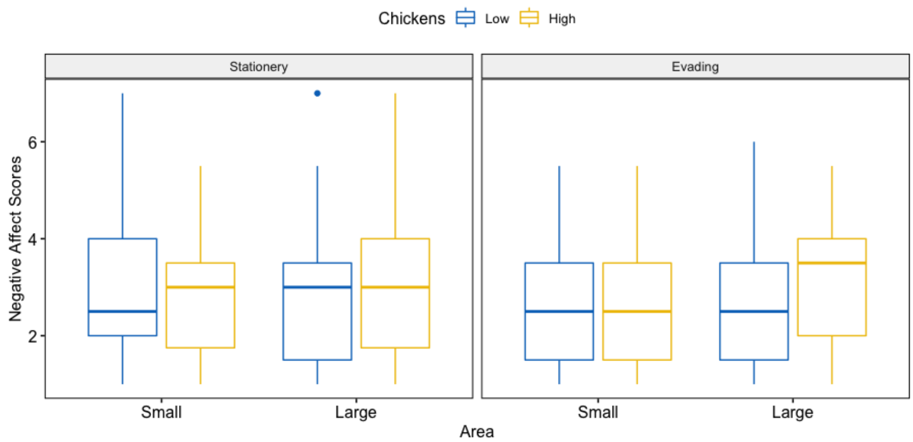


Figure 4.23: The figure shows the **boxplot plot for the Negative Affect scores** across the study conditions. The image shows that scores are similar across all conditions.

Table 4.3: Summary of results of statistical analysis of game parameters on PX.

PX dimension	Game Parameters	F-value	$\eta^2_{partial}$	Effect Description
Valence	G_{No} (main effect)	4.17	0.014	Decreases when number is high.
Arousal	G_{Mov} (main effect)	5.98	0.020	Increases when chickens evade players.
	$G_{No} \times G_{Mov}$ (Interaction effect)	4.61	0.016	Increases when number is low and chickens evade players, as compared level with a low number and chickens are stationary.
Competence	G_{No} (main effect)	10.23	0.034	Decreases when number is high.
	$G_{No} \times G_{Mov}$ (Interaction effect)	4.61	0.066	Increases when the number is low and the chickens are stationary, as compared to levels with a high number of chickens that evade players. Increases when the number of chickens high and the chickens stationary as compared to levels with a high number of chickens that evade players.
Immersion	G_{Mov} (main effect)	12.78	0.031	Increases when chickens evade players.
Flow	G_{No} (main effect)	4.40	0.014	Increases when number is high.
	G_{Mov} (main effect)	12.81	0.018	Increases when chickens evade players.
Tension	G_{No} (main effect)	10.61	0.034	Increases when number is high.
	G_{Mov} (main effect)	4.26	0.014	Increases when chickens evade players.
	$G_{Area} \times G_{No}$ (Interaction effect)	5.92	0.019	Increases when area is large with a high number as compared to levels with a small area and low number. Increases when a large area with a high number as compared to levels with large area and small number.
Challenge	G_{Area} (main effect)	8.58	0.022	Increases when area is large.
	G_{No} (main effect)	54.76	0.142	Increases when number is high.
	G_{Mov} (main effect)	27.31	0.071	Increases when chickens evade players.
Positive Affect	$G_{Area} \times G_{No}$ (Interaction effect)	4.91	0.016	Increases when area is large and number is high as compared to levels with large area with low number.
	$G_{No} \times G_{Mov}$ (Interaction effect)	5.63	0.018	Increases when chickens evade players with high number, as compared to levels where chickens evade players with low number.
Negative Affect	No main or interaction effects			

view data were summarized into different themes, which were confirmed and modified throughout the analyses. The data analysis consisted of open, axial and selective coding. Open coding included repeated readings of the interviews and an in-depth, line-by-line analysis of the data. Using open coding, data were coded under various headings according to their content. In the axial coding, categories were linked together, with sub-categories describing the specific category. In the final selective coding, the categories were linked together, which resulted in a core category[206]. These core categories form the main insights from the qualitative data. The core categories that emerged were: Game Experience, Game Progression, Playing Strategy, Health Benefit, Game Preferences and Design Suggestions. Each one is described below:

Player Experience: The game was positively received by participants in the study. Participants used adjectives such as fun, novel and engaging to describe their experience playing the game. Participants felt highly immersed in the game, 5 of them reported that they were so immersed in the game they lost their sense of direction. While another noted that: "...[I] was so engaged, [I] swore to myself many times...". In one case the participant reported that it felt similar to chasing actual animals or agents with intelligent thought. The majority of participants reported that "near misses" where chickens narrowly avoided getting captured as highly motivating. Finally, it is interesting to note that several participants reported the repeated switching of attention between the AR world (seen through the mobile device) and the non-AR world (around the mobile screen). When asked to elaborate, they reported that it was challenging to use the AR view to navigate around the game area (especially when moving fast). Participants would use the AR view to search for chickens and the non-AR views to navigate through the game space. It is important to note that participants could easily perform fast movements in the same area and moderate walking while looking through the mobile screen, running is what shifted the participant's focus away from the AR viewport to the non-AR world.

Game Progression: The order of the levels were randomized in this study to minimize ordering effects in the PX data associated with each level. However, this approach was confusing to the majority of participants since they expected a system of progressive difficulty in the game. While most expected the game to get more difficult as levels progressed 3 participants noted that finding easier levels after difficult ones reassured their sense of competence in the game. However, extremely easy levels (eg: small area and a few stationary chickens) was considered boring if experienced towards the end of the experiment session. While 2 participants noted that it would be nice to be able to choose between easy or difficult levels, the remaining participants (35 out of 40) expected the game to have a steadily increasing difficulty in the progression of their game experience.

Playing Strategy: There was a common strategy that 5 participants reported using in levels where chickens evaded them. They would herd the chickens to the corner of the levels and then focus on capturing clusters of them from these corners. The other 35 participants did not report any specific strategy while playing the game. These participants reported running to the chicken that was closest to them. If these participants retrospectively noticed a closer chicken they missed, it resulted in some amount of frustration at themselves.

Health Benefit: Several participants (31 out 40) noted that playing the game felt like light exercise, made them feel more active and energetic. While participants reported that this game would be beneficial in motivating their physical activity, it is important to note that participants were referring to game mechanics. In the current state, most participants felt like the game would need a narrative and more complex reward systems for them to use the game regularly.

Game Preferences: Participants were asked about their individual preferences regarding each of the game parameters. Half of them preferred a large area while the other half preferred the smaller one. Only 1 person reported having no preference for large or small levels. It is important to note that participants who preferred larger areas also tended to like sports and other physical activities. Conversely, participants who preferred smaller areas tended to not enjoy running or physical activity in general. Participants' preferences for the number of chickens in a level was not as clear. While participants who enjoy physical activity enjoyed a larger number of chickens as it challenged them more and gave them a more rewarding experience. The other participants reported having no preference for the number. They reported preferring few chickens in large areas and more chickens in smaller areas. Only 2 participants reported preferring fewer chickens to many in all games. Finally, 39 out of 40 participants preferred the chickens evading them rather than being stationary. Only 1 person reported having no preference between running and stationary chickens. It is interesting to note that even participants who did not enjoy running and physical activity still preferred games where chickens ran away from them.

Design Suggestions: The participants had many suggestions for improving the PX of the game. This included improvements to the games audio and visual aesthetics such as better designed models and animations, additional digital content (different types of animals and a narrative for the game), additional interactions with the chickens (eg: trying to capture chickens that can fly). A large portion of participants (26 out of 40) requested social play. They wanted to be able to play with their friends in a co-located AR space either competitively or cooperatively.

This subsection reported the findings from the qualitative interview data. The main categories that emerged from the analysis were Player Experience, Game Progression, Player Strategy, Health Benefit, Game Preferences and Design Suggestions. Participants reported positive experiences while playing the game. Additionally, it was observed that players found the act of chasing and catching chickens highly motivating. Additionally, it was interesting to observe that participants reported switching attention between AR and non-AR environments as they navigated through space. A majority of participants (35 out of 40) reported having expectations that game levels would progress from easy to hard levels (which was not the case). This lead to an easy level being perceived as boring if they appeared towards the end of the experiment. Some participants (3 out of 40) found that easy levels appearing after more challenging ones reassured their sense of competence within the game. Some participants (5 out of 40) reported using a strategy of herding chickens together in clusters to be able to capture them. Players (31 out of 40) also reported that game-play has the effect of a light exercise making them feel active and energetic. In terms of preferences for game parameters, there was a split in preferences for the area and number of chickens parameters, participants (19 out of 40) who enjoy physical activity prefer large areas with a high number of chickens while participants (20 out of 40) who do not like to engage in physical activity preferred smaller areas and did not have any preference regarding the number of chickens. It was interesting to observe that irrespective of their preference for physical activity, most participants (39 out of 40) enjoyed levels where chickens evaded players. Participants also provided a number of design recommendations to improve the aesthetics of the game (eg: more realistic chicken models and additional animals), social play (leader-boards, competitive and cooperative multiplayer gaming) was highly requested by participants in future versions of the game.

4.6.3 Player Experience Classification using Supervised Learning

Supervised learning algorithms were tested to explore the potential of predicting PX of an individual in a game level based on their behaviour and performance. This section first describes the supervised learning pipeline used and then presents the results from this evaluation.

Supervised Learning Pipeline

The dataset is first pre-processed and Player Behaviour Features (PBFs) are extracted from the movement data and performance features are extracted from the game metrics recorded. These features along with the game parameters referred to as Controllable Game Features (CGFs) were used to predict PX. Here, these features are the input of a model that aims to predict the measured dimension of PX (eg: Valence).

In this work, predicting PX is treated as a supervised classification problem. The numerical values for the different PX measures are transformed into binary categorical values. This transformation is based on the median of the reported PX measure by that participant across all conditions in the study. For instance, if the Valence score for a specific game is greater than the median Valence score reported by that player across all games played, then it is set to 1 (and 0 if it is lesser than the median). This player-based transformation of the ground truth is applied as it reduces the differences in PX measures due to subjectivity between participants. Additionally, the median statistic is selected due to the limited number of data samples available per participant. The median is a more robust measure of centrality since the mean is easily influenced by extreme values when there is a low number of data points.

Data Pre-processing: The procedure for data preprocessing of the movement data has been adapted from the previous study. The movement data is first filtered to reduce high-frequency artefacts. The rotation vectors along x, y and z are constructed from the rotation quaternions recorded from each game. Following which velocity, acceleration and jitter are extracted from the position and rotation by taking the first, second and third differences of time series data (for position and rotation). This results in 3D vectors containing data for velocity, acceleration, jerk, angular velocity, angular acceleration and angular jerk along the x, y and z axis; $6 \times 3D$ vectors in total. Finally, the dimensionality of this data is reduced by taking the Euclidean norm for the x, y and z axis for each. This 6-D vector is the final output of the data pre-processing phase for each game, given by $\beta \in \{V, A, J, RV, RA, RJ\}$ and this is done at 64 Hz over the duration of the game.

Feature Extraction: PBFs are extracted from the pre-processed movement data: V, A, J, RV, RA, RJ. Similar to the previous study, the same 10 features that are extracted for each dimension of β resulting in 60 movement features, refer to table 3.4 for the list of features. Performance features such as Completion and Time Remaining are computed from game sessions. Completion is the ratio between the score at the end of the game and the maximum possible score from that game. Time Remaining is the number of seconds left at the end of the game (if the player collects all the chickens before the time limit). Which adds up to a total of 62 PBFs (60 + 2). Feature scaling is applied to these features by subtracting the sample mean and scaling sample variance. Finally, the 3 game parameters referred to as controllable game features (CGFs) are considered as inputs for the model. This results in a total of 65 features that are used as input to the learning algorithms tested in this study.

Player Experience Classification: This study compares the performance of 4 classifiers for this problem: logistic regression, linear discriminant analysis (LDA), support vector machines (SVM) and the XGBoost algorithm. The first 3 algorithms have been tested in the previous study, however, it is unclear if an ensemble tree-based approach such as XGBoost can provide more accurate predictions for this problem. XGBoost and other decision trees based algorithms have been successfully applied to the domain of player modelling in traditional (non-AR) digital games [140, 61, 62]. In [61, 62] the XGBoost classifier was shown more successful than other learning algorithms tested in predicting player's preferences for fun in action-adventure games. Finally, feature scaling is not employed for the XGBoost classifier as it is not required.

The previous study observed that model performance improves through feature selection techniques which is consistent with work conducted by Pedersen et. al [167]. While there are a large number of approaches, the experiment used sequential floating forward selection (SFFS) in this study since it is often used in similar work [167], additionally SFFS was the best performing feature selection technique in the previous study conducted. SFFS is is a bottom-up search algorithm that tries to find the best performing feature set. It starts with the best performing single feature and adds new features from the remaining set such that model performance of the new set generates the best possible overall performance over other potential features for addition. When a forward step is performed the algorithm also checks if a feature from the existing set can be excluded to improve overall model performance.

Evaluation Metrics: In this study, model performance is evaluated using 10-fold cross-validation scores (similar to the previous study). Additionally, in order to test to what extent these models can generalise to unseen players Leave-one-subject-out cross-validation (Loso-CV) is used as an evaluation metric. In Loso-CV data from one participant is used as a test set, the model after feature selection is trained using the remaining data (39 players) and evaluated on the test participant data. This process is repeated 40 times for each player in the data set. The mean accuracy and the standard deviation is used to test the generalisability of the model to new players. This is an important metric to consider for the application of such models in game adaptation systems.

All algorithms were implemented using their default hyperparameter values. The supervised learning pipeline was implemented in python, the SciPy library[222] was used to convert data from quaternions to rotation vectors, the SciKit learn library [168] was used for feature scaling and to implement the logistic regression, LDA and SVM models, and the XGBoost[38] library was used for this study.

Supervised Learning Evaluation Results

The results from supervised learning techniques are presented here, the extracted features are used to predict the different dimensions of PX. It has been a common observation across all dimensions and types of classifiers that the base (all 65 features) performance without feature selection performs poorly. Features selection tends to improve classification accuracies across all the classifiers and dimensions of PX tested. All the accuracies from the 10-fold CV and the LOSO-CV are reported for the models after feature selection. Results for each of the dimensions of PX are reported below.

Valence: Random chance classification for valence was 50.02%. All classifiers perform higher than random chance on 10-fold CV accuracies. However, accuracies for Logistic Regression, LDA and SVM classifiers decrease to close to random chance when LOSO-CV accuracies are computed. The XGBoost classifier was the best performing model tested with a 10-fold CV of $64.79 \pm 7.69\%$ and LOSO-CV of $60.84 \pm 17.66\%$ which is higher than random chance classification. Table 4.4 shows the summary of results for all the classifiers tested for valence.

Table 4.4: Summary of results from supervised learning to classify Valence.

	10-CV Acc	LOSO-CV Acc
Logistic Regression	$63.11 \pm 6.41\%$	$52.16 \pm 20.43\%$
LDA	$64.80 \pm 7.08\%$	$52.84 \pm 8.44\%$
SVM	$64.45 \pm 5.23\%$	$47.86 \pm 5.31\%$
XGBoost	$64.79 \pm 7.69\%$	$60.84 \pm 17.66\%$

Arousal: Random chance classification for arousal was 50.32%. All classifiers perform higher than random chance on 10-fold CV accuracies. However, accuracies for Logistic Regression, LDA and SVM classifiers decrease to close to random chance when LOSO-CV accuracies are computed. The XGBoost classifier was the best performing model tested with a 10-fold CV of $66.12 \pm 6.48\%$ and LOSO-CV of $61.69 \pm 18.96\%$ which is higher than random chance classification. Table 4.5 shows the summary of results for all the classifiers tested for arousal.

Table 4.5: Summary of results from supervised learning to classify Arousal.

	10-CV Acc	LOSO-CV Acc
Logistic Regression	$63.77 \pm 10.43\%$	$53.78 \pm 12.06\%$
LDA	$63.43 \pm 9.99\%$	$50.98 \pm 11.86\%$
SVM	$65.40 \pm 4.50\%$	$54.51 \pm 8.90\%$
XGBoost	$66.12 \pm 6.48\%$	$61.69 \pm 18.96\%$

Competence: Random chance classification for competence was 51.62%. All classifiers perform higher than random chance on 10-fold CV accuracies. Accuracies for Logistic Regression, LDA and SVM classifiers decrease but remain higher than random chance when LOSO-CV accuracies are computed. The XGBoost classifier was the best performing model tested with a 10-fold CV of $73.20 \pm 6.18\%$ and LOSO-CV of $69.61 \pm 17.37\%$. Table 4.6 shows the summary of results for all the classifiers tested for competence.

Table 4.6: Summary of results from supervised learning to classify Competence.

	10-CV Acc	LOSO-CV Acc
Logistic Regression	$73.57 \pm 13.53\%$	$59.96 \pm 14.62\%$
LDA	$72.22 \pm 12.11\%$	$60.60 \pm 13.42\%$
SVM	$74.21 \pm 10.96\%$	$60.60 \pm 13.71\%$
XGBoost	$73.20 \pm 6.18\%$	$69.61 \pm 17.37\%$

Immersion: Random chance classification for immersion was 52.00%. All classifiers perform higher than random chance on 10-fold CV accuracies. Accuracies for Logistic Regression and LDA classifiers decrease but remain higher than random chance when LOSO-CV accuracies are computed, while the SVM classifier shows a performance below random chance for this. The XGBoost classifier was the best performing model tested with a 10-fold CV of $71.13 \pm 12.87\%$

and LOSO-CV of $68.86 \pm 16.95\%$. Table 4.7 shows the summary of results for all the classifiers tested for immersion.

Table 4.7: Summary of results from supervised learning to classify Immersion.

	10-CV Acc	LOSO-CV Acc
Logistic Regression	$67.10 \pm 7.69\%$	$61.18 \pm 14.34\%$
LDA	$67.43 \pm 4.81\%$	$63.33 \pm 15.56\%$
SVM	$70.08 \pm 4.75\%$	$39.43 \pm 13.99\%$
XGBoost	$71.13 \pm 12.87\%$	$68.86 \pm 16.95\%$

Flow: Random chance classification for flow was 52.00%. All classifiers perform higher than random chance on 10-fold CV accuracies. Accuracies for all classifiers decrease but remain higher than random chance when LOSO-CV accuracies are computed. In this case, there does not appear to be a clear best performing classifier for flow classification. Table 4.8 shows the summary of results for all the classifiers tested for flow.

Table 4.8: Summary of results from supervised learning to classify Flow.

	10-CV Acc	LOSO-CV Acc
Logistic Regression	$70.18 \pm 11.60\%$	$61.31 \pm 14.78\%$
LDA	$71.54 \pm 12.29\%$	$60.24 \pm 15.81\%$
SVM	$71.50 \pm 6.96\%$	$60.87 \pm 13.02\%$
XGBoost	$69.82 \pm 7.31\%$	$64.62 \pm 18.16\%$

Tension: Random chance classification for tension was 58.82%. All classifiers perform higher than random chance on 10-fold CV accuracies. Accuracies for SVM and LDA classifiers decrease but remain higher than random chance when LOSO-CV accuracies are computed, while the Logistic Regression classifier shows a performance below random chance for this. The XGBoost classifier was the best performing model tested with a 10-fold CV of $76.52 \pm 5.51\%$ and LOSO-CV of $71.16 \pm 20.69\%$. Table 4.9 shows the summary of results for all the classifiers tested for tension.

Challenge: Random chance classification for tension was 51.62%. All classifiers perform higher than random chance on 10-fold CV accuracies. Accuracies for all classifiers decrease but remain higher than random chance when

Table 4.9: Summary of results from supervised learning to classify Tension.

	10-CV Acc	LOSO-CV Acc
Logistic Regression	$75.87 \pm 4.30\%$	$57.01 \pm 21.20\%$
LDA	$72.81 \pm 5.32\%$	$70.06 \pm 20.23\%$
SVM	$73.82 \pm 2.45\%$	$70.48 \pm 20.79\%$
XGBoost	$76.52 \pm 5.51\%$	$71.16 \pm 20.69\%$

LOSO-CV accuracies are computed. The XGBoost classifier was the best performing model tested with a 10-fold CV of $74.79 \pm 9.01\%$ and LOSO-CV of $71.06 \pm 16.06\%$. Table 4.10 shows the summary of results for all the classifiers tested for challenge.

Table 4.10: Summary of results from supervised learning to classify Challenge.

	10-CV Acc	LOSO-CV Acc
Logistic Regression	$76.14 \pm 4.96\%$	$69.33 \pm 18.36\%$
LDA	$75.80 \pm 7.75\%$	$66.39 \pm 18.98\%$
SVM	$75.17 \pm 7.62\%$	$59.87 \pm 12.98\%$
XGBoost	$74.79 \pm 9.01\%$	$71.06 \pm 16.06\%$

Positive Affect: Random chance classification for positive affect was 51.62%. All classifiers perform higher than random chance on 10-fold CV accuracies. However, accuracy for Logistic Regression classifier decrease to close to random chance when LOSO-CV accuracies are computed and the accuracies for SVM and LDA decrease to below random chance for this. The XGBoost classifier was the best performing model tested with a 10-fold CV of $72.13 \pm 10.40\%$ and LOSO-CV of $67.19 \pm 17.69\%$. Table 4.11 shows the summary of results for all the classifiers tested for positive affect.

Table 4.11: Summary of results from supervised learning to classify Positive Affect.

	10-CV Acc	LOSO-CV Acc
Logistic Regression	$68.75 \pm 7.06\%$	$53.20 \pm 17.51\%$
LDA	$67.12 \pm 7.45\%$	$38.25 \pm 14.52\%$
SVM	$69.75 \pm 5.64\%$	$40.54 \pm 13.74\%$
XGBoost	$72.13 \pm 10.40\%$	$67.19 \pm 17.69\%$

Negative Affect: Random chance classification for negative affect was 52.88%. All classifiers perform higher than random chance on 10-fold CV accuracies. However, accuracies for Logistic Regression and LDA classifiers decrease to below random chance when LOSO-CV accuracies are computed while the other classifiers tested remain higher than random chance for the same. The SVM classifier was the best performing model tested with a 10-fold CV of $69.47 \pm 5.23\%$ and LOSO-CV of $61.45 \pm 16.72\%$. Table 4.12 shows the summary of results for all the classifiers tested for negative affect.

Table 4.12: Summary of results from supervised learning to classify Negative Affect.

	10-CV Acc	LOSO-CV Acc
Logistic Regression	$66.80 \pm 4.64\%$	$37.10 \pm 17.37\%$
LDA	$67.79 \pm 4.46\%$	$41.54 \pm 16.14\%$
SVM	$69.47 \pm 5.23\%$	$61.45 \pm 16.72\%$
XGBoost	$69.50 \pm 8.23\%$	$60.88 \pm 19.89\%$

4.7 Discussion

This chapter describes the design, development and user-centric evaluation of *Running Chickens*, an AR mobile exergame. Additionally, player data is used to create models that predict several measures of PX. Robust predictive models of PX, along with an understanding of the relationship between the various game parameters on PX, would enable the creation of experience-driven adaptive mobile AR exergames.

The need for well-designed game progression systems is observed from the interviews with participants where most of them expect these types of games to increase in difficulty. However, some of them indicated that having some easy levels between challenging levels reassured their sense of competence in the game. This indicates that a simple linear progression of the game challenge may not be the most optimal strategy to create engaging AR exergames. Having an adaptive game that adjusts game parameters based on PX could lead to higher engagement with the game which is similar to the idea of flow in games. The adaptive system would try to find the optimal balance of the game's challenge and the player's skill to put the person in a state of flow.

4.7.1 Game Parameters impact on Player Experience

User evaluations of the game show that all 3 game parameters evaluated in this study have a meaningful impact on PX which is observed in both the statistical testing of the questionnaire data and analysis of the participant interview data. The area parameter had a significant impact on the challenge experienced by the player. Larger areas were more challenging which is observed from statistical analysis and interviews with players. People who consider themselves as physically active enjoyed larger areas while others preferred smaller areas. Increasing the areas between game levels would increase the challenge of the game. In games that are played over large areas, it is important to consider PX as they navigate through space. If the person is moving slowly, they can use the AR viewport for navigation, however, as their speed increases, they rely more on their peripheral vision which makes it difficult to focus on the AR world while running. This attention switching between peripheral vision and device viewport will also reduce the levels of immersion a user experiences. This finding is more relevant for mobile AR experiences however, this could be a factor in HMD AR experiences with a limited field of view.

The number of chickens was an important game parameter as it had a significant impact on the sense of Valence, Competence, Flow, Tension, Challenge and Positive Affect. While people felt less competent, more tensed, had a lower positive affect and were more challenged in games with a large number of chickens, the scores for flow increased as the number of chickens increased. People felt more inflow and more rewarded by increasing the number of chickens in levels. However, when a level was perceived as difficult, a high number of chickens could harm PX as people tend to feel overwhelmed. This explains why scores for tension and challenge increased, positive affect and competence decreased when a large number of chickens were used. This game parameter is particularly interesting since it has the potential for creating high player engagement and increased immersion or flow, tunning this parameter could result in positive PX. The challenge for game designers and adaptive exergame systems is to modulate this parameter across a number of game levels within a single play session to ensure the player's experience is maintained in a flow channel (where the number of chickens is kept large enough to keep players engaged without being too high which would result in players being overwhelmed, frustrated and eventually disengaging with the game).

Similarly, the evasion mechanic, which is proposed in this chapter was found to have a significant effect on several dimensions of PX. Levels with evading chickens resulted in significantly higher arousal, immersion, flow, tension and challenge scores. In some cases, it also leads to lower competence scores. An optimal selection of the evasion speed will ensure higher immersion in the game however if it is too fast players would feel overwhelmed and conversely bored if it is too slow. It is interesting to note that all the participants enjoyed chasing after evading chickens irrespective of their preferences for physical activity. This shows that participants enjoy physically interacting with digital objects (in this case chickens) which is similar to observations made by Reimann[184] who investigated an AR footballs game's game interactions. Considering these results, it is possible that users enjoy embodied digital interactions with digital content that mirrors physical objects in the real world, e.g. kicking a football or catch a chicken that tries to evade the player. These observations show that the evasion mechanic can be useful to other AR exergames to motivate physical activity. It would be interesting to explore to what extent this motivation and positive engagement would sustain over time across many play sessions since it is possible that the novelty of these interactions will reduce over time.

The number of chickens and the evasion mechanic are important parameters for adaptive AR games since they can be tuned for optimal PX. While they both can have a meaningful impact on a person's immersion and state of flow, improper selection of this parameter can result in a negative experience for the player. Increasing the area increases the level of challenge in the game however this is often constraint by the availability of physical space, the other two parameters can be used to create a wide range of levels within the same area.

4.7.2 Modelling Player Experience

Interestingly, the proposed modelling technique can predict dimensions of PX considerably better than random chance. Valence, Positive and Negative Affect were classifiers that performed poorly (close to random chance) indicating that positive ad negative emotions are difficult to model. While other PX dimensions relevant to games such as Flow, Immersion, Competence, Tension and Challenge perform considerably better. This shows that some aspects of PX can be successfully modelled using supervised learning. Among these classifiers competence, challenge and tension showed the highest and most reliable performance. The results of challenge and tension are similar to the results for challenge and frustration from the previous study. It is interesting to note that the experience of competence can be successfully modelled within MAR game environments.

All predictive models developed in this experiment show a high variance in the sample LOSO-CV scores, this indicates that model performance is different for different participants. While PX models perform well for some participants, the same models perform close to random chance (or worse) for others. This can be improved by using a more generalisable dataset to train these models. Additionally, since PX is highly subjective and varies greatly depending on the individual, it would be useful to create individual models that are trained on data from a single person, instead of inter-subject models that are trained using data from several participants (which is the case in this study). However, individual PX models require a large amount of data from a single participant to train them.

Among the different supervised learning classifiers tested, the XGBoost classifier had the highest 10-fold CV and LOSO-CV accuracies across all dimensions of PX tested in this study. For some PX dimensions, namely for competence, flow and challenge the 10-fold accuracies of the XGBoost classifier are outperformed by other classifiers tested (logistic regression, LDA or SVM), however, XGboost remains the best performing classifier when LOSO-CV accuracies are considered. This indicates that XGBoost shows the best generalisability for unseen players. The only exception to this pattern appears in the evaluation of the negative affect model, which showed that the SVM was the more reliable classifier, however, the SVM's LOSO-CV scores were only marginally higher than the scores for the XGBoost model.

Among the other classifiers tested, logistic regression and LDA show comparatively poor performance in both 10-fold CV and LOSO-CV scores. The SVM classifier tended to perform better in most cases, however, there is a significant performance drop between 10-fold CV and LOSO-CV scores. This is especially observed in results for the immersion where the SVM classifier showed 10-fold CV scores of $\approx 70\%$ (higher than random chance) and LOSO-CV scores dropping to $\approx 40\%$ (lower than random chance).

Findings from this user study will be used to create an adaptive version of Running Chickens. The adaptive game would adjust the parameter selection of the next level based on predicted PX from the previous level. Consider an adaptive system that is driven by predictions from the classifier for competence. The adaptive system would adjust the challenge of the next level to induce a state of flow and avoid overwhelming (or underwhelming) the player. For instance, if the classifiers predict high competence it would mean the player is bored and the game should increase the challenge level (by increasing the area, number of chickens or the evasion mechanic). Similarly, adaptive systems could be developed using models for player flow (to ensure the player is in a state of flow within the levels) or tension (to avoid high levels of tension or frustration during gameplay).

For experience-based adaptive systems, it is important to know that models do not have a high ability to generalise to unseen players, which is indicated by the low Loso-CV scores for most models tested. Considering this requirement, the results of this study suggest that the competence model using the XGBoost classifier shows promising performance that can be applied to real-time game adaptation. This is further explored in the next chapter.

4.7.3 Study Limitations

This study has several limitations and gaps that must be addressed. Since data from a limited sample of 40 university students have been used in this study, it is difficult to generalise findings to other demographics of players. Additionally, the supervised learning techniques can show a problem of model bias when a smaller dataset is used. Another challenge of this research is that these types of games are designed to be played outdoors where several factors like climate, quality of the ground, time of day and surroundings can significantly change PX and add noise to the training data for the predictive models. It is interesting to note that the supervised models can predict PX higher than random chance despite the limited dataset used. Using a larger dataset from players of a wider demographic and varied environments would increase prediction performance and improve the generalisability of findings. Future work could also explore using environmental factors as features for the learning algorithm to consider when predicting PX. Finally, recent research in HCI has found some problems with the validity of the internal structure of the GEQ[94, 30, 113]. These studies use a confirmatory factor analysis which does not find any evidence for the original 7-factor structure proposed for the GEQ and that it appears to have unstable results. This limitation is addressed to some extent in this study by providing scores for Cronbach's α for each factor of the GEQ to give the reader a measure of the reliability of the factor following guidelines from Law et al. [113]. Additionally, the next chapter reports a study that uses a different questionnaire, the Player Experience Inventory[220, 6] which has been comprehensively validated by the authors of the scale.

4.8 Chapter Summary

The potential for AR mobiles games to promote physical activity and positive mental health is an emerging field of research, with this chapter contributing to this body of work. Further investigation into this domain would enable the creation of intelligent AR mobile games that have a positive behavioural impact on players. AR exergames is not a new phenomenon however, there is a gap in empirical evaluations of commonly used mechanics in these games. To address this gap, this chapter describes the design, development and evaluation of the Running Chickens game. Additionally, the study shows that player behaviour, measured through mobile sensor data, can be used to predict several dimensions (Flow, Immersion, Competence, Tension and Challenge) of player experience higher than random chance classification. Robust PX models can be used to drive adaptive AR exergames which would be more engaging for players. Mobile AR Game designers are creating immersive digital experiences that are enjoyed by a large number of players, resulting in a positive impact on their quality of life. Research into adaptive AR exergames will allow designers to use the AR platform to create positive behavioural changes in their players through intentional design, rather than as a consequence of gameplay. The next chapter further investigates PX in the Running Chickens game by exploring the impact of game parameters on player motivations as well as the extent to which player motivations can be modelled using supervised learning techniques. While this chapter has investigated player models in mobile exergames for general game experiences using the GEQ, the next chapter will overcome the limitations of the GEQ discussed in this study and investigate player models in this domain when applied to understand people's motivations for gameplay. This will further help mobile AR exergame designers create games that can motivate people towards positive behavioural change. These models built in the next chapter are further evaluated in the context of real-time DDA in games. This is important since it will show to what extent these models will generalise well to new or unseen players and show the validity of using these models in real-time game applications to create an optimal PX for the people.

Chapter 5

Experience-based Difficulty Adaptation in an AR Exergame

The previous chapter presented the second study of this research which investigated supervised learning to predict player experience (PX) and the impact of game parameters on PX in an augmented reality exergame: Running Chickens. The study reported in this chapter consists of 2 experiments that build on the previous study. While the research chapters presented until now (chapters 3 and 4) have described a single experiment per chapter, this chapter combines 2 experiments that build and evaluate PX models in MAR games for DDA. The research activities described here explore to what extent game parameters can impact PX when analysed using the theoretical framework of player motivations. In the first study, the supervised learning pipeline developed in previous studies is used to explore to what extent these models can predict player experience dimensions related to their motivations within AR exergames. In the second study, these PX models are evaluated in the context of dynamic difficulty adjustment (DDA) to assess to what extent these models can be used to personalise the game experience for players.

The validity of PX models must be evaluated based on the objective for which it is made (i.e. is the model meant for DDA, player experience analysis or PCG), both experiments are required to evaluate the suitability of applying these PX models for DDA (which is integral to answering RQ1 and RQ3). Additionally, this thesis focuses on PX modelling so the second experiment of this chapter can be considered as a validation exercise for the models built in experiment 1 of the chapter. Additionally, as argued, DDA in MAR games is a domain that requires much further investigation (which is out of scope for this thesis), the application of MAR DDA presented in this chapter is a simple application of DDA that is built to test the validity of the PX models. Hence it is presented as a second experiment within this chapter as opposed to a separate one.

In the first experiment of this chapter, a user study is conducted using the AR Exergame: Running chickens (the design and development process of the game is described in the previous chapter). In this experiment PX constructs of interest/enjoyment, mastery, autonomy and immersion are used as experience metrics. The user study explores how different game parameters (that generate the different levels of the game) can impact these experience metrics. Additionally, the data collected from this study is used to build and evaluate supervised learning models that predict each of these PX constructs. The experiment method used is similar to the one used in the previous study (reported in chapter 5). Unlike the previous study which focused on general game experience, experiment 1 of this chapter uses PX constructs that are related to motivation and behaviour research. The goal of this experiment is to investigate to what extent these PX constructs can be used in an affective game loop to sustain a player's motivation across their play session and provide an overall positive game experience.

The second experiment reported in this chapter explores to what extent supervised learning models built in the first experiment can be used for game adaptation. The goal of this adaptation is to put players in an optimal affective state and maximise their engagement with the game. The adaptation described here is determined by a decision system that changes the difficulty of the game based on the predicted level of mastery. Predicting the player's perceived mastery for a given level is formulated as a binary classification problem. A supervised learning classification algorithm is trained to predict either high or low mastery from a number of features extracted from mobile sensors and game data for a level. In this experiment, a user study is conducted to empirically compare the player experience of this adaptive game against a non-adaptive version of Running Chickens.

Across the two studies, this chapter investigates to what extent player experience models can be used for real-time game adaptation in MAR exertion games. This is important for personalizing exertion based games since appropriate balancing of game parameters can lead to increased internalisation of game objectives. This would keep players engaged in the game for longer, leading to increased physical activity through engagement with the game.

This chapter is structured as follows: the aims and motivations for this chapter are described in section 5.1. Section 5.2 provides details on the AR exergame game used in both experiments of this chapter. Section 5.3 describes the experiment design, analysis and results for the first experiment of this chapter. The data from this experiment is used to build an experience-based DDA system implemented within the Running Chickens game, this process is described in Section 5.4. This adaptation system is evaluated in the second experiment which is described in section 5.5. Section 5.6 discusses the implications of the results and limitations of the experiments conducted in this chapter. Finally, section 5.7 presents a chapter summary.

5.1 Aims and Motivations

The two experiments reported in this chapter explored to what extent PX constructs related to player motivation can be predicted from player behaviour and game data. Additionally, the potential of these models for experience-driven DDA is evaluated. To further investigate player modelling in AR games, this chapter builds on the results from the study reported in the previous chapter. While the previous study explored the impact of player experience using the GEQ [86], it was discussed that this measure has some limitations where validations of this questionnaire could not support the internal factor structure proposed by the authors of this measurement tool [94, 30, 113].

One of the main aims of this chapter is to overcome this limitation by using PX measurements that have been comprehensively validated by other researchers. Additionally, this chapter focuses on PX dimensions that have been identified as significant to player motivations. This aim is addressed in the first experiment reported in this chapter but using PX measures that have been built from the Player Experience Inventory (PXI) which is a tool that has been developed over a number of rounds of development and validation using 64 game user researchers and population of 529 participants [6]. This experiment focuses on the following dimensions of player motivation (which have been taken from the PXI): Interest/Enjoyment, Mastery, Autonomy and Immersion. The motivations for using each of these PX dimensions are described as follows:

- **Interest/Enjoyment** with the game is one of the main factors of intrinsic motivation, where player behaviour (or in this case, engagement with the game) is satisfying internalised objectives and rewards [136]. This study explores to what extent game parameters can be used to facilitate intrinsic motivation through the enjoyment of engaging with the game.
- **Mastery** of the game is an important factor for motivation and is related to several theories that describe behaviour, motivations and player experience. Self-Determination Theory (SDT) [181, 188] identifies competence (which is similar to mastery) as an important factor that can describe a

person's motivation for a task. It states that when a person's capability to engage with a task matches their existing skills their competence needs are satisfied. This allows them to enhance their skills and expertise. When competence is frustrated it can lead to an experience of failure and helplessness. Mastery is also related to a person's sense of self-efficacy and confidence for the task. Self-efficacy can be understood as a person's belief in their ability to succeed in a particular situation [15] which is a significant mediator of PX in exergames [126]. While confidence is an important component of the ARCS (Attention, Relevance, Confidence and Satisfaction) model that is a popular framework used for motivational design [104, 223] and games for learning [75, 245]. Finally, mastery is important in the theory of flow [43, 36], which describes the importance of a good balance between games challenge and player skills to push players towards an optimal psychological state of flow.

- **Autonomy** which is an important construct identified by SDT, refers to the experience of volition and willingness [181, 188]. It is satisfied when a person's actions within a task are perceived as self-endorsed or self-directed. When autonomy is frustrated it results in the experience of pressure, conflict or being pushed by the activity in an undesired direction. This construct is investigated here since players of the Running Chickens game actively make many choices about how to physically move around and engage with the game levels. It is therefore important to analyse how game parameters can satisfy or frustrate participants autonomy need when playing the game.

- **Immersion** is an important aspect of games, it is the engagement or involvement a person feels through game-play [93, 180]. Immersion has been extensively studied in the context of digital games, however, there is limited understanding of how immersion is mediated in AR games that involve physical activity[200]. While the previous study explored immersion using the GEQ, experiment 1 of this chapter further builds on this by using a different questionnaire to validate these results.

While the PX constructs described above have been used as descriptive measures to understand game experiences, there is limited research on how these dimensions can be predicted using player models in movement-based MAR games. Therefore another aim of this chapter is to bridge this gap by applying supervised learning classification algorithms to predict them.

These player models can be used to adapt game levels based on PX as the player engages with the game. Therefore, the final aim of this chapter is to evaluate to what extent these predictive models can be used to meaningfully adapt player experience. This aim is addressed in the second experiment described in this chapter.

Player motivations in exergames have been investigated by previous researchers taking a case study approach which compares player motivations between different people[126] or using competition to improve motivations [155]. However, further investigation into the impact of game parameters on player motivation would enable the design of optimal exertion trajectories [130], which is an important design consideration of successful exergames. Additionally, predictive models of player motivations would allow for the creations of personalised trajectories that maximise players' motivations for gameplay which would ideally lead to increased physical benefit from the exergame activity.

5.2 The AR Exergame: Running Chickens

This section provides details about the game used in the two experiments described in this chapter. The *Running Chickens* game is the same as the one used in the previous chapter (refer to section 5.2 for a full description of the game). The version of the game used in this chapter had some usability modifications made to it in order to improve the performance of the AR tracking algorithm. It was observed in previous the previous chapter that participants would perform a "hammering" action (jerk the phone up and down) when they were trying to capture the chickens in a level. This hammering action would sometimes lead to loss of tracking from the AR algorithm resulting in a game crash. This was the case especially if the participant was unsuccessful at capturing the chicken causing them to continue this hammering gesture with the phone.

In order to solve this problem, the interaction for collecting a chicken was modified to make capturing the chickens a smoother experience for the player. Instead of calculating distance in 3-dimensional space (x, y and z axes) between the mobile device and a chicken to check for possible collisions, the newer versions of the game would calculate the distance in 2-dimensional space using only the x and z axes (ignoring the y axis). This reduced the need for participants to perform the hammering action with the mobile in order to collect the chickens. Which reduced the vertical jerking motion of the mobile device and resulted in more stable tracking by the AR algorithm. This modified mechanic was tested with two pilot participants to confirm that the mechanic resulted in a more stable game. The other aspects of the game were the same as the version used in the previous chapter. This updated Running Chickens game was used in the remaining experiments conducted as part of this research.

5.3 Experiment 1

This section describes the first experiment conducted in this chapter. The design of this experiment is similar to the one used in the previous study however, different questionnaires were used to collect the player experience data. In this experiment, semi-structured interviews were not conducted since qualitative data from the previous study showed "knowledge saturation"[21]. So it is unlikely to find out any new information by conducting additional interviews, especially when working with a relatively homogeneous sample of university students.

5.3.1 Experiment Design

The experiment followed a within-participant design where a participant faces all conditions of the study, a study condition refers to each level of the Running Chickens game. The study uses the same 3 game parameters (as binary categorical variables) from the previous study which resulted in 8 ($2 \times 2 \times 2$) game levels (or study conditions). The 3 game parameters used are:

- *The Area of the Level (G_{Area}):* 2 sizes of levels are compared. *Large Area (Lrg)* levels are $30m \times 30m$ and *Small Area (Sml)* levels are $15m \times 15m$.
- *Number of Chickens (G_{No}):* 2 amounts of chicken in each level are compared: *Low Number* with 5 and *High Number* with 20 pieces respectively.
- *Chicken Movement (G_{Mov}):* 2 settings of the chicken evasion mechanic are selected for evaluation: *Non-Evasive* condition where the chickens do not evade players, *Evasive* condition where the chickens evade players. The speed of chickens in the evasive condition is the same as in the previous study.

The study consisted of 9 sessions (1 training + 8 experiment), each experiment session took approximately 2 min to complete while the training session took up to 10 min to complete. The training session involved a demonstration by the researcher of the game and interactions. Which was followed by the participant practising the game for 2 levels. For training, a simple level is used with 3 chickens that do not evade the player in small a $5m \times 5m$ area (data from the training session is not used in analysis). In each experiment session, participants played a level of Running Chickens and rated a PX questionnaire.

All experiments were conducted during daylight and adequate weather conditions (no signs of rainfall) in a park near Queen Mary University of London's Mile End campus. This is done to minimise the difference in results that may arise due to different locations or poor lighting and weather conditions. The order in which each participant experienced each session of the experiment was randomised to minimise ordering effects on the data collected. Participants were anonymised using IDs and were compensated £10.00 for their participation in the study. The experiment was conducted using a Google Pixel 3 mobile device.

Procedure

All participants provided informed consent before participating in the study. At the beginning of the study, participants filled up a questionnaire about their background and previous experience in MAR games. After which they were given a training session about the game and the questionnaires used in the study. During training, the researcher first demonstrated how the game works to the participant over one level of the game, after which participants played 2 training levels and filled up the PX questionnaire after each. During training, the researcher was present with them and they were encouraged to ask any questions about the game or the study procedure.

Once the training was completed, the participants were left alone in the park (while the researcher waited by the entrance of the area), to minimise the effect of the researcher's presence of the data collected. During this time, participants experienced the 8 study sessions of the experiment. In each session, the participant played one of the game levels and filled up the PX questionnaire. Participants were asked to take a 2-5 min break between sessions to minimise the effects of physical fatigue from the previous session on the data collected from the next one.

While playing Running Chickens, if the participants experienced any technical issues (the main one being the AR algorithm losing tracking of the environment), they were asked to proceed to the next session and the data from this session was not used in data analysis. At the end of the study, participants were debriefed about the objectives of the research, all the questions were answered and the experiment was concluded. The experiment took 40-60 min for each participant (depending on the length of the breaks they took during the study).

5.3.2 Data Collection

This study used the same data collection methods employed in the previous experiment (qualitative data from semi-structured interviews was not collected). During the study, PX, player movement data (measured through the mobile sensors) and game metrics were collected. As each game level was a study condition, each participant contributed 8 games to the data set. This resulted in a total of 200 games played across the study. However, due to some technical crashes, only 185 games were recorded and used in data analysis.

The data set collected is used to analyse the impact of game parameters on PX. Additionally, it is also used to train and evaluate supervised learning classification algorithms to predict PX from the mobile sensor data and game metrics. This section describes the questionnaires and sensors used as well as the demographic information of the sample of participants in this study.

Participants

Participants were recruited using university mailing lists which included PhD, Masters and Undergraduate students from the Electronic Engineering and Computer Science at Queen Mary University of London. The study sample consisted of 25 volunteers (7 female and 18 male) aged between 18-44 (8 participants were aged between 18-24, 11 were 25-29, 6 were above 30) who took part in this study (summarized in table 5.1). When asked about prior experience playing AR games 14 subjects had no prior experience. In the remaining 11 subjects: 5 reported having only one experience in the past, 5 played a few times before, and 1 participant played AR games regularly (summarised in table 5.2). It is worth noting that the study aimed to collect data from 40 participants however, due to the COVID 19 pandemic, data collection was stopped at 25 for safety concerns.

Table 5.1: Summary of participants' ages.

Age range	Number of participants
18-24	8
25-29	11
Above 30	6

Table 5.2: Summary of participants' previous experience with AR games.

Previous AR experience	Number of participants
No experience	14
Played only once before in the past	5
Played only a few times in the past	5
Played AR games regularly	1

Questionnaire Data

The study used ratings based questionnaires to collect PX data. The dimensions of PX measured in this study were Interest/Enjoyment, Mastery, Autonomy and Immersion. Interest/Enjoyment is measured using the subscale (with the same name) from the Intrinsic Motivation Inventory (IMI) [136]. The interest/enjoyment subscale consists of a 6 item questionnaire that uses a 7 point Likert scale providing their level of agreement with the different statements. Mastery, Autonomy and Immersion are measured using the corresponding subscales of the Player Experience Inventory (PXI) [220, 6] which consists of a 9 item questionnaire (3 for each) using a 7 point Likert scale as well. The complete versions of the IMI and the PXI were not used in order to minimise the possibility of questionnaire fatigue on the participants as they are required to fill in this questionnaire 10 times across the study (2 training + 8 experiment sessions).

Player Behaviour Data

Player behaviour in game sessions is measured through mobile sensor data which is similar to the previous study. This player movement data is the position and rotation of the mobile device during the game and is recorded at a frequency of $64Hz$ as a 7-D vector: $\alpha \in \{P_X, P_Y, P_Z, R_X, R_Y, R_Z, R_W\}$of positions (in meters) and rotations (in quaternions). Additionally, the timing and score data for each level is also recorded. Refer to section 4.5.3 in the previous chapter for additional details.

5.3.3 Analysis and Results

Similar to the previous study, the resulting data set from the experiment was used to understand the impact of different game parameters on PX and to explore the potential of supervised learning to predict dimensions of PX based on player movement and game-based data. The methods used are the same as the previous study. The data analysis and results have been described below in this section.

Analysis of Game Parameters influence on Player Experience

Similar to the previous study, the relationship between game parameters on PX is analysed using the three-way ANOVA with a p-value of 0.05. Residual analysis was performed to test for the assumptions of the three-way ANOVA. Normality within groups was assessed using QQplots. In some cases the normality assumption was not met however, as indicated by Blanca et. al [28] the ANOVA test is still a valid option for analysis. It is worth noting the Blanca et. al. proved this hypothesis using a Monte Carlo simulation study. The assumptions of homogeneity of variances was assessed by Levene's test. A posthoc Tukey test was also conducted if significant interaction effects were observed from the ANOVA test. The results of each of the PX dimensions are reported below:

Interest/Enjoyment: A good reliability was found for the interest/enjoyment scores, Cronbach's $\alpha = 0.91$. Residuals were not normally distributed ($p < 0.05$) and there was homogeneity of variances ($p > 0.05$). Inspection of the QQplot by experiment conditions showed most groups fall approximately along the reference line shown in figure 5.1, indicating that data within groups show minor deviations from normality. No significant main or interaction effects were observed from the game parameters on interest/enjoyment scores. Figure 5.2 shows the box plot of the Interest/Enjoyment scores for the experiment conditions, which shows that scores are similar across all conditions.

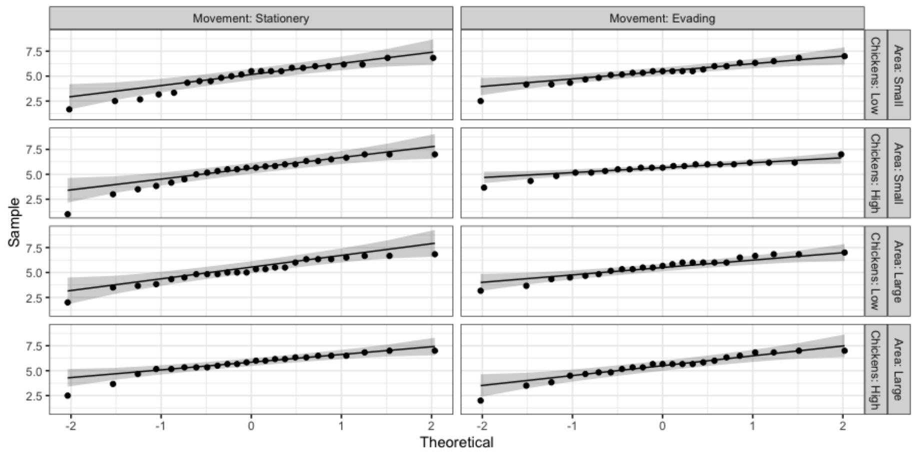


Figure 5.1: The figure shows the **QQplots for the Interest/Enjoyment scores** across the study conditions which show minor deviations from normality across the different conditions.

Mastery: A good reliability was found for the mastery scores, Cronbach's $\alpha = 0.85$. Residuals were not normally distributed ($p < 0.05$) and there was homogeneity of variances ($p > 0.05$). Inspection of the QQplot by experiment conditions showed most groups fall approximately along the reference line shown in figure 5.3, indicating that data within groups show minor deviations from normality. There was a significant main effect for the number of chickens, $F(1, 177) = 8.28$, $p = 0.004$ with an effect size $\eta^2_{partial} = 0.015$. There was also a significant main effect for the evasion mechanic, $F(1, 177) = 11.04$, $p = 0.001$ with an effect size $\eta^2_{partial} = 0.053$. Finally, there was a significant 2-way interaction effect was also observed between the number of chickens and the evasion mechanic, $F(1, 177) = 7.02$, $p = 0.008$ with an effect size $\eta^2_{partial} = 0.033$. Posthoc

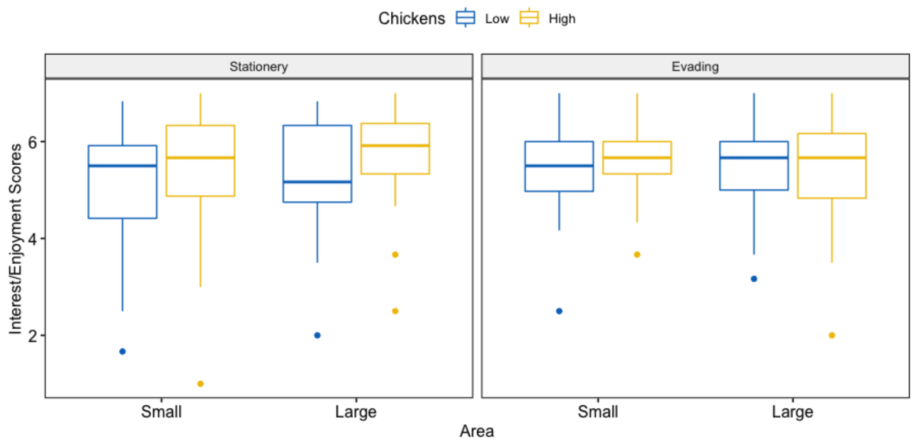


Figure 5.2: The figure shows the **boxplot plot for the Interest/Enjoyment scores** across the study conditions. The image shows that scores are similar across all conditions.

testing showed that mastery scores were significantly lower when the number of chickens was high as compared to when it was low ($p = 0.004$). Mastery scores were significantly lower when chickens evade players as compared to when they were stationary ($p = 0.001$). Levels with a high number of chickens that evade players have significantly lower mastery scores as compared to levels with a low number of chickens that are stationary ($p < 0.001$). In levels with a high number of chickens, mastery scores were lower when chickens evade players as compared to when they are stationary ($p < 0.001$). Finally, in levels where chickens evade players, a high number of chickens had significantly lower mastery scores as compared to a low number of chickens ($p < 0.001$). These effects are observed in figure 5.4, which shows the plot of the mastery scores across the study conditions.

Autonomy: A good reliability was found for the autonomy scores, Cronbach's $\alpha = 0.88$. Residuals were not normally distributed ($p < 0.05$) and there was homogeneity of variances ($p > 0.05$). Inspection of the QQplot by experiment conditions showed most groups fall approximately along the reference line shown in figure 5.5, indicating that data within groups show minor deviations from normality. No significant main or interaction effects were observed from the game parameters on autonomy scores. Figure 5.6 shows the box plot of the autonomy scores for the experiment conditions, which shows similar scores across all conditions.

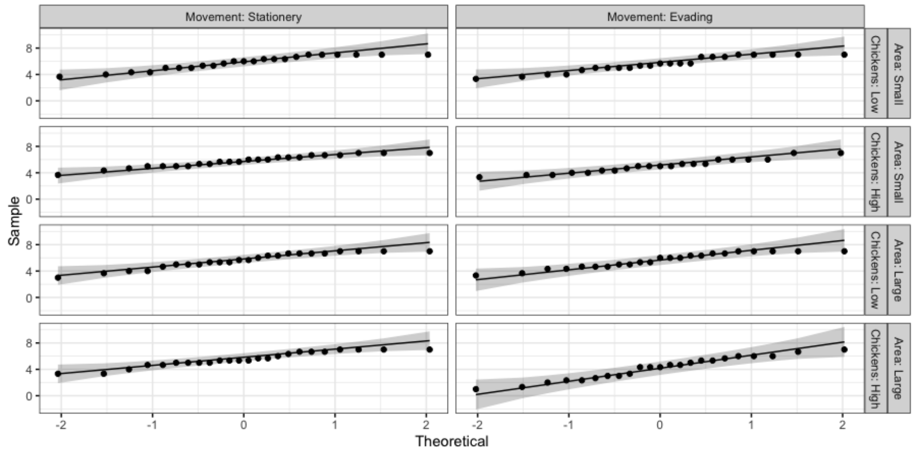


Figure 5.3: The figure shows the **QQplots for the Mastery scores** across the study conditions which show minor deviations from normality across the different conditions.

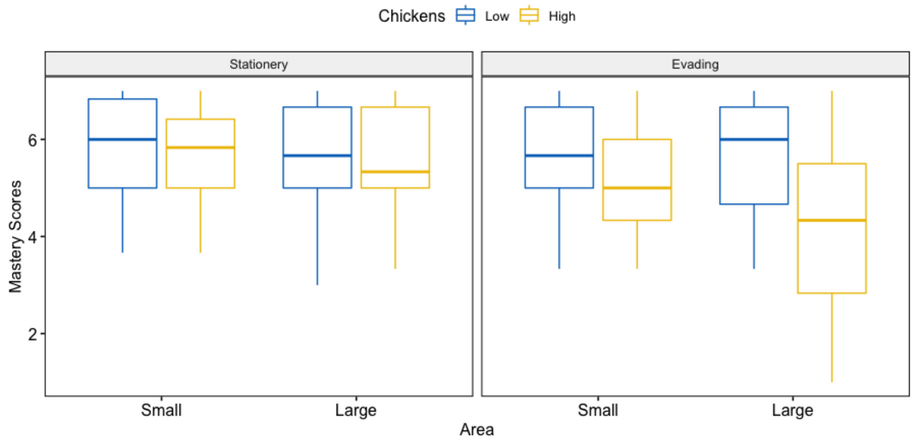


Figure 5.4: The figure shows the **boxplot plot for the Mastery scores** across the study conditions. The image shows that when the number of chickens is high mastery scores are lower. Additionally, when chickens evade players, mastery scores are lower. Levels with a high number of chickens that evade players have significantly lower mastery scores as compared to levels with a low number of chickens that are stationary. Finally, in levels where chickens evade players, a high number of chickens had significantly lower mastery scores as compared to a low number of chickens.

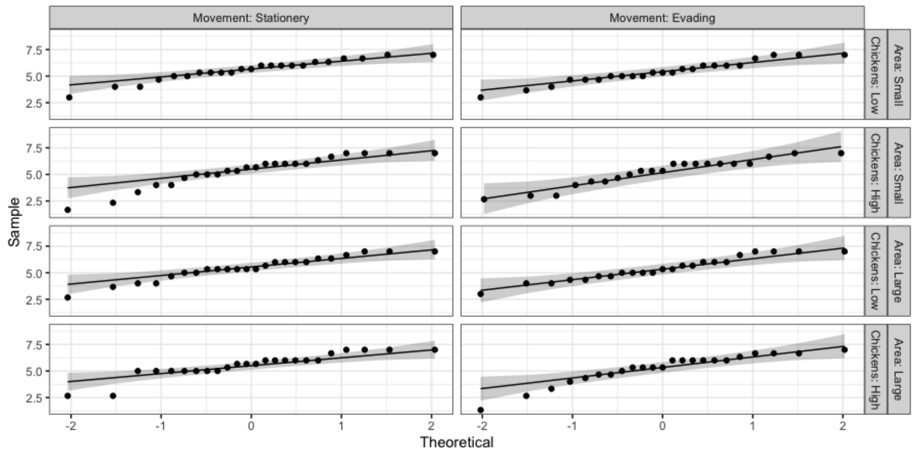


Figure 5.5: The figure shows the **QQplots for the Autonomy scores** across the study conditions which show minor deviations from normality across the different conditions.

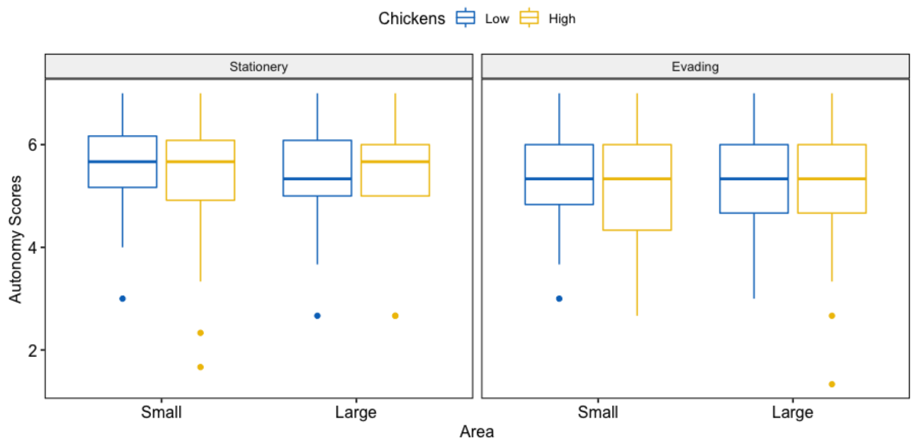


Figure 5.6: The figure shows the **boxplot plot for the Autonomy scores** across the study conditions. The image shows that scores are similar across all conditions.

Immersion: A moderate reliability was found for immersion scores, Cronbach's $\alpha = 0.61$. Residuals were not normally distributed ($p < 0.05$) and there was homogeneity of variances ($p > 0.05$). The QQplot by experiment conditions showed most groups fall approximately along the reference line shown in figure 5.7, indicating that data within groups show minor deviations from normality. A significant main effect for number of chickens, $F(1, 177) = 4.06$, $p = 0.045$ with effect size $\eta^2_{partial} = 0.022$. Posthoc testing showed that when number of chickens was high, immersion scores were significantly higher as compared to when number of chickens was low ($p = 0.045$). These effects are observed in figure 5.8, which shows the plot of the immersion scores across study conditions.

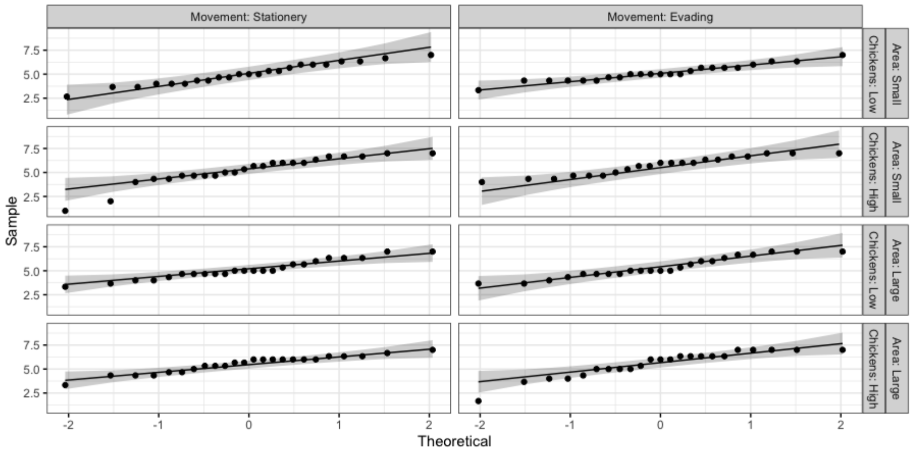


Figure 5.7: The figure shows the **QQplots for the Immersion scores** across the study conditions which show minor deviations from normality across the different conditions.

This subsection reported the results of statistical analysis of the impact of CGFs on player motivations. Table 5.3 summarises the results from the statistical analysis.

Player Experience Classification using Supervised Learning

This subsection presents results from supervised learning techniques. The same supervised learning pipeline is used to predict the dimensions of PX. Details on the complete pipeline are provided in subsection 4.6.3. Predictions of the different dimensions of PX is treated as a supervised classification problem by transforming the ratings based measures into binary categorical data through the player-based transformation described in the previous chapter. Using the data-processing and feature extraction methods described previously 65 features are extracted for each game. These features are used as an input for logistic

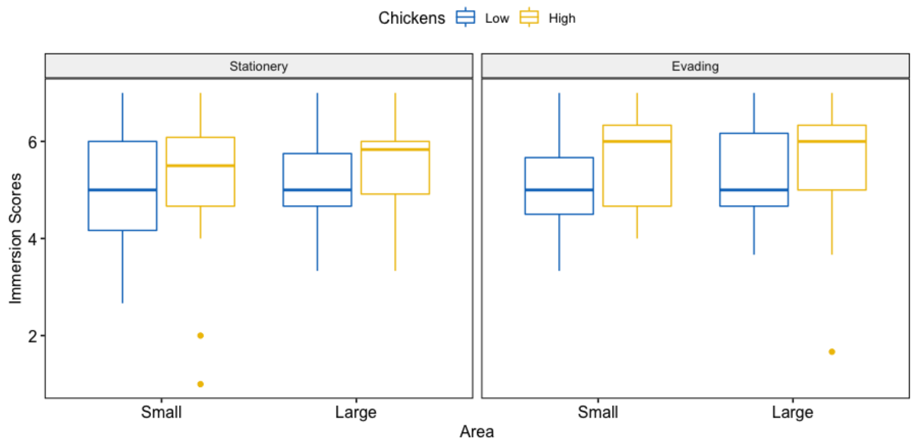


Figure 5.8: The figure shows the **boxplot plot for the Immersion scores** across the study conditions. The image shows that when the number of chickens is high immersion scores are higher.

Table 5.3: Summary of results of statistical analysis of CGFs on PX.

PX dimension	CGFs	F-value	$\eta^2_{partial}$	Effect Description
Interest/Enjoyment	No main or interaction effects			
Mastery	G_{Num} (main effect)	8.28	0.015	Decreases when number is high.
	G_{Mov} (main effect)	11.04	0.053	Decreases when chickens evade players.
	$G_{No} \times G_{Mov}$ (Interaction effect)	7.02	0.033	Decreases when number is high and chickens evade players as compared to levels with a low number of chickens that are stationary. Decreases when number is high and chickens evade players as compared to levels with a high number of chickens that are stationary. Decreases in levels where number is high and chickens evade players as compared to levels where numbers are low and chickens evade players.
Autonomy	No main or interaction effects			
Immersion	G_{Num} (main effect)	4.06	0.022	Increases when number is high

regression, LDA, SVM and XGBoost models to predict the different dimensions of PX. In this study, Logistic Regressions and LDA are linear models, while the SVM[1] and XGBoost classifiers are nonlinear models.

Feature selection is performed using the SFFS algorithm to improve model performance. Similar to the previous study, model performance improves with feature selection. The 10-fold CV and Loso-CV scores are used as evaluation metrics for each model. All the accuracies are reported for the models after feature selection. Results for each of the dimensions of player motivations are reported below.

Interest/Enjoyment: Random chance classification for interest/enjoyment was 50.05%. All classifiers perform higher than random chance on 10-fold CV accuracies. However, accuracies for Logistic Regression, LDA and SVM classifiers decrease to close to random chance when LOSO-CV accuracies are computed. The XGBoost classifier was the best performing model tested with a 10-fold CV of $68.12 \pm 10.09\%$ and LOSO-CV of $62.27 \pm 21.06\%$ which is higher than random chance classifications. Table 5.4 shows the summary of results for all classifiers tested for Interest/Enjoyment.

Table 5.4: Summary of results from supervised learning to classify Interest/Enjoyment.

	10-CV Acc	LOSO-CV Acc
Logistic Regression	$62.63 \pm 13.24\%$	$56.12 \pm 9.99\%$
LDA	$64.35 \pm 9.26\%$	$55.15 \pm 10.21\%$
SVM	$67.54 \pm 7.70\%$	$44.84 \pm 8.84\%$
XGBoost	$68.12 \pm 10.09\%$	$62.27 \pm 21.06\%$

Mastery: Random chance classification for interest/enjoyment was 51.78%. All classifiers perform higher than random chance on 10-fold CV and LOSO-CV accuracies. The XGBoost classifier was the best performing model tested with a 10-fold CV of $81.72 \pm 10.63\%$ and LOSO-CV of $71.65 \pm 19.98\%$ which is higher than random chance classifications. Table 5.5 shows the summary of results for all classifiers tested for Mastery.

Autonomy: Random chance classification for autonomy was 53.50%. All classifiers perform higher than random chance on 10-fold CV accuracies. However, accuracies for the Logistic Regression classifier decreases to close to random

[1] The SVM classifier in this study uses an RBF kernel so it is a nonlinear classifier in this study.

Table 5.5: Summary of results from supervised learning to classify Mastery.

	10-CV Acc	LOSO-CV Acc
Logistic Regression	$73.59 \pm 7.79\%$	$67.21 \pm 11.18\%$
LDA	$73.09 \pm 10.46\%$	$61.83 \pm 12.61\%$
SVM	$75.78 \pm 9.90\%$	$59.22 \pm 12.46\%$
XGBoost	$81.72 \pm 10.63\%$	$71.65 \pm 19.98\%$

chance when LOSO-CV accuracies are computed. Accuracies for LDA, SVM and XGBoost perform higher than random chance for LOSO-CV. The SVM classifier was the best performing model tested with a 10-fold CV of $71.34 \pm 7.17\%$ and LOSO-CV of $63.02 \pm 15.51\%$. Table 5.6 shows the summary of results for all classifiers tested for Autonomy.

Table 5.6: Summary of results from supervised learning to classify Autonomy.

	10-CV Acc	LOSO-CV Acc
Logistic Regression	$64.94 \pm 6.91\%$	$52.38 \pm 15.61\%$
LDA	$68.07 \pm 9.86\%$	$61.97 \pm 16.72\%$
SVM	$71.34 \pm 7.17\%$	$63.02 \pm 15.51\%$
XGBoost	$68.15 \pm 9.35\%$	$64.85 \pm 14.95\%$

Immersion: Random chance classification for immersion was 52.95%. All classifiers perform higher than random chance on 10-fold CV accuracies. However, accuracies for Logistic Regression and LDA classifiers decrease to close to random chance when LOSO-CV accuracies are computed, while the SVM and XGBoost models perform higher than random chance for the same. The XGBoost classifier was the best performing model tested with a 10-fold CV of $72.51 \pm 9.07\%$ and LOSO-CV of $69.81 \pm 12.85\%$. Table 5.7 shows the summary of results for all classifiers tested for Immersion.

Table 5.7: Summary of results from supervised learning to classify Immersion.

	10-CV Acc	LOSO-CV Acc
Logistic Regression	$72.92 \pm 9.41\%$	$52.45 \pm 18.22\%$
LDA	$71.34 \pm 10.28\%$	$55.75 \pm 16.85\%$
SVM	$71.81 \pm 8.09\%$	$62.59 \pm 13.87\%$
XGBoost	$72.51 \pm 9.07\%$	$69.81 \pm 12.85\%$

Results from the analysis of different supervised learning models for the prediction of the different dimensions of PX. Results show that linear classifiers (logistic regressions and LDA) do not generalise well to unseen participants. Among the nonlinear classifiers tested, the XGBoost classifier performs the best showing a good generalisability to unseen participants. Mastery was the best performing model as it could be predicted to a reasonably high level of accuracy followed by immersion and then autonomy. Interest/Enjoyment was the most difficult to dimension to predict reliably.

5.3.4 Experiment Summary

The experiment described in this section builds on the previous study to investigate modelling PX within AR exergames as well as explore the relationship between AR game parameters and dimensions of PX. The results show that mastery is an important dimension of PX since predictive models created using the XGBoost classifier perform reliably and generalise to new players. Another important finding in this chapter is that exergame parameters of the Running Chickens game such as the number of chickens and the evasion mechanic are important features for adaptation as they can influence the levels of mastery and immersion players experience within the game. The results observed here will allow for the creation of personalised adaptive AR exergames that use the player models created to maximise a player's motivations as they play the game. However, it remains unclear to what extent these models can be used successfully for real-time game adaptation. To address this gap, the mastery model (using the XGBoost classifier) is used to create a dynamic difficulty adaptive version of the Running Chickens game. This adaptive game is presented section 5.4 and the experimental evaluation of this adaptive game is described in section 5.5 of this chapter. The development details of the game adaptation system are presented in the next sections.

5.4 Game Adaptation System

This section documents the design and implementation of the adaptation engine built using the mastery classifier. The mastery model was selected for this as the previous experiment found that it could be reliable predicted using movement and game features. The choice of using the mastery model is further justified by exiting research in PX in games, it is am important component of SDT [181, 188], self-efficacy [15, 126], the ARCS model [104, 223] and flow in games [43, 36] (refer to section 5.1 for details). The mastery classifier uses player movement and game metrics to predict if the player felt a high or a low sense of mastery.

Data from the previous experiment is used to train the mastery classifier offline. As a player engages with each level of the game, data from that level is used by the model to return a prediction for mastery. The prediction is used by the game adaptation system to modify the difficulty of the game. The data analytics pipeline for mastery prediction is described in the next subsection. The adaptation system, described in the subsection below, aims to sustain the player in an optimal affective state in order to improve the game engagement. Figure 5.9 shows an overview of the data pipeline that is used for mastery predictions.

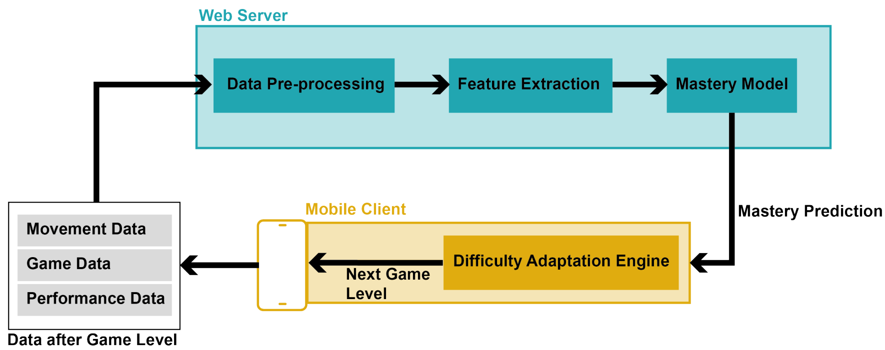


Figure 5.9: The figure shows the overview of the adaption pipeline for Mastery-based dynamic difficulty adaptation

5.4.1 Analytic Pipeline for Mastery Prediction

The analytics pipeline used for mastery prediction is developed using the XG-Boost classifier evaluated in the previous study. The model uses 26 features that are extracted from player movement data, game performance data and game parameters to return a binary prediction of either high or low mastery. Only 26 (out of the 65 features) are used for mastery prediction as these are the main features that be identified through feature selection using the SFFS algorithm. A benefit of using the XGBoost model (and other ensembles of decision tree models) is that it provides estimates about the importance of the features the trained model uses for prediction. Figure 5.10 shows the graph showing the importance of the different features used for prediction. It is observed that the model weights player movement features as most important for mastery prediction (the time of maximum acceleration and time of maximum jitter is weighted as most important). The model weights player performance (measured through the completion feature which is the ratio between the game score and the maximum possible score from the game) as moderately important.

Game parameters (number of chickens and the evasion mechanic) are weighted as least important for mastery prediction.

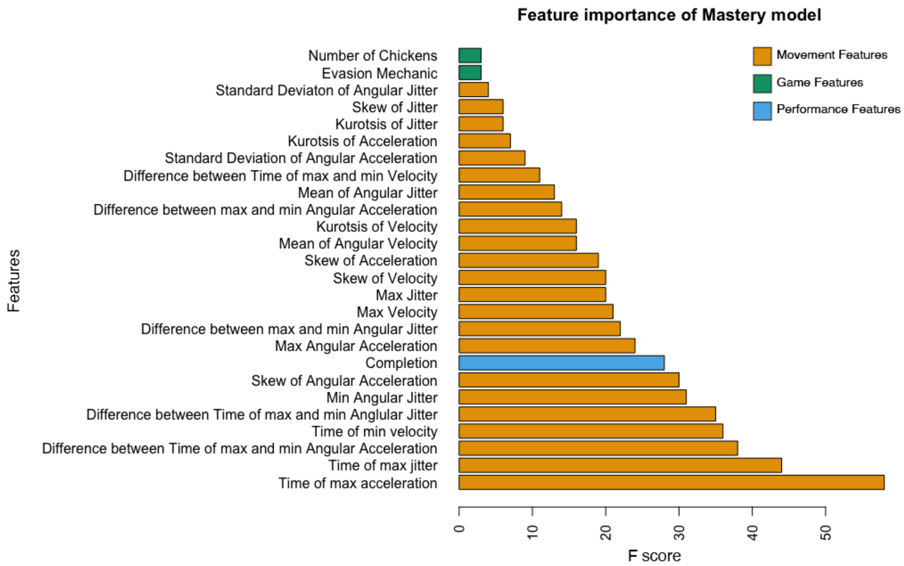


Figure 5.10: The figure shows the Feature Importance scores for the XGBoost Mastery Model

The goal of the model is to return a mastery prediction for each level of the game so that the difficulty of the next game level can be optimised to ensure the player is sustained in an optimal state of high mastery.

As the player engages with a game level, the position and rotation of the mobile device are logged at a frequency of 64 Hz. At the end of the game level the movement data, performance metrics (score and game time) and game parameters are sent to a web server for mastery predictions. The web server handles data preprocessing, feature extraction and mastery prediction. The method for data preprocessing and feature extraction is the same as the one described in the previous chapter. The main difference is that the webserver applied this data pipeline for a single game instead of a batch of games as was the case in the previous studies. Once the 27 features are extracted, it is used as input to the trained mastery model which returns a mastery prediction that is sent back to the game client running on the mobile device. The model was trained offline using data collected from experiment 1 (section 5.3) of this chapter.

This prediction pipeline has been implemented in python as a rest API using the Flask library. This web application is hosted on a Heroku server. The data sent to the server as well as the predictions generated are saved from the server for analysis of the model. Testing the prediction pipeline showed that the service takes 0.72 ± 0.64 seconds to return a prediction from the server. The relatively high variation observed in the response time is mainly due to the speed of the internet connection of the mobile device. Since this is an outdoor game, players would likely use their mobile internet which differs between service providers. The mastery prediction is used by the game client running on the mobile device as input for the adaptation engine which is described in the next subsection.

5.4.2 Difficulty Adaptation System

The adaptation system uses the predicted mastery from the model for experience-based DDA. In order to create an adaptive game, a subset of the game levels are used such that these levels form a linear difficulty progression. For this 4 (out of the 8) levels are selected where the difficulty of the game moves from an easy level to successively more difficult ones. Data from the experiment reported in chapter 5 is used to select this subset of levels since game challenge is one of the dependent variables analysed in this study. The game levels used by the adaption system are:

- **Level 1:** This level has a small game area with a low number of chickens that do not evade players.
- **Level 2:** This level has a large game area with a low number of chickens that do not evade players.
- **Level 3:** This level has a large area with a low number of chickens that evade players.
- **Level 4:** This level has a large game area with a high number of chickens that evade players.

The difficulty of the game levels moves from easy to difficult as a player moves from level 1-4 in the subset of game levels. This is observed from the Challenge scores (which is one of the subscales of the GEQ) collected from experiment 2 (chapter 4). This difficulty progression is seen in figure 5.11 which shows the boxplot of the challenge scores across these 4 levels.

The decision process for difficulty adaptation uses the mastery prediction that is returned from the webserver to change the game level. All players start the game in level 1, if the system predicts a high level of mastery after gameplay, the system increases the difficulty of the game by selecting the next level. If the system predicts a high level of mastery after the player experiences level 4 of the game, the difficulty remains constant. Similarly, if the system predicts a low level of mastery, the difficulty is reduced by selecting the previous level.

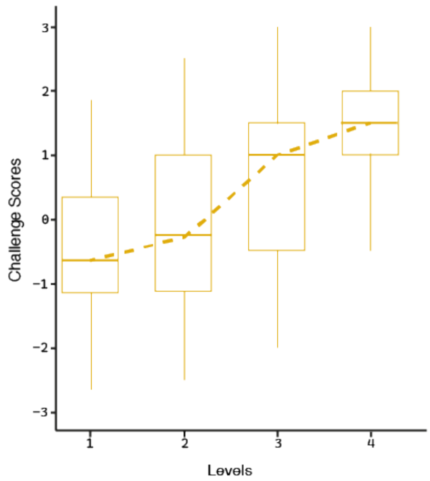


Figure 5.11: The figure shows the boxplot of the challenge scores across the 4 levels of the game game. The difficulty progression is shown with the dotted line.

If the system predicts a low level of mastery from level 1, the difficulty is kept constant.

The goal of this adaption system is to keep the players in an optimal state of mastery experienced from the game. This minimises the chances of the player getting demotivated due to the game being too difficult or easy for them. This adaptive system is hypothesised to improve player's confidence and make them more likely to engage with the physical activities involved in gameplay. This is further linked to the theory of game flow [43, 36], so it is expected that the adaptive system will increase the immersive nature of the game by maintaining an optimal balance between the game's challenge and the player's skills. In this research, skill is measured through perceived mastery in the game.

The Running Chickens game (which was built in Unity) was modified to create the adaptive game. This version of Running Chickens is the first adaptive AR game based on body movement. The effectiveness of this adaptation system in improving player engagement and confidence is evaluated in a user study that is described in the next section of this chapter.

5.5 Experiment 2

This section describes the experiment conducted to evaluate the effectiveness of the mastery-based adaptation system. Mastery-based difficulty adaptation is hypothesised to improve the players' confidence in the game and increase the immersive nature of the game. This study investigates this hypothesis through a controlled experiment with players.

The experiment compares this adaptive game against a non-Adaptive version (used as the control condition) of the Running Chickens game. Participants played both versions of the game and provided both quantitative (from questionnaires) and qualitative (from semi-structured interviews) about their experience in each of the study conditions.

Due to COVID-19 social distancing measures, this study was conducted remotely. Participants were contacted online and requested to travel to a specific park for the experiment (details provided in the next subsection).

5.5.1 Experiment Design

The experiment followed a within-participant design where each participant experienced all conditions of the study. The experiment compared the adaptive version of the Running Chickens game against a non-adaptive version (control condition), which resulted in 2 study conditions. Details about each condition are described below:

- **Adaptive Condition:** In this condition, participants played the adaptive version of the game (described in section 5.4) for 8 rounds.
- **Non-Adaptive Condition:** In this condition participants played the running chickens game without any DDA techniques applied. Participants would start the game at level 1 and play each of the levels twice moving across levels 1-4. This condition is considered as having linear difficulty since participants start with an easy level of the game and move progressively to more difficult levels of the game. This linear difficulty is considered as a control condition since it has been traditionally used in a number of digital games and best aligns with players' mental models of difficulty progression within these types of games (as indicated from qualitative data presented in section 5.6.2). Linear difficulty has also been used by other researchers as a control condition to evaluate the effectiveness of DDA systems in games[131].

Each session consisted of 8 game rounds, each one took approximately 90 seconds to complete. This included playing the game level for up to a minute as well as answering a rating based PX questionnaire (about that specific level) in each game round. At the end of each session, participants filled up a post-session PX questionnaire evaluating that version of the game (across all game rounds). Participants were also given a training session on how to play the game before they experienced the experiment conditions. The same training levels from previous studies were used in the experiment.

Similar to previous studies, all experiments were conducted during daylight and adequate weather conditions to minimise environmental factors from biasing the data collected. In this case, 2 parks were selected as sites for the study. One park was the same one used in the previous experiment while the second park was the Russell Square park (in Bloomsbury, London). An additional experiment site is used to give participants an option of travelling to a venue that suits their convenience. The order in which participants experience each experiment condition was counterbalanced to minimise ordering effects in the data collected. Participants were anonymised using IDs and were compensated £15.00 for their participation in the experiment. The sample of participants was restricted to people who owned or had access to an android device purchased after 2015. This was done since the Running Chickens game was developed only for the android platform and due to Covid-19 restrictions lab phones could not be used.

Procedure

All participants provided informed consent before participating in the study. Since this was a remote study, participants who provided informed consent were sent the questionnaires and games used in the study via email. The participants would need to install 3 applications to the mobile device: 1 training game and 2 versions of the Running chickens game. Participants were also asked to travel to one of the 2 chosen sites for the experiment in order to participate in the study during daylight. Participants were also provided with an anonymised ID (which they would use as an identifier when filling up the questionnaires) and the order in which they were expected to play the 2 experiment sessions. Participants were informed to start the study at least 2 hours before sunset to ensure appropriate lighting conditions for the entire study. They were also advised not to do the study if it was raining or if the park was slippery (just after rainfall). In addition to email information, the researcher set up a video call to give them a brief on what they were expected to do across the study. Participants were also asked to fill up a pre-experience questionnaire about their background and experience with MAR games before they arrive at the experiment site.

At the beginning of the study, once they had arrived at the experiment site, participants were given a training session about the game and the questionnaires used in the study. In the training session, participants were requested to watch a video provided that shows the researcher demonstrating how the game is played and how to go about answering the questionnaires in the study.

Following which participants, were requested to play through 2 levels in the training game provided to them. Participants were informed that they could contact the researcher via phone or text if they had any doubts or problems with the game.

After training, participants played through the two experiment sessions which were given to them as two different android applications. The order in which they were expected to do the sessions were assigned to them at the beginning of the study. In each session, participants played 8 rounds of the adaptive or non-adaptive version of Running Chickens at the end of each round participants filled up a short PX questionnaire about their experience in that level. Participants were asked to take a break of up to 5 min between the rounds of the game to minimise their physical fatigue across the study. At the end of the 8 game rounds, participants filled up a longer post-session questionnaire about their experience across that game version (all 8 game rounds). Once the participant had completed both experiment conditions, a semi-structured interview was conducted with them to collect qualitative data about their experience across both conditions of the study.

While playing the game, if the AR algorithm lost environment tracking participants were advised to proceed to the next game round. If this happened in the adaptive game, the same level was given to the participant in the next round (without any changes to the game difficulty). If a participant experienced 3 or more instances of this problem, their data was excluded from the analysis.

At the end of the semi-structured interview, participants were debriefed, all their questions were answered and the study was concluded. The experiment took approximately 60-90 min for each participant to complete (once they had arrived at the study site).

5.5.2 Data Collection

The study collected both qualitative and quantitative player experience data from participants. This data was used to compare differences in PX between the two study conditions and as ground truth to evaluate model performance. Player movement data, scores and timing information was logged from the mobile device (similar to previous studies) for each game level, this data was used to compute the mastery predictions by the webserver. These mastery predictions were also logged for analysis to evaluate the performance of the model on unseen players.

Participants

Participants were recruited using the same mailing lists used in the previous study. In addition to this participants were also recruited from social media groups for student accommodations managed by the University of London. These social media groups were selected since the COVID-19 lockdown measures in the UK prevented many students from travelling to the university campus.

The locations of the 2 study sites selected for the experiment coincided with these 2 streams of recruitment. Students who are living around the Queen Mary campus could select the park that was used in the previous study. While students living in the University of London accommodations could use the Russell Square park (which is near the accommodations targeted). This is done to ensure that participants do not need to travel long distances for the experiment which was in line with health and safety guidelines released by the UK government at the time the experiment was conducted.

The study sample consisted of 24 volunteers (12 male and 12 female) aged between 18-34 (9 participants had ages in the range of 18-24, 9 were in the 25-29 age range and 6 were above 30), summarised in table 5.8. When asked about prior experience playing AR games, 13 participants had no prior experience. In the remaining 11 subjects, 3 reported having played only one experience in the past and 8 reported having played a few times in the pasted. None of the participants in this sample reported playing AR games regularly. Refer to table 5.9 for a summary of the participants' previous AR experiences.

Table 5.8: Summary of participants' ages.

Age range	Number of participants
18-24	9
25-29	9
Above 30	6

Table 5.9: Summary of Participants' Previous experience with AR games.

Previous AR experience	Number of participants
No experience	13
Played only once before in the past	3
Played only a few times in the past	8
Played AR games regularly	0

Questionnaires Data

The study used ratings based questionnaires to collect PX data. These questionnaires were given to players after each level and at the end of each session. After each level players filled up a questionnaire which comprised of the mastery subscale of the PXI [220, 6] which consisted of 3 questions (rated on a 7 point Likert scale). The participants' mastery experience was collected for each level as this data was used as ground truth to evaluate the performance of the mastery prediction model. Players answered this questionnaire based on their mastery experience in that level.

At the end of the study, players answered a post-session questionnaire. For this, the complete version of the PXI [220, 6] was used. This questionnaire consists of 10 subscales that relate to both the functional and psychological consequences of gameplay. The subscales for the functional consequences are ease of control, (clear) goals and rules, challenge, (clear) progress feedback and audio-visual appeal. The psychological consequences are meaning, curiosity, mastery, immersion and autonomy. This data was used to compare PX between the two game versions.

Qualitative Data

Qualitative data was gathered from participants in the form of a short semi-structured interview which was conducted once both versions of the game had been played. Participants were asked to compare the versions of Running Chickens. Participants were also asked about their preferences between the two versions of the game in terms of enjoyment and challenge. The goal of this interview is to compare any PX differences between both game versions and to check to what extent the effects of the mastery-based DDA adaptation could be perceived by players.

5.5.3 Analysis and Results

The resulting data set from the experiment was used to compare the adaptive and non-adaptive versions of Running Chickens based on player experience. Statistical analysis is conducted to compare the game versions using the post-session PXI data for each game version. Qualitative responses from semi-structured interviews were analysed using thematic analysis to understand PX differences and individual preferences between both game versions. Finally, data from the short post-level mastery questionnaire was used as ground truth to test the performance of the mastery prediction model.

Statistical Analysis of Game Adaptation

The effects of mastery-based DDA on PX was analysed by comparing PXI scores from the adaptive game against scores from the non-adaptive game (control condition). The data set did not meet the assumption of normality which was evaluated using a Shapiro-Wilk test (conducted for each of the 10 subscales of the PXI). Since the data was found to be non-normal, non-parametric statistical tests were used to analyse the data. A Wilcoxon signed-rank test (with a p-value of 0.05) was used to check for statistical differences between game versions for each of the 10 sub-scales of the PXI. The results for each of the PXI sub-scales are reported below:

Ease of Control: The was no significant difference in the scores of ease of control between both game versions ($p = 0.275$). The boxplot of the ease of control scores is shown for both study conditions is shown in figure 5.12, which shows that scores are similar across both conditions.

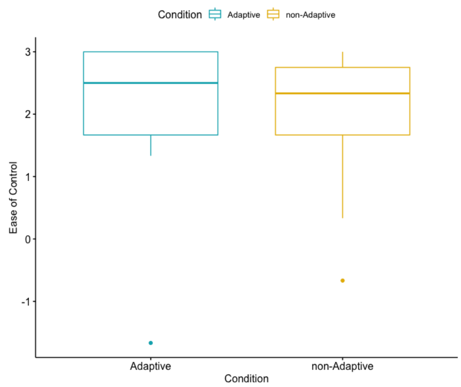


Figure 5.12: The figure shows the **boxplot plot for the Ease of control scores** across the study conditions. The image shows that scores are similar across both conditions.

Goals and Rules: There was no significant difference in the scores for Goals and Rules between both game versions ($p = 0.305$). The boxplot of the Goals and Rules scores is shown in figure 5.13, which shows that scores are similar across both conditions.

Challenge: There was no significant difference in the scores for Challenge between both game versions ($p = 0.837$). The boxplot of the Challenge scores is shown in figure 5.14, which shows that scores are similar across both conditions.

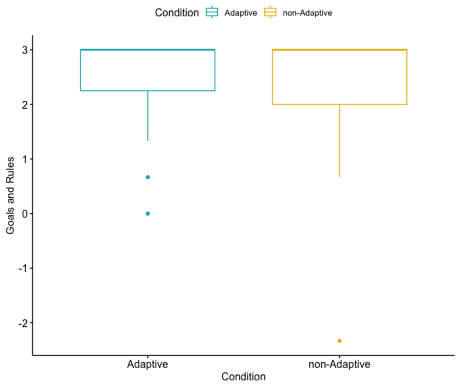


Figure 5.13: The figure shows the **boxplot plot for the Goals and rules scores** across the study conditions. The image shows that scores are similar across both conditions.

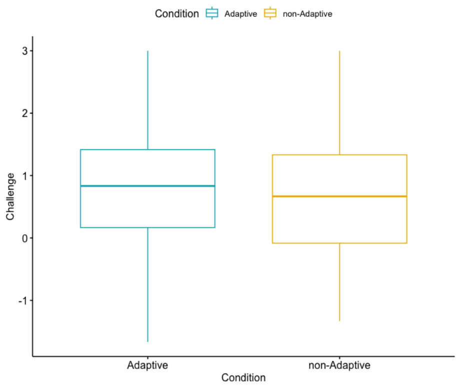


Figure 5.14: The figure shows the **boxplot plot for the Challenge scores** across the study conditions. The image shows that scores are similar across both conditions.

Progress Feedback: There was no significant difference in the scores for Progress Feedback between both game versions (p = 0.305). The boxplot of the Progress Feedback scores is shown in figure 5.15, which shows that scores are similar across both conditions.

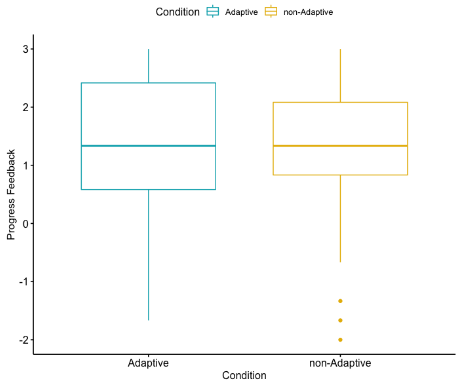


Figure 5.15: The figure shows the **boxplot plot for the Progress feedback scores** across the study conditions. The image shows that scores are similar across both conditions.

Audiovisual Appeal: There was a difference approaching significance in the scores for Audiovisual Appeal between both game versions (p = 0.055) with a small effect size (r = 0.394). Participants scored the audiovisual appeal of the adaptive game higher than the non-adaptive version. The boxplot of the Audiovisual Appeal scores is shown in figure 5.16, which shows that scores for the adaptive condition are marginally higher.

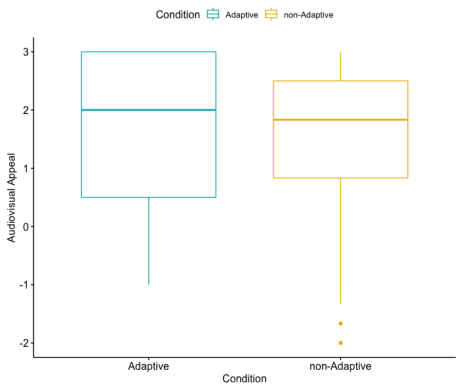


Figure 5.16: The figure shows the **boxplot plot for the Audiovisual appeal scores** across the study conditions. The image shows that scores in the adaptive condition are marginally higher.

Meaning: There was no significant difference in the scores for Meaning between both game versions (p = 0.948). The boxplot of the Meaning scores is shown in figure 5.17, which shows that scores are similar across both conditions.

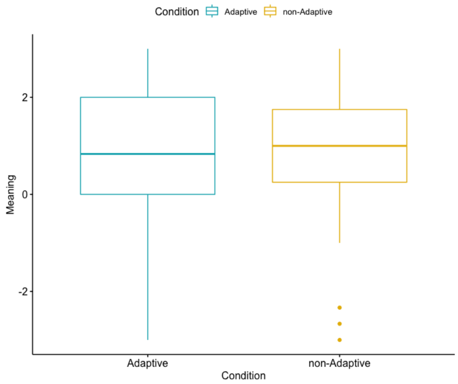


Figure 5.17: The figure shows the **boxplot plot for the Meaning scores** across the study conditions. The image shows that scores are similar across both conditions.

Curiosity: There was no significant difference in the scores for Curiosity between both game versions (p = 0.118). The boxplot of the Curiosity scores is shown in figure 5.18, which shows that scores are similar across both conditions.

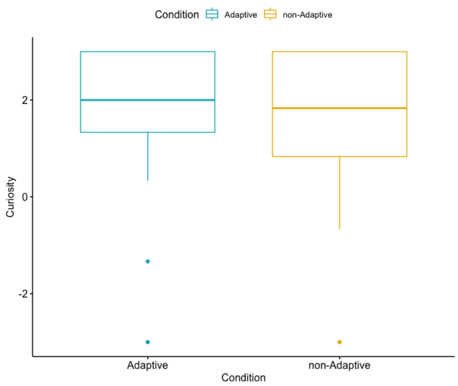


Figure 5.18: The figure shows the **boxplot plot for the Curiosity scores** across the study conditions. The image shows that scores are similar across both conditions.

Mastery: There was a significant difference in the scores for Mastery between both game versions (p = 0.011) with a moderate effect size (r = 0.540). Participants scores their perceived mastery for the adaptive game higher than the non-adaptive version. The boxplot of the mastery scores is shown in figure 5.19, which shows that scores for the adaptive condition are higher.

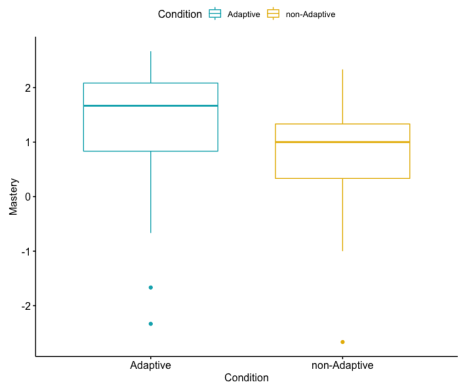


Figure 5.19: The figure shows the **boxplot plot for the Mastery scores** across the study conditions. The image shows that mastery scores are higher in the adaptive condition.

Immersion: There was no significant difference in the scores for Immersion between both game versions (p = 0.685). The boxplot of the Immersion scores is shown in figure 5.20, which shows that scores are similar across both conditions.

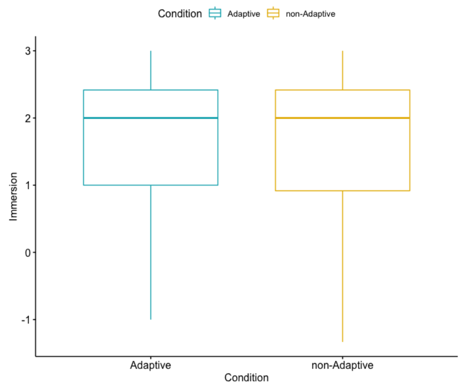


Figure 5.20: The figure shows the **boxplot plot for the Immersion scores** across the study conditions. The image shows that scores are similar across both conditions.

Autonomy: There was no significant difference in the scores for Autonomy between both game versions (p = 0.943). The boxplot of the Autonomy scores is shown in figure 5.21, which shows that scores are similar across both conditions.

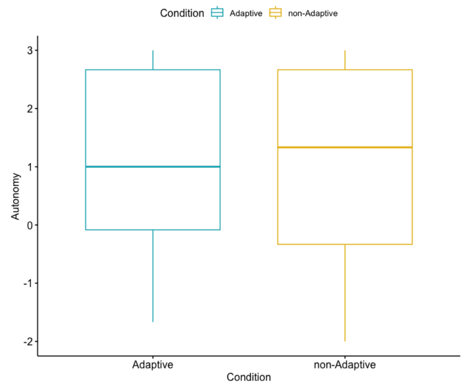


Figure 5.21: The figure shows the **boxplot plot for the Autonomy scores** across the study conditions. The image shows that scores are similar across both conditions.

This subsection reported the results from the statistical comparison of the adaptive and non-adaptive game versions based on PX. The analysis found that there was a significant difference in mastery scores between both games where players felt more masterful in the adaptive game condition. There was also a difference approaching significance in the audiovisual appeal scores where players found the audiovisual styling of the adaptive game more appealing.

Qualitative Analysis

A semi-structured interview was conducted once participants had played both versions of the game to understand how the adaptive game impacts player experience. The interview lasted about 5-10 min depending on the participant and focused on the perceived differences between both game versions. Participants were also asked to compare both game versions based on their experiences of challenge and enjoyment.

Participants made some general comments about their experience across both game versions. The most common observation was that players found the evasion mechanic as highly enjoyable and immersive. One of the participants reported that *”...when they (chickens) run away, you get really into the game. When I ran after them, I forgot I was in the park and that there were other people around me..”*. Another common observation was that it was difficult to navigate around the space when using the mobile’s view-port to navigate around the space. Finally, one participant said that the audio feedback from the game made them feel highly self-conscious in the public park. This person preferred playing the game on mute as they did not want to draw attention to themselves. This person also noted that they avoid exercising in public spaces as they do not want to be judged by other people for their physical inability.

It was interesting to note that most participants could not perceive the effects of the difficulty adaptation. One of the participants remarked that the game difficulty in the adaptive version seemed to change depending on how they performed in the previous level. While another participant reported that the adaptive game seemed to get easier when they experienced physical fatigue from gameplay.

The majority of participants (15 out of 24) found the adaptive game less challenging. Among the remaining participants, 5 found the non-adaptive game less challenging and 4 found no differences in challenge across both versions.

Participants who found the adaptive game as less challenging reported that the game would present them with easier levels when they felt either tired to overwhelmed playing a difficult level in the game. The other participants who reported the adaptive version as more challenging mainly attributed this due to the spatial distribution of chickens in each level. When chickens were spaced far apart from each other participants found the game more challenging as they needed to increase the amount of physical exertion needed to navigate around the game level.

Participants responses about game enjoyment showed that 13 (out of 24) found the adaptive game more enjoyable while 7 (out of 24) found the non-adaptive game as more enjoyable and the remaining 4 enjoyed both versions equally. When comparing participant comparisons about game challenge and enjoyment, to types of players emerged in the sample: *Challenge seekers* and *Reward Seekers*. In the sample of participants, 10 were found to be *Challenge seekers* while 8 were found to be *Reward seekers* and 6 participants could not be categorised into either of these groups.

Challenge seekers preferred the difficulty to increase constantly as the game progressed. When the adaptive game presented this player type with easier levels (if low mastery was predicted), they found this mechanic frustrating as they wanted to continue experiencing the game at a high level of difficulty (irrespective of their performance in the difficult levels of the game). One participant in this group reported, *"...the first session (adaptive game) was a lot easier but I don't like that, I prefer these games to keep giving me difficult levels so that I can push myself and improve my fitness..."* This group of people had a high sense of self-efficacy towards physical activity. Participants sense of self-efficacy was inferred by asking about their relationship towards physical activity. If participants regularly engaged in physical activity and enjoyed such activities were considered as having high self-efficacy. These participants reported either being physically active or motivated towards improving their physical fitness.

Reward seekers as the name implies seemed highly motivated by game rewards. This group reported being overwhelmed or discouraged when game difficulty constantly increased. These players found the adaptive condition more enjoyable since easier levels were presented to them when they felt frustrated with their performance. This group was motivated by getting as high a score as possible with the least effort. This group of people had a comparatively lower sense of self-efficacy toward physical fitness and tended to be less physically active in their daily lives.

A final observation was the importance of players physical fatigue during gameplay. In spite of being encouraged to take breaks between levels, some players reported feeling fatigued towards the end of each experiment session. They reported that they were not able to play as effectively when they were tired. When player fatigue was high the adaptation rules helped keep players motivated to game-play. It is important to note that this observation was made by *Reward Seekers*. This group found linear difficulty progression as frustrating as it had a negative impact on their game enjoyment. One of the participants remarked, *"...the second session (non-adaptive game) was really bad for me, I was really tired in the last few rounds but the game seemed to get more difficult. I felt really out of shape and was barely able to survive the last level..."*

Analysis of Classification Performance of Mastery Model

To evaluate the performance of real-time mastery prediction, participants were asked to evaluate their sense of mastery using (using the sub-scale of the PXI) at the end of each level. The mastery scores for each level of the adaptive condition was used as ground truth and compared to the predictions logged from the XGBoost model for that level. The ratings-based mastery scores were transformed to binary labels using the player based transformation technique introduced in the previous chapter. The resulting data-set consisted of mastery labels from 187 games (data from 5 games were discarded due to loss of tracking during the game). In these games, 107 were labelled as high mastery and 80 as low mastery which showed a mild imbalance between both classes in the data set. Based on the distribution of these labels, random chance classification accuracy was computed at 51.04%. The classification accuracy of the mastery model was 69.51% which is higher than random chance. The confusion matrix illustrated in figure 5.22 shows that the models perform well at identifying true positives (cases where the participant and classifier reported a high level of mastery). However, the classifier tended to misclassify cases of low mastery as indicated by the high false-positive rate. Due to class imbalance, the analysis of the precision and recall scores of the mastery classifier was also computed. In this case, precision is the measure of how well the classifier performs at identifying cases of high mastery that were actually high mastery cases which was found to be 0.702. While recall is the measure of how well the classifier performs at predicting cases of high mastery over all actual cases of high mastery which was found to be 0.813. The comparatively low score of precision confirms the model's bias of misclassifying true cases low mastery as high. The F1 score, which combines the precision and recall measures was found to be 0.753 for the mastery model. Overall the analysis indicated the prediction model works well for unseen players. However, there were a number of cases to misclassification in low mastery. This causes the adaptation system to increase the game difficulty for already overwhelmed players.

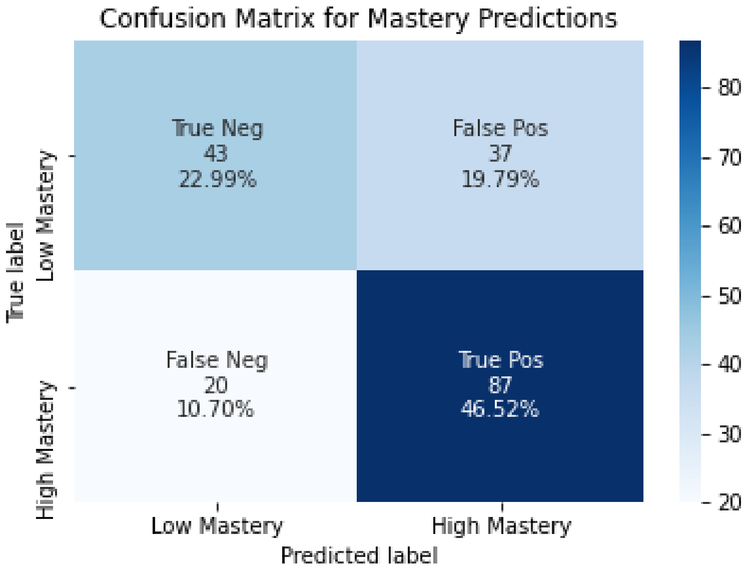


Figure 5.22: The figure shows the confusion matrix for Mastery predictions

5.5.4 Experiment Summary

The experiment described in this section builds on the previous one by evaluating the potential of using player experience models to dynamically adjust the difficulty of MAR exergames. This involved the creation of a game adaptation engine that uses mastery-based predictions to adjust the game's difficulty. The goal of this adaptation is to gradually increase game difficulty while maintaining an optimal balance of player mastery. This experience-based adaptation mechanic is evaluated against the non-adaptive game (with linear difficulty progression) in a controlled experiment with players. Statistical analysis indicated that players felt a significantly higher sense of mastery in the adaptive game and found the audio and visuals of the adaptive game more appealing (difference was approaching significance). Qualitative analysis of the participants' interviews indicated that the adaptive game was less challenging for players however, two distinct player types emerged. *Challenge Seekers*, who preferred the game to constantly increase the game difficulty and felt a sense of failure/frustration when adaptive game decreased the game difficulty. *Reward Seekers*, preferred the adaptive game as they were highly motivated by getting a high score in games and reported that game adaptations prevented them from feeling overwhelmed during gameplay. Qualitative results also indicated that the most significant determinant of player type was a person's self-efficacy towards physical activity. *Challenge Seekers* had a high self-efficacy to physical activity and enjoyed engaging with high-intensity physical activity. While *Reward Seekers* had a low self-efficacy for physical activity and avoided heavy physical activity. Finally, analysis of the classification performance of the mastery model showed that the model was good at predicting cases of high mastery however, tended to misclassify cases of low mastery as false positives.

5.6 Discussion

This chapter reports the results of 2 experiments that explore the impacts of experience-based difficulty adaptation in a MAR exergame. In the first experiment, a user-centric evaluation of the Running Chickens exergame parameters on dimensions of player experience is conducted. Additionally, player data is used to create models that can predict several dimensions of player motivations. This study builds on the study reported in the previous chapter by exploring player measures that are important in motivating players in engaging with the game. The mastery model created from data collected in the first experiment is used to create an adaptive game. The adaptive system adjusts difficulty based on predicted mastery. The goal of the adaptive mechanic is to keep players in an optimal state of mastery in the game, keeping them motivated towards gameplay. The second experiment reported evaluates this experience-based adaptive mechanic in a controlled experiment with players. This experiment showed that players models constructed from player movement and game-based data can be used to meaningfully personalise PX in MAR exergames.

5.6.1 Game Parameters impact on Player Experience

While the previous chapter explored broader dimensions of PX, the first experiment of this chapter takes a more focused approach of exploring player experience dimensions that are relevant to game motivations, which were: interest/enjoyment, mastery, autonomy and immersion.

Interest/enjoyment here is considered as a measure of intrinsic motivation for the player. This measure is considered as qualitative interviews with players in the previous study indicated that they found game interactions enjoyable and motivating. The study did not find any significant impact of game parameters on a player's interest/enjoyment. These results could mean that these game parameters do not directly influence intrinsic motivation however, they act as extrinsic motivators. Additionally, it can be argued that the study design of a controlled experiment used here is not the best way to assess intrinsic motivation. It would be useful to consider longitudinal study designs to understand how these game parameters can impact intrinsic motivations over time.

Mastery is the second measure considered as it is similar to competence which is an important factor for player motivations. Competence was explored in the previous chapter, the studies in this chapter explored mastery to investigate to what extent results of the previous chapter are generalizable and reproducible. It is important to note, that mastery is contextualized as 'A sense of competence and mastery derived from playing the game'[6], this is a psychosocial consequence of gameplay. This is different from the mastery of controls referred to by Przybylski et. al. [181] which refers to the learned ability to effortlessly control game interactions through a digital interface. In the first experiment, the number of chickens and the evasion mechanic parameter had a significant impact on a player's sense of mastery. It is observed that as the number of chickens increase the perceived mastery decreases. Additionally, when chickens evade players mastery significantly decreases. These results support the analysis of competence from the previous chapter. The results indicate that players feel masterful in simpler game levels (eg: levels with a small number of chickens that are stationary). However, qualitative data from the previous study indicated that participants enjoyed the evasion mechanic and expected the game progression to increase in challenge. The results here suggest that simple levels can be used at any point during the progression of the game to satisfy a player's need for competence from the game system. This would ensure that players remain engaged in the game.

Autonomy is the next measure considered as it is another important dimension of motivation in games. This measure is explored in this study since it is our hypothesis that some game parameters (specifically, the number of chickens and the evasion mechanic) can modulate the number of choices and sense of freedom a player experiences during gameplay. Understanding the impact of game parameters on autonomy would further explain how dimensions of PX such as immersion and flow were modulated in the previous study. However, this study was unable to validate this hypothesis as no significant effects were observed from the game parameters on autonomy.

Finally, immersion is the last measure considered as it was explored in the previous study. The impact of game parameters on player immersion would be important for designing game progression systems to keep players immersed in the game. In this study, results show that the number of chickens has a significant impact on player immersion. The immersion scores increase when the number of chickens is high. These results suggest that the number of chickens is an important parameter for game design; while a high number can lead to more challenging levels, it is important in keeping players immersed in the game. This is because when chicken numbers are high, players are kept more involved in the game since there are more game objects for players to interact within a level. It is possible that players enjoy physical interactions with AR content closer to their location as opposed to needing to travel across large areas to interact with AR content.

In the previous chapter, it was found that participants were more immersed in the AR world when they were collecting chickens that were close to them however, when they had to navigate through the space they shifted their attention to the non-AR world (which tended to happen in larger areas). Since a high number of chickens leads to a higher density of chickens within the space, players are more focused within the AR world with fewer attention shifts to the non-AR world. This could be another reason for higher immersion scores for the conditions with the high number of chickens, fewer attention shifts would allow for players to be more immersed within the game world.

5.6.2 Modelling Player Experience

Predictive models evaluated in the first experiment show that some dimensions of player experience evaluated here can be used to create experience-based adaptive games. Similar to the previous chapter it is observed that the XGBoost classifier performed the best for both 10-fold CV and LOSO-CV evaluation metrics. However, LOSO-CV scores across all dimensions showed a higher variance which indicates that the predictive models show a high accuracy only for a few participants.

The dimensions of interest/enjoyment showed low accuracy which is similar to results observed for Valence from the previous chapter. While previous studies do not explore interest/enjoyment, it is similar to dimension Fun in the first study and the dimensions of Positive Affect and Valence in the second study. It was previously shown that these dimensions of experience were difficult to model reliably based on movement and game-based data.

Autonomy and Immersion could be modelled with higher accuracy. However, in both cases, the logistic regression model does not generalise well to new players which are observed by the LOSO-CV scores for each performing close to random chance. For the LDA models, immersion does not generalise well (with accuracy scores for LOSO-CV performing close to random chance), however, autonomy shows a better performance with LOSO-CV scores performing higher than random chance. In both cases, the SVM (which is a nonlinear classifier) and the XGBoost (which is an ensemble classifier) performs well. This shows that for Autonomy and Immersion dimensions simple linear classifiers are not sufficient to create reliable predictive models. The previous study investigated modelling immersion, in both studies, it is observed that the XGBoost model performs the best and produces the most generalisable predictive models. However, the linear classifiers (Logistic regression and LDA) perform better in the previous chapter. This is possibly due to the larger dataset used.

Finally, predictive models for mastery performed well producing models that generalise to unseen players (as indicated by the LOSO-CV scores for each of the models evaluated). The results observed here are similar to results from the competence models investigated in the previous study. It is interesting to note that the models for mastery perform marginally better than models for competence despite the larger dataset used in the previous study. This could be the case since the results for mastery from the PXI showed high reliability (shown by the Cronbach's α) as compared to the competence measured from the GEQ. This shows that when using the ground truth transformation method proposed in this research to convert Likert ratings to binary categorical variables, it is important that that the ratings responses show high reliability. However, it is unclear what is the ideal reliability of responses needed for the categorical transformation of the data, future work must investigate this. Analysis of the feature importance of the XGBoost mastery classifier showed that game features were least important in predicting mastery as compared to movement and performance features, this is particularly interesting as it suggests that with enough data, mastery can be predicted irrespective of game parameters. This is an important aspect in the development of general PX models that can be applied across several MAR games. The results of the second experiment showed that the XGBoost model for mastery generalises well to new players, however, the model tended to be biassed towards predicting a high level of mastery. This is meant that the model was good at predicting true positives cases (where participants felt high mastery and the model predicted the same) however, it tended to falsely classify participants who felt low mastery as feeling masterful. This has serious implications for experience-based adaptation since incorrectly increasing the difficulty for players who do not feel competent in a level could further overwhelm them.

5.6.3 Mastery-based Dynamic Difficult Adaptation

The effects of mastery-based adaption was explored in the second study of this chapter. The adaptive mechanic ensured that a player's experience was personalised based on their mastery while playing the game. The results indicated that players feel more masterful when this adaptation is applied which was confirmed by both qualitative and quantitative data. It was also interesting to observe that players found the game aesthetics more appealing in the adaptive condition (as indicated from the audiovisual appeal scores). However, this effect was only approaching significance. It is possible that when players feel more masterful in the game, they are more receptive to game aesthetics as they are less overwhelmed by game objectives. This potential of using difficulty adaption to influence the aesthetic appreciation of the game must be further investigated in future work.

The results confirmed the hypothesis that mastery-based game adaptation leads to players feeling a high sense of competence in the game. However, another hypothesis that this game adaptation would lead to players being more engaged or immersed in the game could not be validated since the statistical analysis of immersion scores was not significant. Additionally, the extent to which players enjoyed the adaptive game was dependant on their self-efficacy towards physical activity. For players with low self-efficacy named *Reward Seekers* the adaptation mechanics worked as expected. The difficulty of the adaptive game was adjusted depending on their predicted mastery, this ensured that their sense of confidence during gameplay was maintained. This group of players do not tend to engage in physical activity and were overwhelmed in more difficult game levels. Linear difficulty progression tended to have a negative impact on game enjoyment for this group and it was important for them to experience low difficulty levels after facing difficult levels. This group of people were highly influenced by the score at the end of the level, they felt their performance in the adaptive game were much better than the non-adaptive one.

People with high self-efficacy to physical activity named *Challenge Seekers* did not respond as intended to the mastery-based game adaptation. This group was less influenced by the game score. They were motivated to push themselves to higher levels of physical exertion through gameplay. They interpreted the difficulty adaptation as the game pandering to them. When presented with an easy level after a more challenging one, they felt it was a result of their failure in the previous level. This group was more interested in personal improvement in difficult levels of the game. This shows that for this group of players the existing rules for game adaptation cannot be used to improve game engagement. For these players, it would be better to maintain game difficulty if low mastery was predicted giving them the chance to improve their performance. Reducing the game's difficulty should only be considered if repeated plays of the same level resulting in low mastery being predicted by the classifier. The optimal rules for game adaptation for *Challenge Seekers* is an open research question that must be investigated in future work.

Finally, when considering these player types that have been observed in this study, it is important to note that players need not be only *Challenge seekers* and not *Reward seekers* (or vice-versa). In reality, a specific player can be a mix of both these player types, or they can oscillate between these two player types depending on their motivations for play or situational and environmental context.

This experiment shows that for exergames it is essential for adaptive systems to consider peoples preference for physical activity and their motivations for game-play. For real-time game adaptation, this is an important factor to be considered. It would be useful for these systems if predictive models could use player behaviour data from gameplay to automatically classify the player type (*Challenge* vs. *Reward Seekers*). This would allow for different adaptation rules to be applied that can better engage that player type.

5.6.4 Limitations of the Study

The most notable limitation of both experiments in this chapter was the limited sample of participants in each study. The sample comprised of mainly university students which makes the results difficult to generalise to other demographics of players.

Another limitation is that the sample of participants has a limited experience with mobile AR games. Since this genre of games is not yet a well-established type of consumer games, the PX impact of the AR game mechanics tested in this research could change as players become more familiar with this medium of gameplay. This increased familiarity could also have an impact on their in-game behaviour (or the way players move in the game). Since this behaviour data is the main input for the predictive models of player experience, it is unclear to what extent model performance could change. Future work can take two potential approaches to overcome this limitation. Either, participants can be pre-screened to ensure that there is a good balance of experienced vs non-experienced players in the sample with respect to AR games, this approach would have the advantage of being able to investigate the difference between these two categories of players. The second approach could use longitudinal studies where the participant sample would play the game across several sessions, this approach would help investigate how these game mechanics can motivate players as they get more familiar with the game.

Another limitation of this study is that participants' existing level of physical fitness could be a confounding variable. Since player motivations for a game that promotes physical exertion are dependant on their levels of physical fitness. This aspect of motivation is of interest to exergame research and is referred to as the idea of self-efficacy [126]. Self-efficacy can be understood as a person's belief in their ability to succeed in a particular situation [15]. The impact of players' different levels of physical fitness and self-efficacy towards physical exertion in mobile AR exergames should be investigated to find out how different game mechanics can impact experiential outcomes and motivations to play for different types of people. This would be beneficial for the design of personalised exergame experiences for these different types of players. While the qualitative analysis of data from the second study showed that a players self-efficacy towards physical activity was an important factor that mediated their enjoyment in adaptive AR exergames, it is unclear how this difference in self-efficacy would impact the statistical analysis conducted to understand game parameters impact on player experience (in experiment 1) and for experience-based game adaptation (in experiment 2).

The predictive models created in the first experiment are evaluated using 10-fold CV and LOSO-CV from the dataset collected in experiment 1. It is unclear how these models will generalise to unseen players. The second experiment addresses this limitation (to some extent) by evaluating the mastery model with new players. However, since the sample consisted of university students in both experiments it is unclear how these models will perform with a more generalised sample of participants.

The second experiment validated the mastery model for experience-based difficulty adaptation. However, the adaptive system was simple as it considered only 4 game levels. It is unclear to what extent this mastery-based adaption would perform when applied to a complex game design space.

Finally, it is worth considering the limitations of conducting this research work during the COVID-19 pandemic which had implications on both of the experiments reported in this chapter. In the first experiment, data collection had to be stopped after 25 participants which fell short of the intended goal of 40 participants. The second study had to be conducted remotely which had implications on the data collected. It was noticed that some participants did not take sufficient breaks across the experiment (as they were instructed) which increased their physical fatigue. This physical fatigue had severe implications on the PX data since these players had more energy in the first session of the study. Having a researcher present would have ensured more consistency in the experience of the second study across all participants.

5.7 Chapter Summary

The 2 experiments conducted in this chapter builds on the previous study to investigate approaches to player experience modelling for dynamic difficulty adaption in mobile AR exergames. The first experiment explored modelling player motivations within AR exergames as well as exploring the relationship between AR exergames and dimensions of player motivations. The results observed are used for the creation of personalised adaptive AR exergames that use player models created to maximise a player's motivations as they play the game. The results show that mastery is an important dimension of player motivations since it can be predicted through player models and these models generalise well to new players. Another important finding from the experiment is that exergame parameters of the Running Chickens game such as the number of chickens and the evasion mechanic are important features for adaptation as they can influence the levels of mastery and immersion player experience within a game. The second experiment investigated the application of the mastery model for dynamic difficulty adaptation. The experiment found that the mastery adaption improved people's sense of mastery in the game and the adaptive game's audio and visuals were more appealing for players. However, the adaptation was enjoyable and rewarding for players with a low sense of self-efficacy for physical activity. Players with a high sense of self-efficacy for physical activity found the adaptation frustrating as they disliked the game getting easier. In the second study, it is observed that mastery-based adaptation can perform well and generalise to new players however, it is clear that different player types require different adaption techniques respectively. The current adaptation system proposed in this chapter is ideal for people who do not engage in physical activity as it keeps them feeling confident while playing the game.

Chapter 6

Discussion and Conclusions

This last chapter summarises the main findings and contributions made in this thesis. The chapter is structured as follows: section 6.1 summarises the main findings across the four experiments conducted in this research work, putting the main findings in context with previous relevant research. Section 6.2 summarises the main contributions of this research work. Section 6.3 details the main limitations of the studies described in this thesis. Section 6.4 proposes directions of future research work. Finally, section 6.5 concludes with some reflections and lessons learned while researching player experience modelling in mobile AR games.

6.1 Discussion and Implications

This thesis aimed to address three research questions:

RQ1: *To what extent can player movement and game metric data be used to predict PX in mobile AR games?*

RQ2: *What is the impact of commonly used AR exergame parameters on PX?*

RQ3: *Can these predictive models of PX be used for dynamic difficulty adaptation in mobile games to improve PX?*

These research questions were addressed across the four experiments reported in this thesis. First, **RQ1** was investigated using two mobile AR games: *AR Treasure Hunt* which is an exploration-based game, and *Running Chickens* which is an exertion game. This RQ was addressed in the first three experiments (chapters 3, 4 and 5), which collected PX and corresponding player movement and game metric data to build and evaluate supervised learning models that predict several dimensions of PX. Second, **RQ2** was investigated using the *Running*

Chickens exergame across two studies (chapters 4, 5), using both quantitative and qualitative methods. In these studies, statistical tests were conducted to evaluate the effects of the game area, the number of rewards and evasiveness of the chickens (independent variables) on player experience (dependent variable). Additionally, qualitative data from player interviews (conducted as part of one of the studies reported in chapter 4) further contributed to this investigation. Finally, **RQ3** was addressed in the final study reported in this thesis (chapter 5), where the mastery model built in the previous experiment was used to develop an adaptive version of the *Running Chickens* game. The model is used to adjust the difficulty of the game to ensure that players are not overwhelmed by the game difficulty using the principles of Flow theory [44, 46, 208, 209, 43, 36]. This adaptive game was evaluated against a non-adaptive version of *Running Chickens* using both qualitative and quantitative approaches. Statistical analysis was conducted to evaluate mastery-driven game adaption (independent variable) effects on PX (dependent variable). At the same time, qualitative analysis of interviews was used to understand the experiential differences between the adaptive and non-adaptive versions for players.

The rest of this section discusses the main findings of the research, which has been grouped into the following topics: 1) Player Experience Modelling in mobile AR games, 2) Impact of Exergame parameters on Player Experience, and 3) Experience-driven Dynamic Difficulty Adaptation in Mobile AR games. These topics closely map to the three research questions investigated in this thesis and are discussed in the subsections below.

6.1.1 Player Experience Modelling in mobile AR games

During the first three studies reported in this thesis, player movement and game metrics data was used to predict several dimensions of player experience. The first study took a preference learning approach which aimed to model a number of emotion preferences: *Fun*, *Excitement*, *Boredom*, *Challenge* and *Frustration*. Player experience modelling was formulated as a preference learning problem where given data from two games; the model would predict which game the player felt the specific emotion more, e.g., *Did the player find the first game more fun than the second game? or vice-versa*. Results from the study indicated the *Challenge* and *Frustration* could be modelled to a high degree of accuracy while *Fun*, *Boredom* and *Excitement* were more difficult to predict reliably. These results are similar to findings from [167], who investigated predicting these emotion preferences in the *Super Mario* game. Interestingly, this pattern of challenge and frustration experiences being more predictable exists across both traditional (non-AR) games and movement-based mobile AR games. Given the difference in interactions and immersive nature of MAR games compared to non-AR games, it was expected that the emotion dimensions for player preference investigated in the first experiment would result in different dimensions being more predictable. However, it was found that similar patterns emerge as to findings from preference modelling in non-AR games, which is likely due to the experimental method selected. This method was selected following advice from researchers in affective computing [135, 238] who argue for the use of ranking based measures over rating based measures. Martinez and Yannakakis [135] suggest that due to the subjectivity and non-linear nature of ratings based measures makes the treatment of ratings as real value numbers fundamentally flawed. However, it is worth noting that while this is a strong case to use ranking approaches for research in affective computing, the emotion ranking-based measures have several limitations when applied to the domain of PX. For instance, a possible explanation for the high predictability of *challenge* and *frustration* is that when players were asked to evaluate a game task, they often find it difficult to disassociate game challenge and frustration. This overlap between emotion dimensions of preference has disadvantages for applications of these player models for game adaptation since it would be impossible to create a challenging but not frustrating game. As discussed in chapter 3, people are different in their ability 'to represent their experiences as categorically distinct events' [17] and this ability is influenced by context and language abilities. This highlights the difficulty in defining emotions in social communication, which has further implications for PX models that rely on a participant's understanding of these emotion definitions for ground truth. Thus this thesis advises other researchers to consider using traditional ratings-based approaches to establish ground truth for PX models.

For this reason, the next two experiments used theories developed by HCI researchers to understand PX. Instead of using constructs based on emotions, PX is often investigated using experiential constructs such as game competence, immersion, flow. However, these PX constructs are conventionally measured using ratings based questionnaires that have been developed and validated using exploratory and confirmatory factor analysis. The ratings are not treated as real values to overcome the disadvantage of applying ratings-based measures for predictive modelling. Instead, they are transformed into categorical variables. This approach is similar to [139] and [65]. In [139] the authors transformed ratings based measures into rank-based measures; however, their approach was developed to make comparisons between two different players, not different games played by the same person (which is the problem domain investigated in this thesis). In [65] the author transforms ratings-based measured into categorical variables to predict experiential differences of the same player across different levels of a game. However, they note that this transformation does not account for inter-player subjectivity when using the rating scale. For this reason, this research work proposed a new system of numerical transformation of rating measures that used the median statistic of each player to transform ratings into categorical variables, which were then modelled as a classification problem. This transformation technique was applied in the second and third experiments of this thesis, and the results from the leave-one-subject-out evaluation scores show that models built using this data can be generalised to new players. However, formal comparative evaluations of this transformation technique versus previous ones have not been conducted and must be addressed in future work. Finally, it is worth noting that the chosen questionnaire must be comprehensively validated for PX analysis. The need for questionnaire validation was observed and discussed in chapter 4 where the internal inconsistencies of the GEQ's structure negatively impacted model performance in that study.

Experiment 2 (chapter 4) aimed to model the player experience dimensions from the affective slider [22] and the GEQ [86]. The affective slider measures affect dimensions of arousal and valence, while the GEQ is made up of several dimensions relevant to PX: Competence, Sensory and Imaginative Immersion, Flow, Tension, Challenge, Negative Affect and Positive Affect. Results show that all these experiential dimensions can be predicted better than random chance levels. The predictive models of Competence, Challenge and Tension were the best performing models, while models for valence, positive affect and negative affect perform poorly.

Competence, Challenge and Tension, while being different experiences, are similar in that they depend on the player's response to game difficulty. These results are further validated by observations from the first study where frustration and challenge (the best performing models) are also related to game difficulty. This hypothesis is further validated in experiment 3 (chapter 6), which aimed to model 4 dimensions of PX: Interest/Enjoyment, Immersion, Mastery and Autonomy. This study showed that mastery was the best performing model, while Interest/Enjoyment and Autonomy performed poorly. This study further confirms that experiences associated with game difficulty can be predicted using player movement and game metrics data in mobile AR games.

Comparing results across the first three experiments conducted indicate that experiences related to a player's perceived skills in the game can be predicted to a reliable extent. For example, experiences such as mastery and (high) competence usually indicate that the player perceives themselves as highly skilled in the game task. While experiences such as tension, frustration and (high) challenge usually indicate that the player perceived themselves as lacking skill for the game task. These findings are particularly relevant for applications in game adaptation since these models can be used to put the player in a state of flow. Since one of the core requirements for flow is an optimal balance between task difficulty and skill. Reliably predicting a player's perceived skill from a game level allows the game system to tune difficulty to reach this optimal balance between player skill and game difficulty. This will ensure that players have an engaging experience without getting overwhelmed (or underwhelmed) by the game experience.

It is worth noting that this approach to game adaptation is potentially advantageous compared to performance-based adaptation since performance metrics such as game score are objective measures that do not account for player subjectivity. For instance, two different players with the same game score from a specific level may have very different perceptions of the game's difficulty or their skill at that level, so they will require different game adaptation steps to put them in a flow state. This approach is tested in the final experiment (chapter 6), which uses the mastery model developed in the previous experiment, the results of which are discussed in the final subsection of the section.

6.1.2 Impact of Exergame parameters on Player Experience

While several mobile AR exergames have been investigated in previous research [112, 119], these activities have been mainly case studies that validate these exergames with players. It is observed that these games use similar game parameters, which were: game area and the number of rewards in a level. Each of these game parameters can be varied to create a spectrum of player experiences across different levels of these games. While these research activities show the potential benefit of exergames, they do not provide information on how varying these game parameters will impact PX. This information is especially relevant to game designers working on similar games and designing adaptation systems. This research work addressed this gap in Experiments 2 (chapter 3) and 3 (chapter 4), which used the *Running Chickens* game where these game parameters were used to generate different game levels. Each of these experiments had players evaluating their experience across various game levels (with different settings of these game parameters) using ratings based questionnaires and qualitative interviews.

Experiment 2 found that game area had an impact on a player's experience of Tension and Challenge. Increasing the game area results in an increase in tension and challenge scores. At the same time, the number of game rewards impacted a player's experience of Valence, Competence, Flow, Tension, Challenge, and Positive Affect. Increasing the number of game rewards in a level was generally associated with positive experiences such as Valence and Flow. However, a large number of game rewards also results in an increased perceived challenge and tension due to the physical exertion required to obtain these rewards. Qualitative analysis of player interviews shows that PX was largely mediated by their preferences around each of the game parameter settings, which were in turn mediated by their self-efficacy towards physical activity. Players with a high self-efficacy towards physical activity preferred larger game areas as they enjoyed the increased physical exertion required to engage in these types of levels. In contrast, players with a low self-efficacy towards physical activity preferred smaller game areas and had a mixed response about their preference towards the number of game rewards. A portion of this group reported enjoying a high number of rewards in smaller levels and a low number of rewards in larger levels. This indicates that although this group does not enjoy physical activity, they are still motivated by the rewards of the game. They feel rewarded when they physically exerted themselves to obtain game rewards; however, they can get easily overwhelmed if the amount of physical exertion required to play the game becomes too high (which is the case when the game area is large and there are a high number of rewards). These observations were further validated in experiment 3 (chapter 5), which found that game area impacts a player's sense of mastery, where increasing the game area reduced a player's sense of mastery in the game level. The number of game rewards impacted a player's sense of mastery and immersion in the game. Increasing the number of game rewards resulted in a higher sense of immersion and decreased perceived mastery.

These experiments show that exergame designers must consider their players' attitudes, motivations and self-efficacy towards physical activity. People who enjoy physical activity will enjoy challenging levels with a large game area and a high number of rewards. While people who do not engage in physical activity prefer easier levels, it is worth noting that this group reported enjoying the game despite needing to physically exert themselves. However, they were motivated by the game's reward system. These results are similar to results from Mcvean and Robertson [126] who investigate mobile exergames motivations and behaviour in school children. While the authors explored location-based games not in AR, it is interesting to observe the similarity in-game motivations between players. Additionally, the number of rewards in game levels is an important mediator of player immersion and perceived competence in the game. While increasing the number of rewards tends to get players more engaged in the game, if the player has a low self-efficacy to physical activity or if they are reaching exhaustion (from the physical exertion), it could have a negative result by overwhelming players resulting in experiences of frustration or tension. However, the number of rewards can be varied across levels to ensure that the player does not get overwhelmed during the game. Increases in the game area parameter generally tended to increase experiences of challenge, tension and decreases in perceived competence. However, this is an important game parameter for controlling the amount of physical exertion required to play the game. For mobile AR exergames to result in positive fitness outcomes for players, the game area parameter must be increased over time, so players gradually increase their physical activity over longitudinal game-play. It is advisable to make small gradual increases to this parameter over time to not overwhelm players. Qualitative results suggest that for players with high self-efficacy for physical activity, this gradual increase in the game area can be faster than players with low self-efficacy towards physical activity. Exergame adaptation techniques that successfully increase physical activity are important for future work that will further benefit designers.

Finally, in addition to the mobile AR game parameters discussed above, this research work investigated a novel parameter referred to as the *evasion mechanic*, the speed at which game collectables could evade capture. This is an important feature of target acquisition games, such as *Running Chickens*. The mechanic was found to have a highly immersive and enjoyable effect on players despite their self-efficacy towards physical activity. This could be mainly due to the novelty of the mobile AR platform. It is worth noting that players who volunteered in the studies found the game objective of capturing chickens that run away from them unique and pleasurable. However, it is unclear how this mechanic can engage with players over time once novelty effects have depleted. From the results observed in this research, it can be concluded that in mobile AR exergames, mechanics that increase the level of presence and immersion a player experiences can motivate them towards high levels of physical exertion in short-term play sessions.

6.1.3 Experience-driven Dynamic Difficulty Adaptation in Mobile AR games

From the discussion in subsection 6.1.1, it can be concluded that PX models built using supervised learning algorithms can be used to predict experiential constructs related to a player's perceived skill in the game to a reliable extent. Given this conclusion, this subsection reflects on findings from experiment 4 (chapter 5), which evaluates the use of these PX models for DDA in mobile AR games. The mastery model developed was applied to a game adaptation engine that personalised the experience for individuals.

Research into player motivations in games has shown that a player's sense of mastery or competence is an important factor to keep players motivated to play the game[181, 188]. While this previous work has primarily focused on non-AR games, this research extends these findings to MAR games. The results from the statistical analysis of mastery-based game adaption demonstrated that the players in the adaptive game felt a higher sense of mastery for the adaptive version and tended to find audiovisual aesthetics in the adaptive game more appealing. These results suggest that players feel more confident in the adaptive game and do not feel overwhelmed by the game environment, reducing the cognitive load on players and allowing them a higher sense of aesthetic appreciation of the game. These findings are important since they show that mastery-based game adaption can ensure that players feel competent.

This research work investigated a simple adaptation system that is built based on Flow theory[44, 46, 208, 209, 43, 36], where the mastery prediction is used to adjust the difficulty of the game in such a way that difficulty is increased if the player feels a high sense of mastery and decreased if the predicted mastery is low. This mechanic aims to keep players in a flow channel, which previous research suggests should be more immersive for players. However, these findings could not be validated since the statistical analysis conducted did not show any differences in immersion scores across the adaptive and non-adaptive versions of the game. This can be explained by the qualitative insights from player interviews which revealed two different player types: *Challenge Seekers* and *Reward Seekers*. *Challenge Seekers* had high self-efficacy toward physical activity and perceived themselves as being physically active. In comparison, *Reward Seekers* had a low sense of self-efficacy for physical activity and generally tended to perceive themselves as unskilled in physical activity. These two diverse types of players had different experiential responses to the adaption system. For the *Reward Seeker* group, the adaptation system seems to work as desired. They reported finding the adaptive game more enjoyable and engaging than the non-adaptive version (which tended to result in an overwhelming experience). While *Challenge Seekers*, did not find the game adaptation enjoyable. This group was more focused on increasing their physical exertion through game-play; thus, they found decreases in difficulty frustrating. They attributed the game reducing difficulty (usually after a challenging level) to their inability to perform well at the game level. These results show that this adaption system based on Flow theory can engage players who are not physically active. Adaptive mobile AR exergames can engage these types of players in casual play, resulting in improved physical benefits for this group. However, exergame adaptive mechanics for physically active players by translating principles of Flow theory does not appear to engage them in the same way as the other group. This is because peoples perception of skill in exergames is related to their self-efficacy towards physical activity. So people with a high-self efficacy will perceive themselves as being highly skilled in the game level, so decreases in difficulty will result in a negative or frustrating experience for them.

6.2 Contributions

This work contributes towards the advancement of mobile AR games and shows that these games can leverage several data streams such as sensor data and game metrics to model player experience. Furthermore, this work investigates movement data extracted from mobile IMU sensors and game data for real-time player experience prediction and personalisation. This personalisation is especially important for exergames that promote physical fitness. This PhD thesis makes novel contributions to the existing literature in player modelling, mobile AR exergames and Dynamic Difficulty Adaptation in games.

This research evaluates two commonly used AR exergame patterns: area of the game level and the number of rewards (e.g., collectables). Experiment results showed that game area positively impacted experiences of tension and challenge. In general, when playing local AR games, players find navigating over large areas challenging, leading to a negative experience if they are already overwhelmed or do not generally engage in physical activities. This research showed that the number of rewards positively impacts experiences of Arousal, Flow, Tension, Challenge, Positive Affect and Immersion. In general, participants found a high number of rewards motivating as it enabled them to get immersed in the game world; however, if they are approaching fatigue or overwhelmed by the game difficulty, this can have a negative experience on players. Finally, both game patterns impact the amount players must navigate around a physical area during gameplay. This research found that as players move around a space, their attention switches between the AR world (through the mobile viewport) and the non-AR world (using their peripheral vision). This attention switching occurs to enable safe movement through the physical space during gameplay. However, this attention switching can break a player's immersion in the game.

This work proposes a novel methodology for Player Experience Prediction in mobile AR games. The player experience prediction system measures the player's movement, game parameters (e.g., the settings of different game mechanics) and player performance to infer their player experience in the game. Investigations of these modelling techniques suggest that experiences related to competence or perceived mastery can be predicted to a high degree of reliability.

This research further contributes a novel adaptation engine that uses a player's predicted mastery in a game level to adjust the game's difficulty. The adaption system shows how player models can be used in real-time to provide a personalised experience in MAR exertion games. User evaluations of this adaption system showed that this system is more beneficial for players who do not usually engage in physical activity. However, this is not the case for physically active people, as they are motivated by self-improvement since they do not respond positively to reductions in in-game difficulty.

6.3 Limitations

This section describes the main limitations of the research presented in this thesis and recommendations to overcome them.

6.3.1 Challenges in game type and interactions

Player experience modelling in mobile AR games is highly dependent on behaviour cues from players. This research work uses player movement as a source for player behaviour. It is important to note that different types of games and interactions afforded by the experience will influence player movements within the game. In this work, *AR Treasure Hunt* (exploration game) and *Running Chickens* (exergame) cause players to move around their physical spaces in very different ways despite using similar game parameters (area and number of rewards) to generate game levels. The experiment results suggest that the player experience prediction pipeline does generalise to both these game types. However, genres of mobile AR games exist that involve minimal player movement, a popular type of such games are referred to as AR table-top games. As the name suggests, the game experience is usually overlaid onto a physical table in the real world. The player would stand or sit around the table and use their mobile devices as a magic window to the game. In this case, it is unclear to what extent the modelling techniques introduced in this thesis will successfully generalise to these table-top games. Future work must investigate the potential of this player experience modelling pipeline on table-top games. Additionally, this thesis does not explore location-based MAR games, where players usually move between locations without using the mobile's viewport until they arrive at their desired destination. While modelling approaches used presented in this thesis can be used to model PX during AR gameplay events (when the player uses the device viewport to interact with the game), it does not account for travelling between game event locations which would have a significant impact on PX in location-based MAR games.

6.3.2 Challenges to user studies for AR games

This research work relied on conducting experiments with human participants to collect PX and related behaviour data. Opportunity sampling was used to recruit participants for the studies. To investigate the impact of game parameters on player experience, it is desirable to get a sample of participants of varied experiences with mobile AR games. However, it was a common observation that most participants found through opportunity sampling either do not have any experience or have very limited (tried these types of games once or twice in the past). This imbalance in the sample of participants has impacted the results of these studies. It is worth considering that when using mixed reality interfaces, there is a novelty effect that influences PX. The main influence of this novelty effect is observed in the investigations around game parameters and immersion. For instance, players found the evasion game mechanic in *Running Chickens* highly immersive. It is unclear how PX will change as novelty effects reduce over time and players get more familiar with gameplay and interactions. It is possible that as novelty effects decrease, player behaviour will also change. This will impact the predictive models of PX since it relies on this behaviour data. There are two ways this limitation can be overcome in future work. First, a pre-screening questionnaire can be used to ensure a study sample covers a uniform distribution of experienced and new players. Second, a longitudinal study design that involves players logging their experience and behaviour over some time (e.g., over several weeks) would allow for data collection as novelty effects decrease. This second approach is particularly interesting as it can inform researchers about how PX and behaviours change over time. This is particularly important for the design of exertion games that require regular gameplay for positive physical health outcomes.

Another limitation of this work is related to the investigation of exergames. Findings from this work suggest that player experience in exergames largely depends on people's self-efficacy towards physical activity. Different levels of self-efficacy towards physical activity influence the experiential impact of the game parameters investigated and the impact of game adaptation on player experience. However, the quantitative analysis conducted in this research work does not consider self-efficacy as a control variable in the analysis. This is mainly because inferences around player self-efficacy were made using qualitative data from participants. Future work must consider self-efficacy in the quantitative analysis of PX by gathering ratings-based measures about self-efficacy towards physical activity from players. The Self-Efficacy for Physical Activity (SEPA) scale [142] is an example of such a questionnaire that can be used in future work.

Finally, it is worth noting that the results observed and contributions made across this research mainly apply to people with normative physical and mental abilities. This does not include players with specific accessibility needs, such as players with reduced motor abilities (due to ageing, physical injury or diseases) or players with developmental challenges such as Autism Spectrum Disorder.

Players with these types of special needs would potentially have different bodily expressions in response to physical mobile AR experiences. Thus PX models driven by mobile device movement need to be optimised for each of these player types. This is a challenging problem since there isn't a one-size-fits-all approach to making these PX models accessible for non-normative players. Future work must involve these groups of players to understand their bodily expression in MAR games and improve the generalisability of PX models by using data collected from such research activities. Interesting questions emerge here about effective ways of making PX models more generalisable to all groups of players. For instance, can a single model be trained for all player types (given enough data), or would individual PX models for each player type be a more optimal strategy? These research challenges must be addressed in order to make PX models effective for sensitive player groups and for serious applications of MAR games.

6.3.3 Challenges in PX prediction

An important limitation of this work is the use of subject-independent models used for player experience prediction. Parsons and Reinebold[163] discuss that subject-independent methods are harder to create than subject-dependent methods as they have to work for all subjects. This limitation is observed in experiments 2 (chapter 3) and 3 (chapter 4), in which model accuracies were highly varied between participants. Exploring subject-dependant models (trained on data from a single participant) would overcome this limitation. However, it is worth noting that this is an expensive solution as it would require a large amount of data to be collected for each participant. An interesting alternative would be to take a hybrid approach, where subject-independent models are used for new players, and reinforcement techniques would be built into the system so that models can be personalised over time for different players. Finally, models created in this research work were built using limited data. The performance of these models is expected to improve with larger training data sets.

Additionally, modelling experiences and emotions are especially challenging, due to the difficulty in defining and measuring these experiential constructs. While a simple way to overcome this issue would be to model objective aspects of players i.e. their behaviour, this approach does not help game designers understand the experiential impact of their design decisions. This research explored the ranking of simple emotions, and traditional ratings-based questionnaires to measure player experience - the specific advantages and disadvantages of each have been discussed in section 6.1.1. Measuring and defining PX is an active field of research, the modelling techniques introduced in this research would further benefit from advancements in our ability to reliably measure PX.

Finally, it is worth considering the particular challenges with creating player models from movement data, since bodily expression is highly subjective. Additionally, in a video game context, the player's expression of their experience through body movement is often very subtle making it hard for even a human observer to be able to describe the player's experience. While this research did not use observational studies as a formal methodology, a researcher was present for the first 3 experiments. Observations of participant behaviour confirm this difficulty in recognising a player's emotion-state during gameplay. This is further confirmed by previous background research in this domain [193]. Based on this research the author advises interested readers to consider using players' self-reported measures of PX as compared to observer agreement for this domain. Additionally, the movement data measured from this type of gameplay has limited information about PX, while this research has demonstrated that experiences related to skill and mastery can be effectively modelled for predictive purposes, more complex experiences such as happiness are more difficult to model using this information medium.

6.3.4 Challenges to adaptation for AR exergames

Similar to the previous section, an important limitation of this work is that the game adaptation system was designed in a subject independent approach. Therefore, the adaptation rules do not take the inter-subject difference into account. From the evaluation of the game adaption system conducted in experiment 4 (chapter 5), it is clear that even a small opportunity sample consists of 2 types of players: *Challenge seekers* with high self-efficacy for physical activity and *Reward seekers* with low efficacy for physical activity. The adaption rules work well for *Reward seekers* who found mastery-based adaption enjoyable and rewarding. However, this adaptation system had a negative effect on *Challenge seekers.* Future work must address this limitation by exploring custom adaption rules for each of these player types. This approach has implications for the design of the adaption engine since it would be essential for the system to automatically classify player types from behaviour and game-based data. While this research work has identified two categories of player types in mobile AR exergames, future studies must explore this problem space to discover an appropriate player archetype and corresponding adaption principles in mobile AR exergames.

6.4 Future Work

This section outlines directions of future work following the findings in this thesis. The previous section outlined the limitations of the current work along with suggestions to overcome them. Future work needs to address these limitations to further increase the generalise-ability of the thesis findings.

It is also important to investigate more complex game design spaces. Across the research work presented in this thesis, game parameters are considered as discrete categorical variables; this results in relatively small game spaces as compared to consumer games: *AR treasure hunt* comprised of 4 game levels and *Running Chickens* comprised of 8 levels. When increasing the complexity of these game spaces, game parameters can be treated as continuous variables instead of categorical ones. Future work must investigate to what extent the player modelling techniques developed can be used for continuous game parameter representations. In addition to this, using continuous game parameters can extend the DDA system proposed in this work since difficulty adjustments would need to optimise the continuous game parameters instead of increasing or decreasing discretised difficulty levels.

In addition to this, future work must incorporate learning from physical exercise training programs into the exergame design and adaptation strategy to further improve the effectiveness of the exergame as a physical intervention. One such popular program that yields itself well to the *Running Chickens* game design is High-intensity Interval Training (HIIT). A HIIT program is generally characterised by bursts of high-intensity physical activity followed by bursts of rest intervals. However, there are no universal rules for optimal work and rest intervals for HIIT training; these factors are usually dependent on the individual's fitness. Research in exercise and physiology has found that HIIT programs can reduce cardiovascular risk factors [77, 35, 124]. Additionally, repeated periods of exercise followed by periods of recovery may be a more achievable and enjoyable alternative to high volume continuous exercise [124, 123]. Thus, the *Running Chickens* game can be extended to include principles of HIIT, where the time a player spends running around collecting chickens and time of rest between levels can be optimised for maximum physical improvement outcomes. However, research into HIIT based exergames is still in its early days. Nevertheless, early research shows that exergames that incorporate HIIT can reduce the amount of perceived exertion and can improve enjoyment, flow and motivation as compared to traditional HIIT training [132].

Additionally, the player modelling techniques investigated in this research can be used to optimise the exertion and rest interval timings to best suit the physical objectives of the player. While player mastery recognition can ensure that players feel confident while engaging with the exergame, it would be further beneficial if mobile sensors can be used to detect player fatigue. Since HIIT training requires participants to continue periods of exercise and rest until they reach physical fatigue, it would be useful for exergame systems to detect player fatigue in real-time so that the game experience can be tailored to players' level of fatigue. While some research exists in fatigue detection of drivers [78, 218], and factory works [117], this problem has not been investigated within the domain of mobile exergames.

Finally, this thesis has investigated player modelling applications in mobile AR games; it is important to extend this work to HMD based AR games. Since current AR HMDs have similar sensors to mobile devices, this thesis's player modelling pipeline can be applied to these environments. However, it is unclear how differences in the movement data captured from HMDs (instead of mobiles) will affect the findings of this research work. Additionally, player experiences of immersion and flow will change for HMD based experiences. This thesis found that it is difficult for players to navigate the physical environment due to attention switching between the device view-port and their peripheral vision; HMDs may improve player performance since the AR view-port is overlaid onto the player players field of vision.

6.5 Closing Remarks

This thesis focuses on applications of player experience modelling for DDA in mobile AR games, especially focusing on exergames aimed at making a positive impact on players' physical health. This research finds that experiences of player competence or mastery can be modelled to a reasonable degree of reliability, and these models can be used to personalise mobile AR games experiences using DDA techniques. Furthermore, this research work highlights the importance of using mobile sensor data to measure player behaviour in these types of games since this behaviour data can be used to get real-time feedback on a player's experience as they engage with the game. Finally, this research involved the design and development of 2 mobile AR games. It is worth noting that designing a video game experience is not a trivial task. It is a long process that involves many professionals such as graphic designers, sound engineers, developers and interaction designers. Due to the complexity of creating a high-quality game, the researcher encourages collaboration between experts to create a successful, engaging game. In the case of exergames, it is especially important to involve fitness experts while designing the game and the adaptation mechanics. Finally, it is worth considering the challenges in conducting research that involves machine learning techniques that require large amounts of data. In these cases, the researcher must either rely on existing data-sets or gather this data themselves - which is even more challenging for mobile AR games research due to the challenges discussed. Due to these challenges, it is important for strong collaborations between researchers and the game industry.

Bibliography

[1] Arcore. URL https://developers.google.com/ar/.

[2] Niantic. URL https://www.harrypotterwizardsunite.com/.

[3] Catch pokémon in the real world with pokémon go! URL https://www.pokemongo.com/.

[4] Vuforia: Ptc. URL https://www.vuforia.com/.

[5] Wikitude augmented reality: the world's leading cross-platform ar sdk. URL https://www.wikitude.com/.

[6] V. V. Abeele, K. Spiel, L. Nacke, D. Johnson, and K. Gerling. 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, 2020.

[7] O. Ahlqvist. Location-based games. *International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology*, pages 1–4, 2016.

[8] N. Alderman. Zombies, run! 2012.

[9] T. Althoff, R. W. White, and E. Horvitz. Influence of pokémon go on physical activity: study and implications. *Journal of medical Internet research*, 18(12):e315, 2016.

[10] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson. Emotion sensors go to school. In *AIED*, volume 200, pages 17–24. Citeseer, 2009.

[11] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias. Estimation of behavioral user state based on eye gaze and head pose—application in an e-learning environment. *Multimedia Tools and Applications*, 41(3):469–493, 2009.

[12] S. Azad, C. Saldanha, C. H. Gan, and M. O. Riedl. Mixed reality meets procedural content generation in video games. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2016.

[13] R. T. Azuma. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, 6(4):355–385, 1997. doi: 10.1162/pres.1997.6.4.355. URL `https://doi.org/10.1162/pres.1997.6.4.355`.

[14] A. Baldwin, D. Johnson, P. Wyeth, and P. Sweetser. A framework of dynamic difficulty adjustment in competitive multiplayer video games. In *2013 IEEE International Games Innovation Conference (IGIC)*, pages 16–19. IEEE, 2013.

[15] A. Bandura. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2):191, 1977.

[16] R. Banse and K. R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614, 1996.

[17] L. F. Barrett. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1):20–46, 2006.

[18] R. A. Bartle. *Designing virtual worlds*. New Riders, 2004.

[19] Y. Baveye, J.-N. Bettinelli, E. Dellandréa, L. Chen, and C. Chamaret. A large video database for computational models of induced emotion. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 13–18. IEEE, 2013.

[20] D. Bernhardt and P. Robinson. Detecting affect from non-stylised body motions. In *International conference on affective computing and intelligent interaction*, pages 59–70. Springer, 2007.

[21] D. Bertaux. From the life-history approach to the transformation of sociological practice. *Biography and society: The life history approach in the social sciences*, pages 29–45, 1981.

[22] A. Betella and P. F. Verschure. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one*, 11 (2):e0148037, 2016.

[23] P. Bhattacharyya, Y. Jo, K. Jadhav, R. Nath, and J. Hammer. Brick: A synchronous multiplayer augmented reality game for mobile phones. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, page INT009. ACM, 2019.

[24] D. Bian, J. Wade, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar. Design of a physiology-based adaptive virtual reality driving platform for individuals with asd. *ACM Transactions on Accessible Computing (TACCESS)*, 12(1):2, 2019.

[25] N. Bianchi-Berthouze. What can body movement tell us about players engagement. *Measuring Behavior12*, 2:94–97, 2012.

[26] N. Bianchi-Berthouze. Understanding the role of body movement in player engagement. *Human–Computer Interaction*, 28(1):40–75, 2013. doi: 10.1080/07370024.2012.688468. URL `https://www.tandfonline.com/doi/abs/10.1080/07370024.2012.688468`.

[27] N. Bianchi-Berthouze and A. Kleinsmith. A categorical approach to affective gesture recognition. *Connection science*, 15(4):259–269, 2003.

[28] M. J. Blanca, R. Alarcón, J. Arnau, R. Bono, and R. Bendayan. Non-normal data: Is anova still a valid option? *Psicothema*, 29(4):552–557, 2017.

[29] M. E. Bratman, D. J. Israel, and M. E. Pollack. Plans and resource-bounded practical reasoning. *Computational intelligence*, 4(3):349–355, 1988.

[30] F. Brühlmann and G.-M. Schmid. How to measure the game experience? analysis of the factor structure of two questionnaires. In *Proceedings of the 33rd annual acm conference extended abstracts on human factors in computing systems*, pages 1181–1186, 2015.

[31] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM, 2004.

[32] A. Camurri, R. Trocca, and G. Volpe. Interactive systems design: A kansei-based approach. In *Proceedings of the 2002 conference on New interfaces for musical expression*, pages 1–8. National University of Singapore, 2002.

[33] A. Camurri, I. Lagerlöf, and G. Volpe. Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International journal of human-computer studies*, 59(1-2):213–225, 2003.

[34] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe. Multimodal analysis of expressive gesture in music and dance performances. In *International gesture workshop*, pages 20–39. Springer, 2003.

[35] V. Carson, R. Rinaldi, B. Torrance, K. Maximova, G. Ball, S. Majumdar, R. Plotnikoff, P. Veugelers, N. Boule, P. Wozny, et al. Vigorous physical activity and longitudinal associations with cardiometabolic risk factors in youth. *International journal of obesity*, 38(1):16–21, 2014.

[36] J. Chen. Flow in games (and everything else). *Communications of the ACM*, 50(4):31–34, 2007.

[37] M. Chen, Q. Tang, S. Xu, P. Leng, and Z. Pan. Design and evaluation of an augmented reality-based exergame system to reduce fall risk in the elderly. *International journal of environmental research and public health*, 17(19):7208, 2020.

[38] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[39] C. Clavel, J. Plessier, J.-C. Martin, L. Ach, and B. Morel. Combining facial and postural expressions of emotions in a virtual character. In *International Workshop on Intelligent Virtual Agents*, pages 287–300. Springer, 2009.

[40] T. Cole, P. Cairns, and M. Gillies. Emotional and functional challenge in core and avant-garde games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, pages 121–126. ACM, 2015.

[41] C. Conati and H. Maclaren. Modeling user affect from causes and effects. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 4–15. Springer, 2009.

[42] C. Conati, A. Gertner, and K. Vanlehn. Using bayesian networks to manage uncertainty in student modeling. *User modeling and user-adapted interaction*, 12(4):371–417, 2002.

[43] B. Cowley, D. Charles, M. Black, and R. Hickey. Toward an understanding of flow in video games. *Computers in Entertainment (CIE)*, 6(2):1–27, 2008.

[44] M. Csikszentmihalyi. Flow and the psychology of discovery and invention. *HarperPerennial, New York*, 39, 1997.

[45] M. Csikszentmihalyi. *Beyond boredom and anxiety.* Jossey-Bass, 2000.

[46] M. Csikszentmihalyi. Toward a psychology of optimal experience. In *Flow and the foundations of positive psychology*, pages 209–226. Springer, 2014.

[47] L. Cui, S. Li, and T. Zhu. Emotion detection from natural walking. In *International Conference on Human Centered Computing*, pages 23–33. Springer, 2016.

[48] E. L. Deci and R. M. Ryan. Intrinsic motivation. *The corsini encyclopedia of psychology*, pages 1–2, 2010.

[49] A. Denisova and P. Cairns. Adaptation in digital games: the effect of challenge adjustment on player performance and experience. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, pages 97–101. ACM, 2015.

[50] A. Denisova and P. Cairns. Player experience and deceptive expectations of difficulty adaptation in digital games. *Entertainment Computing*, 29: 56–68, 2019.

[51] J. Dennerlein, T. Becker, P. Johnson, C. Reynolds, and R. W. Picard. Frustrating computer users increases exposure to physical factors. In *Proceedings of the international Ergonomics association, Seoul, Korea*, 2003.

[52] K. Desai, K. Bahirat, S. Ramalingam, B. Prabhakaran, T. Annaswamy, and U. E. Makris. Augmented reality-based exergames for rehabilitation. In *Proceedings of the 7th International Conference on Multimedia Systems*, pages 1–10, 2016.

[53] S. D'Mello and A. Graesser. Automatic detection of learner's affect from gross body language. *Applied Artificial Intelligence*, 23(2):123–150, 2009.

[54] L. Dormehl. 8 major milestones in the brief history of virtual reality, 2018. URL `https://www.digitaltrends.com/cool-tech/history-of-virtual-reality/`.

[55] P. Dourish. *Where the action is.* MIT press Cambridge, 2001.

[56] A. Drachen, A. Canossa, and G. N. Yannakakis. Player modeling using self-organization in tomb raider: Underworld. In *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*, pages 1–8. IEEE, 2009.

[57] A. Drachen, M. S. El-Nasr, and A. Canossa. *Game Analytics: Maximizing the Value of Player Data*. Springer, 2013.

[58] P. Ekman and W. V. Friesen. Head and body cues in the judgment of emotion: A reformulation. *Perceptual and motor skills*, 24(3 PT 1):711–724, 1967.

[59] L. A. Feldman. Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of personality and social psychology*, 69(1):153, 1995.

[60] C.-N. Fiechter and S. Rogers. Learning subjective functions with large margins. In *ICML*, pages 287–294, 2000.

[61] A. Fortin-Côté, C. Chamberland, M. Parent, S. Tremblay, P. Jackson, N. Beaudoin-Gagnon, A. Campeau-Lecours, J. Bergeron-Boucher, and L. Lefebvre. Predicting video game players' fun from physiological and behavioural data. In *Future of Information and Communication Conference*, pages 479–495. Springer, 2018.

[62] A. Fortin-Côté, N. Beaudoin-Gagnon, C. Chamberland, F. Desbiens, L. Lefebvre, J. Bergeron, A. Campeau-Lecours, S. Tremblay, and P. L. Jackson. Funii: The physio-behavioural adaptive video game. In *International Conference on Human-Computer Interaction*, pages 14–28. Springer, 2019.

[63] N. H. Frijda. *The emotions*. Cambridge University Press, 1986.

[64] J. Frommel, F. Fischbach, K. Rogers, and M. Weber. Emotion-based dynamic difficulty adjustment using parameterized difficulty and self-reports of emotion. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, pages 163–171. ACM, 2018.

[65] D. Gabana Arellano. *Games 4 VRains: affective gaming for working memory training in virtual reality*. PhD thesis, Queen Mary University of London, 2019.

[66] M. Georgeff, B. Pell, M. Pollack, M. Tambe, and M. Wooldridge. The belief-desire-intention model of agency. In *International workshop on agent theories, architectures, and languages*, pages 1–10. Springer, 1998.

[67] K. M. Gerling, M. Miller, R. L. Mandryk, M. V. Birk, and J. D. Smeddinck. Effects of balancing for physical abilities on player performance, experience and self-esteem in exergames. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 2201–2210. ACM, 2014.

[68] K. Gilleade, A. Dix, and J. Allanson. Affective videogames and modes of affective gaming: assist me, challenge me, emote me. *DiGRA 2005: Changing Views–Worlds in Play.*, 2005.

[69] K. M. Gilleade and A. Dix. Using frustration in the design of adaptive videogames. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pages 228–232. ACM, 2004.

[70] L. Gong, T. Wang, C. Wang, F. Liu, F. Zhang, and X. Yu. Recognizing affect from non-stylized body motion using shape of gaussian descriptors. In *Proceedings of the 2010 ACM symposium on applied computing*, pages 1203–1206. ACM, 2010.

[71] J. F. Grafsgaard, K. E. Boyer, and J. C. Lester. Predicting facial indicators of confusion with hidden markov models. In *International Conference on Affective computing and intelligent interaction*, pages 97–106. Springer, 2011.

[72] J. Gratch and S. Marsella. Evaluating a computational model of emotion. *Autonomous Agents and Multi-Agent Systems*, 11(1):23–43, 2005.

[73] G. Guo, J. Segal, H. Zhang, and W. Xu. Armove: A smartphone augmented reality exergaming system for upper and lower extremities stroke rehabilitation: Demo abstract. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, pages 384–385, 2019.

[74] J. Gutman. A means-end chain model based on consumer categorization processes. *Journal of marketing*, 46(2):60–72, 1982.

[75] K.-C. Hao and L.-C. Lee. The development and evaluation of an educational game integrating augmented reality, arcs model, and types of games for english experiment learning: an analysis of learning. *Interactive Learning Environments*, pages 1–14, 2019.

[76] Y. Hao, S. He, J. Wang, X. Liu, W. Huang, et al. Dynamic difficulty adjustment of game ai by mcts for the game pac-man. In *2010 sixth international conference on natural computation*, volume 8, pages 3918–3922. IEEE, 2010.

[77] J. Hay, K. Maximova, A. Durksen, V. Carson, R. L. Rinaldi, B. Torrance, G. D. Ball, S. R. Majumdar, R. C. Plotnikoff, P. Veugelers, et al. Physical activity intensity and cardiometabolic risk in youth. *Archives of pediatrics & adolescent medicine*, 166(11):1022–1029, 2012.

[78] J. He, S. Roberson, B. Fields, J. Peng, S. Cielocha, and J. Coltea. Fatigue detection using smartphones. *Journal of Ergonomics*, 3(03):1–7, 2013.

[79] A. Henrysson, M. Billinghurst, and M. Ollila. Ar tennis. In *ACM SIGGRAPH 2006 Emerging Technologies*, SIGGRAPH '06, New York, NY, USA, 2006. ACM. ISBN 1-59593-364-6. doi: 10.1145/1179133.1179135. URL `http://doi.acm.org/10.1145/1179133.1179135`.

[80] N. Hocine, A. Gouaïch, S. A. Cerri, D. Mottet, J. Froger, and I. Laffont. Adaptation in serious games for upper-limb rehabilitation: an approach to improve training outcomes. *User Modeling and User-Adapted Interaction*, 25(1):65–98, 2015.

[81] C. Holmgård, G. N. Yannakakis, K.-I. Karstoft, and H. S. Andersen. Stress detection for ptsd via the startlemart game. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 523–528. IEEE, 2013.

[82] C. Holmgård, G. N. Yannakakis, H. P. Martinez, and K.-I. Karstoft. To rank or to classify? annotating stress for reliable ptsd profiling. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 719–725. IEEE, 2015.

[83] C. Holmgård, G. N. Yannakakis, H. P. Martínez, K.-I. Karstoft, and H. S. Andersen. Multimodal ptsd characterization via the startlemart game. *Journal on Multimodal User Interfaces*, 9(1):3–15, 2015.

[84] J. Höysniemi, P. Hämäläinen, L. Turkki, and T. Rouvi. Children's intuitive gestures in vision-based action games. *Communications of the ACM*, 48(1):44–50, 2005.

[85] R. Hunicke. The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pages 429–433. ACM, 2005.

[86] W. IJsselsteijn, K. Poels, and Y. A. De Kort. The game experience questionnaire: Development of a self-report measure to assess player experiences of digital games. *TU Eindhoven, Eindhoven, The Netherlands*, page 42, 2008.

[87] A. Inc. Augmented reality. URL `https://developer.apple.com/augmented-reality/`.

[88] K. Isbister and C. DiMauro. Waggling the form baton: Analyzing body-movement-based design patterns in nintendo wii games, toward innovation

of new possibilities for social and emotional experience. In *Whole body interaction*, pages 63–73. Springer, 2011.

[89] K. Isbister and N. Schaffer. *Game usability: Advancing the player experience*. CRC press, 2008.

[90] K. Isbister, U. Schwekendiek, and J. Frye. Wriggle: an exploration of emotional and social effects of movement. In *CHI'11 extended abstracts on human factors in computing systems*, pages 1885–1890. ACM, 2011.

[91] M. Ishihara, S. Ito, R. Ishii, T. Harada, and R. Thawonmas. Monte-carlo tree search for implementation of dynamic difficulty adjustment fighting game ais having believable behaviors. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE, 2018.

[92] S. A. Jackson and R. C. Eklund. Assessing flow in physical activity: the flow state scale-2 and dispositional flow scale-2. *Journal of sport & exercise psychology*, 24(2), 2002.

[93] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton. Measuring and defining the experience of immersion in games. *International journal of human-computer studies*, 66(9):641–661, 2008.

[94] D. Johnson, M. J. Gardner, and R. Perry. Validation of two game experience scales: the player experience of need satisfaction (pens) and game experience questionnaire (geq). *International Journal of Human-Computer Studies*, 118:38–46, 2018.

[95] T. Johnstone and K. R. Scherer. Vocal communication of emotion. *Handbook of emotions*, 2:220–235, 2000.

[96] P. N. Juslin and K. R. Scherer. Vocal expression of affect. *The new handbook of methods in nonverbal behavior research*, pages 65–135, 2005.

[97] D. Kahneman and J. Riis. Living, and thinking about it: Two perspectives on life. *The science of well-being*, 1:285–304, 2005.

[98] T. Kannetis and A. Potamianos. Towards adapting fantasy, curiosity and challenge in multimodal dialogue systems for preschoolers. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 39–46. ACM, 2009.

[99] A. Kapoor, W. Burleson, and R. W. Picard. Automatic prediction of frustration. *International journal of human-computer studies*, 65(8):724–736, 2007.

[100] D. Kappen, P. Mirza-Babaei, and L. Nacke. Gamification of older adults' physical activity: an eight-week study. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[101] M. Karg, K. Kuhnlenz, and M. Buss. Recognition of affect based on gait patterns. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(4):1050–1061, 2010.

[102] T. A. Kato, A. R. Teo, M. Tateno, M. Watabe, H. Kubo, and S. Kanba. Can pokémon go rescue shut-ins (hikikomori) from their isolated world. *Psychiatry Clin Neurosci*, 71(1):75–76, 2017.

[103] M. Kegeleers, S. Miglani, G. M. Reichert, N. Z. Salamon, J. T. Balint, S. G. Lukosch, and R. Bidarra. Star: superhuman training in augmented reality. In *Proceedings of the First Superhuman Sports Design Challenge: First International Symposium on Amplifying Capabilities and Competing in Mixed Realities*, pages 1–6, 2018.

[104] J. M. Keller. *Motivational design for learning and performance: The ARCS model approach.* Springer Science & Business Media, 2009.

[105] D. J. Ketchen and C. L. Shook. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, 17(6):441–458, 1996.

[106] A. Kleinsmith and N. Bianchi-Berthouze. Recognizing affective dimensions from body posture. In *International conference on affective computing and intelligent interaction*, pages 48–58. Springer, 2007.

[107] A. Kleinsmith, P. R. De Silva, and N. Bianchi-Berthouze. Grounding affective dimensions into posture features. In *International Conference on Affective Computing and Intelligent Interaction*, pages 263–270. Springer, 2005.

[108] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed. Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(4):1027–1038, 2011.

[109] J. Koivisto, A. Malik, B. Gurkan, and J. Hamari. Getting healthy by catching them all: A study on the relationship between player orientations and perceived health benefits in an augmented reality game. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[110] R. Koster. *Theory of fun for game design.* " O'Reilly Media, Inc.", 2013.

[111] C. D. Kounavis, A. E. Kasimati, and E. D. Zamani. Enhancing the tourism experience through mobile augmented reality: Challenges and prospects. *International Journal of Engineering Business Management*, 4:10, 2012.

[112] T. H. Laine and H. J. Suk. Designing mobile augmented reality exergames. *Games and culture*, 11(5):548–580, 2016.

[113] E. L.-C. Law, F. Brühlmann, and E. D. Mekler. Systematic review and validation of the game experience questionnaire (geq)-implications for citation and reporting practice. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, pages 257–270, 2018.

[114] N. Lazzaro. Why we play: affect and the fun of games. *Human-computer interaction: Designing for diverse users and domains*, 155, 2009.

[115] I. Levstein and L. J. Justice. Csi4fs®: A markerless augmented reality game–a novel approach to crime scene investigation training. In *Emerging Technologies in Virtual Learning Environments*, pages 238–257. IGI Global, 2019.

[116] B. Li, C. Zhu, S. Li, and T. Zhu. Identifying emotions from non-contact gaits information based on microsoft kinects. *IEEE Transactions on Affective Computing*, 9(4):585–591, 2016.

[117] P. Li, R. Meziane, M. J.-D. Otis, H. Ezzaidi, and P. Cardou. A smart safety helmet using imu and eeg sensors for worker fatigue detection. In *2014 IEEE International Symposium on Robotic and Sensors Environments (ROSE) Proceedings*, pages 55–60. IEEE, 2014.

[118] R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[119] R. W. Lindeman, G. Lee, L. Beattie, H. Gamper, R. Pathinarupothi, and A. Akhilesh. Geoboids: A mobile ar application for exergaming. In *2012 IEEE International Symposium on Mixed and Augmented Reality-Arts, Media, and Humanities (ISMAR-AMH)*, pages 93–94. IEEE, 2012.

[120] S. E. Lindley, J. Le Couteur, and N. L. Berthouze. Stirring up experience through movement in game play: effects on engagement and social behaviour. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 511–514. ACM, 2008.

[121] J. Linowes and K. Babilinski. *Augmented Reality for Developers: Build practical augmented reality applications with Unity, ARCore, ARKit, and Vuforia*. Packt Publishing Ltd, 2017.

[122] C. Liu, P. Agrawal, N. Sarkar, and S. Chen. Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. *International Journal of Human-Computer Interaction*, 25(6):506–529, 2009.

[123] G. R. Logan, N. Harris, S. Duncan, and G. Schofield. A review of adolescent high-intensity interval training. *Sports Medicine*, 44(8):1071–1085, 2014.

[124] G. R. M. Logan, N. Harris, S. Duncan, L. D. Plank, F. Merien, and G. Schofield. Low-active male adolescents: a dose response to high-intensity interval training. *Medicine & Science in Sports & Exercise*, 48 (3):481–490, 2016.

[125] P. Loveday and J. Burgess. Flow and pokémon go: the contribution of game level, playing alone, and nostalgia to the flow state. *E-Journal of Social & Behavioural Research in Business*, 8(2):16–28, 2017.

[126] A. Macvean and J. Robertson. Understanding exergame users' physical activity, motivation and behavior over time. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1251–1260, 2013.

[127] T. W. Malone. What makes things fun to learn? heuristics for designing instructional computer games. In *Proceedings of the 3rd ACM SIGSMALL symposium and the first SIGPC symposium on Small systems*, pages 162–169. ACM, 1980.

[128] R. L. Mandryk and M. S. Atkins. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International journal of human-computer studies*, 65(4):329–347, 2007.

[129] R. L. Mandryk, K. M. Inkpen, and T. W. Calvert. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & information technology*, 25(2):141–158, 2006.

[130] J. Marshall, S. Benford, S. Pijnappel, et al. Expanding exertion gaming. *International Journal of Human-Computer Studies*, 90:1–13, 2016.

[131] A. L. Martin-Niedecken and U. Götz. Go with the dual flow: evaluating the psychophysiological adaptive fitness game environment "plunder planet". In *Joint International Conference on Serious Games*, pages 32–43. Springer, 2017.

[132] A. L. Martin-Niedecken, A. Mahrer, K. Rogers, E. D. de Bruin, and A. Schättin. “hiit” the exercube: Comparing the effectiveness of functional high-intensity interval training in conventional vs. exergame-based training. *Frontiers in Computer Science*, 2:33, 2020.

[133] H. P. Martínez and G. N. Yannakakis. Genetic search feature selection for affective modeling: a case study on reported preferences. In *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, pages 15–20. ACM, 2010.

[134] H. P. Martínez and G. N. Yannakakis. Mining multimodal sequential patterns: a case study on affect detection. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 3–10. ACM, 2011.

[135] H. P. Martinez, G. N. Yannakakis, and J. Hallam. Don’t classify ratings of affect; rank them! *IEEE transactions on affective computing*, 5(3): 314–326, 2014.

[136] E. McAuley, T. Duncan, and V. V. Tammen. Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport*, 60 (1):48–58, 1989.

[137] D. J. McCaffery and J. Finney. The need for real time consistency management in p2p mobile gaming environments. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pages 203–211. ACM, 2004.

[138] M. McCartney. Margaret mccartney: game on for pokémon go. *BMJ: British Medical Journal (Online)*, 354, 2016.

[139] D. Melhart, A. Azadvar, A. Canossa, A. Liapis, and G. N. Yannakakis. Your gameplay says it all: Modelling motivation in tom clancy’s the division. *arXiv preprint arXiv:1902.00040*, 2019.

[140] D. Melhart, A. Liapis, and G. N. Yannakakis. Towards general models of player experience: A study within genres. In *2021 IEEE Conference on Games (CoG)*, pages 1–8. IEEE, 2021.

[141] A. Melzer, I. Derks, J. Heydekorn, and G. Steffgen. Click or strike: Realistic versus standard game controls in violent video games and their effects on aggression. In *International Conference on Entertainment Computing*, pages 171–182. Springer, 2010.

[142] A. S. Mendoza-Vasconez, B. Marquez, T. J. Benitez, and B. H. Marcus. Psychometrics of the self-efficacy for physical activity scale among a latina women sample. *BMC Public Health*, 18(1):1–10, 2018.

[143] P. Milgram and F. Kishino. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77(12):1321–1329, 1994.

[144] T. Miron-Shatz, A. Stone, and D. Kahneman. Memories of yesterday's emotions: Does the valence of experience affect the memory-experience gap? *Emotion*, 9(6):885, 2009.

[145] J. D. Morris. Observations: Sam: the self-assessment manikin; an efficient cross-cultural measurement of emotional response. *Journal of advertising research*, 35(6):63–68, 1995.

[146] F. Muller and N. Berthouze. Evaluating exertion games experiences from investigating movement-based games. Springer, 2010.

[147] A. Mulloni, D. Wagner, and D. Schmalstieg. Mobility and social interaction as core gameplay elements in multi-player augmented reality. In *Proceedings of the 3rd international conference on Digital Interactive Media in Entertainment and Arts*, pages 472–478. ACM, 2008.

[148] J. Munoz, G. N. Yannakakis, F. Mulvey, D. W. Hansen, G. Gutierrez, and A. Sanchis. Towards gaze-controlled platform games. In *2011 IEEE Conference on Computational Intelligence and Games (CIG'11)*, pages 47–54. IEEE, 2011.

[149] L. Nacke and C. A. Lindley. Flow and immersion in first-person shooters: measuring the player's gameplay experience. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share*, pages 81–88. ACM, 2008.

[150] J. Nielsen. Usability 101: Introduction to usability.

[151] J. Nijhar, N. Bianchi-Berthouze, and G. Boguslawski. Does movement recognition precision affect the player experience in exertion games? In *International Conference on Intelligent Technologies for interactive entertainment*, pages 73–82. Springer, 2011.

[152] A. Nijholt. Bci for games: A 'state of the art'survey. In *International Conference on Entertainment Computing*, pages 225–228. Springer, 2008.

[153] T. Nilsen, S. Linton, and J. Looser. Motivations for augmented reality gaming. *Proceedings of FUSE*, 4:86–93, 2004.

[154] A. Nishchyk, W. Geentjens, A. Medina, M. Klein, and W. Chen. An augmented reality game for helping elderly to perform physical exercises at home. In *International Conference on Computers Helping People with Special Needs*, pages 233–241. Springer, 2020.

[155] M. Nunes, L. Nedel, and V. Roesler. Motivating people to perform better in exergames: Competition in virtual environments. In *Proceedings of the 29th annual ACM symposium on applied computing*, pages 970–975, 2014.

[156] M. Orozco, J. Silva, A. El Saddik, and E. Petriu. The role of haptics in games. In *Haptics rendering and applications*. IntechOpen, 2012.

[157] A. Ortony, G. L. Clore, and A. Collins. The cognitive structure of emotions. *Cambridge Uni*, 1988.

[158] J. Paavilainen, H. Korhonen, K. Alha, J. Stenros, E. Koskinen, and F. Mayra. The pokémon go experience: A location-based augmented reality mobile game goes mainstream. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 2493–2498, 2017.

[159] V. Paelke, C. Reimann, and D. Stichling. Foot-based mobile interaction with games. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pages 321–324. ACM, 2004.

[160] F. Palero, C. Ramirez-Atencia, and D. Camacho. Online gamers classification using k-means. In *Intelligent Distributed Computing VIII*, pages 201–208. Springer, 2015.

[161] B. Pang, L. Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[162] H. Park, J.-I. Park, U.-M. Kim, and W. Woo. Emotion recognition from dance image sequences using contour approximation. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 547–555. Springer, 2004.

[163] T. D. Parsons and J. L. Reinebold. Adaptive virtual environments for neuropsychological assessment in serious games. *IEEE Transactions on Consumer Electronics*, 58(2):197–204, 2012.

[164] M. Pasch, N. Bianchi-Berthouze, B. van Dijk, and A. Nijholt. Movement-based sports video games: Investigating motivation and gaming experience. *Entertainment Computing*, 1(2):49–61, 2009.

[165] H. M. Paterson, F. E. Pollick, and A. J. Sanford. The role of velocity in affect discrimination. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23, 2001.

[166] C. Pedersen, J. Togelius, and G. N. Yannakakis. Modeling player experience in super mario bros. In *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*, pages 132–139. IEEE, 2009.

[167] C. Pedersen, J. Togelius, and G. N. Yannakakis. Modeling player experience for content creation. *IEEE Transactions on Computational Intelligence and AI in Games*, 2(1):54–67, 2010.

[168] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[169] L. Perumal. Quaternion and its application in rotation using sets of regions. *International Journal of Engineering and Technology Innovation*, 1(1):35, 2011.

[170] M. G. Petersen. Aesthetics of interaction–a pragmatist perspective. In *Aesthetic Approaches to Human-Computer Interaction, Workshop, NordiCHI 2004, 24. okt., Tampere, Finland*, pages 31–32. Department of Computer Science, Aarhus University, 2005.

[171] M. G. Petersen, O. S. Iversen, P. G. Krogh, and M. Ludvigsen. Aesthetic interaction: a pragmatist's aesthetics of interactive systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 269–276. ACM, 2004.

[172] R. W. Picard. Affective computing-mit media laboratory perceptual computing section technical report no. 321. *Cambridge, MA*, 2139, 1995.

[173] R. W. Picard and W. Rosalind. Toward agents that recognize emotion. *VIVEK-BOMBAY-*, 13(1):3–13, 2000.

[174] J. Plass, B. Homer, S. Pawar, C. Brenner, and A. MacNamara. The effect of adaptive difficulty adjustment on the effectiveness of a game to develop executive function skills for learners of different ages. *Cognitive Development*, 49:56–67, 2019.

[175] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford. Perceiving affect from arm movement. *Cognition*, 82(2):B51–B61, 2001.

[176] F. E. Pollick, V. Lestou, J. Ryu, and S.-B. Cho. Estimating the efficiency of recognizing gender and affect from biological motion. *Vision research*, 42(20):2345–2355, 2002.

[177] J. Posner, J. A. Russell, and B. S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3): 715–734, 2005.

[178] S. J. Preece, J. Y. Goulermas, L. P. Kenney, and D. Howard. A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *IEEE Transactions on Biomedical Engineering*, 56(3):871–879, 2008.

[179] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[180] K. Procci and C. Bowers. An examination of flow and immersion in games. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 55, pages 2183–2187. SAGE Publications Sage CA: Los Angeles, CA, 2011.

[181] A. K. Przybylski, C. S. Rigby, and R. M. Ryan. A motivational model of video game engagement. *Review of general psychology*, 14(2):154–166, 2010.

[182] J. C. Quiroz, M. H. Yong, and E. Geangu. Emotion-recognition using smart watch accelerometer data: preliminary findings. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 805–812. ACM, 2017.

[183] N. Ravaja, T. Saari, M. Salminen, J. Laarni, and K. Kallinen. Phasic emotional reactions to video game events: A psychophysiological investigation. *Media psychology*, 8(4):343–367, 2006.

[184] C. Reimann. Kick-real, a mobile mixed reality game. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pages 387–387. ACM, 2005.

[185] C. Reimann and V. Paelke. Adaptive mixed reality games. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pages 302–305. ACM, 2005.

[186] T. J. Reynolds and J. Gutman. Laddering theory, method, analysis, and interpretation. *Journal of advertising research*, 28(1):11–31, 1988.

[187] J. Robison, S. McQuiggan, and J. Lester. Evaluating the consequences of affective feedback in intelligent tutoring systems. In *2009 3rd international conference on affective computing and intelligent interaction and workshops*, pages 1–6. IEEE, 2009.

[188] R. M. Ryan and E. L. Deci. Basic psychological needs theory: Satisfaction and frustration of autonomy, competence, and relatedness in relation to psychological well-being and full functioning. *Self-determination theory, EL Deci & RM Ryan*, pages 239–271, 2017.

[189] R. M. Ryan, C. S. Rigby, and A. Przybylski. The motivational pull of video games: A self-determination theory approach. *Motivation and emotion*, 30(4):344–360, 2006.

[190] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 305–312. ACM, 2011.

[191] M. Santoso. Markerless augmented reality technology for real-space basketball simulation. In *Consumer Electronics (ICCE), 2018 IEEE International Conference on*, pages 1–3. IEEE, 2018.

[192] T. Sapiński, D. Kamińska, A. Pelikant, and G. Anbarjafari. Emotion recognition from skeletal movements. *Entropy*, 21(7):646, 2019.

[193] N. Savva and N. Bianchi-Berthouze. Automatic recognition of affective body movement in a video game scenario. In *International Conference on Intelligent Technologies for interactive entertainment*, pages 149–159. Springer, 2011.

[194] N. Savva, A. Scarinzi, and N. Bianchi-Berthouze. Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience. *IEEE Transactions on Computational Intelligence and AI in games*, 4(3):199–212, 2012.

[195] J. Schell. *The Art of Game Design: A book of lenses.* AK Peters/CRC Press, 2014.

[196] K. R. Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[197] K. R. Scherer, A. Schorr, and T. Johnstone. *Appraisal processes in emotion: Theory, methods, research.* Oxford University Press, 2001.

[198] N. Shaker, G. Yannakakis, and J. Togelius. Towards automatic personalized content generation for platform games. In *Sixth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2010.

[199] N. Shaker, G. N. Yannakakis, J. Togelius, M. Nicolau, and M. O'neill. Evolving personalized content for super mario bros using grammatical evolution. In *AIIDE*, 2012.

[200] L. A. Shaw, B. C. Wuensche, C. Lutteroth, J. Buckley, and P. Corballis. Evaluating sensory feedback for immersion in exergames. In *Proceedings of the Australasian Computer Science Week Multiconference*, pages 1–6, 2017.

[201] B. F. Skinner. *The behavior of organisms: An experimental analysis.* BF Skinner Foundation, 1990.

[202] A. M. Smith, C. Lewis, K. Hullett, G. Smith, and A. Sullivan. An inclusive taxonomy of player modeling. *University of California, Santa Cruz, Tech. Rep. UCSC-SOE-11-13*, 2011.

[203] C. Southerton. Zombies, run! rethinking immersion in light of nontraditional gaming contexts. In *Transmedia practice: A collective approach*, pages 131–141. Brill, 2014.

[204] A. E. Staiano and S. L. Calvert. Exergames for physical education courses: Physical, social, and cognitive benefits. *Child development perspectives*, 5 (2):93–98, 2011.

[205] M. Stephenson and J. Renz. Agent-based adaptive level generation for dynamic difficulty adjustment in angry birds. *arXiv preprint arXiv:1902.02518*, 2019.

[206] A. L. Strauss. *Qualitative analysis for social scientists.* Cambridge university press, 1987.

[207] H. Sung and C.-M. Wang. X-portal: Mixed reality for body engagement in new first person game experience. In *2018 IEEE International Conference on Applied System Invention (ICASI)*, pages 1073–1074. IEEE, 2018.

[208] P. Sweetser and P. Wyeth. Gameflow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, 3(3):3–3, 2005.

[209] P. Sweetser, D. M. Johnson, and P. Wyeth. Revisiting the gameflow model with detailed heuristics. *Journal: Creative Technologies*, 2012(3), 2012.

[210] C. T. Tan and D. Soh. Augmented reality games: A review. *Proceedings of Gameon-Arabia, Eurosis*, 2010.

[211] U. Technologies. Unity. URL `https://unity3d.com/`.

[212] B. H. Thomas. A survey of visual, mixed, and augmented reality gaming. *Computers in Entertainment (CIE)*, 10(1):3, 2012.

[213] T. Tijs, D. Brokken, and W. IJsselsteijn. Creating an emotionally adaptive game. In *International Conference on Entertainment Computing*, pages 122–133. Springer, 2008.

[214] T. J. Tijs, D. Brokken, and W. A. Ijsselsteijn. Dynamic game balancing by recognizing affect. In *International Conference on Fun and Games*, pages 88–93. Springer, 2008.

[215] J. Togelius, R. De Nardi, and S. M. Lucas. Making racing fun through player modeling and track evolution. 2006.

[216] S. Tognetti, M. Garbarino, A. Bonarini, and M. Matteucci. Modeling enjoyment preference from physiological responses in a car racing game. In *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games*, pages 321–328. IEEE, 2010.

[217] S.-p. Tsai. Augmented reality enhancing place satisfaction for heritage tourism marketing. *Current Issues in Tourism*, pages 1–6, 2019.

[218] W. Tu, L. Wei, W. Hu, Z. Sheng, H. Nicanfar, X. Hu, E. C.-H. Ngai, and V. C. Leung. A survey on mobile sensing based mood-fatigue detection for drivers. In *Smart City 360°*, pages 3–15. Springer, 2016.

[219] W. M. van den Hoogen, W. A. IJsselsteijn, and Y. A. de Kort. Exploring behavioral expressions of player experience in digital games. In *Proceedings of the workshop on Facial and Bodily Expression for Control and Adaptation of Games ECAG*, pages 11–19, 2008.

[220] V. Vanden Abeele, L. E. Nacke, E. D. Mekler, and D. Johnson. Design and preliminary validation of the player experience inventory. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, pages 335–341, 2016.

[221] R. Vicencio-Moreira, R. L. Mandryk, and C. Gutwin. Now you can compete with anyone: Balancing players of different skill levels in a first-person shooter game. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2255–2264. ACM, 2015.

[222] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[223] J. Visser and J. M. Keller. The clinical use of motivational messages: An inquiry into the validity of the arcs model of motivational design. *Instructional science*, 19(6):467–500, 1990.

[224] T. Vogt and E. André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *2005 IEEE International Conference on Multimedia and Expo*, pages 474–477. IEEE, 2005.

[225] G. S. Von Itzstein, M. Billinghurst, R. T. Smith, and B. H. Thomas. Augmented reality entertainment: Taking gaming out of the box. *Encyclopedia of Computer Graphics and Games*, pages 1–9, 2017.

[226] S. Vorwerg, O. Stamm, and U. Müller-Werdan. Movement training in augmented reality for older adults with essential hypertension: a qualitative study determining training requirements. *Games for Health Journal*, 9(6):436–445, 2020.

[227] D. Watson, L. A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[228] A. Webster. Pokémon go spurred an amazing era that continues with sword and shield, 2019. URL `https://www.theverge.com/2019/2/28/18243332/pokemon-go-sword-shield-franchise-history-niantic-nintendo-switch`.

[229] J. Westlake. Exploring the potential of using augmented reality and virtual reality for stem education. In *Learning Technology for Education Challenges: 8th International Workshop, LTEC 2019, Zamora, Spain, July 15–18, 2019, Proceedings*, page 36. Springer, 2019.

[230] E. Witkowski. Running with zombies: Capturing new worlds through movement and visibility practices with zombies, run! *Games and Culture*, 13(2):153–173, 2018.

[231] H.-K. Wu, S. W.-Y. Lee, H.-Y. Chang, and J.-C. Liang. Current status, opportunities and challenges of augmented reality in education. *Computers & education*, 62:41–49, 2013.

[232] Y.-H. Yang and H. H. Chen. Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):762–774, 2010.

[233] G. N. Yannakakis. Preference learning for affective modeling. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6. IEEE, 2009.

[234] G. N. Yannakakis and J. Hallam. Evolving opponents for interesting interactive computer games. *From animals to animats*, 8:499–508, 2004.

[235] G. N. Yannakakis and J. Hallam. Modeling and augmenting game entertainment through challenge and curiosity. *International Journal on Artificial Intelligence Tools*, 16(06):981–999, 2007.

[236] G. N. Yannakakis and J. Hallam. Entertainment modeling through physiology in physical play. *International Journal of Human-Computer Studies*, 66(10):741 – 755, 2008. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2008.06.004. URL `http://www.sciencedirect.com/science/article/pii/S107158190800075X`.

[237] G. N. Yannakakis and J. Hallam. Ranking vs. preference: a comparative study of self-reporting. In *International Conference on Affective Computing and Intelligent Interaction*, pages 437–446. Springer, 2011.

[238] G. N. Yannakakis and H. P. Martínez. Ratings are overrated! *Frontiers in ICT*, 2:13, 2015.

[239] G. N. Yannakakis and J. Togelius. Experience-driven procedural content generation. *IEEE Transactions on Affective Computing*, 2(3):147–161, 2011.

[240] G. N. Yannakakis and J. Togelius. *Artificial intelligence and games*, volume 2. Springer, 2018.

[241] G. N. Yannakakis, H. H. Lund, and J. Hallam. Modeling children's entertainment in the playware playground. In *2006 IEEE Symposium on Computational Intelligence and Games*, pages 134–141. IEEE, 2006.

[242] G. N. Yannakakis, H. P. Martínez, and A. Jhala. Towards affective camera control in games. *User Modeling and User-Adapted Interaction*, 20(4): 313–340, 2010.

[243] G. N. Yannakakis, P. Spronck, D. Loiacono, and E. André. Player modeling. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013.

[244] G. N. Yannakakis, R. Cowie, and C. Busso. The ordinal nature of emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 248–255. IEEE, 2017.

[245] M.-H. Ying and K.-T. Yang. A game-based learning system using the arcs model and fuzzy logic. *JSW*, 8(9):2155–2162, 2013.

[246] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2008.

[247] Z. Zhang, Y. Song, L. Cui, X. Liu, and T. Zhu. Emotion recognition based on customized smart bracelet with built-in accelerometer. *PeerJ*, 4:e2258, 2016.

[248] A. Zook, S. Lee-Urban, M. O. Riedl, H. K. Holden, R. A. Sottilare, and K. W. Brawner. Automated scenario generation: toward tailored and optimized military training in virtual environments. In *Proceedings of the international conference on the foundations of digital games*, pages 164–171. ACM, 2012.

[249] A. E. Zook and M. O. Riedl. A temporal data-driven player model for dynamic difficulty adjustment. In *Eighth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2012.

Appendix A

Experiment-1 materiel

This appendix contains materials used for experiment-1 reported in chapter 3.

A.1 Ethics Approval

Queen Mary
University of London

Queen Mary, University of London
Room W117
Queen's Building
Queen Mary University of London
Mile End Road
London E1 4NS

Queen Mary Ethics of Research Committee
Hazel Covill
Research Ethics Administrator
Tel: +44 (0) 20 7882 7915
Email: h.covill@qmul.ac.uk

c/o Dr Laurissa Tokarchuk
CS 302, Peter Landin Building
School of Electronic Engineering
& Computer Science
Queen Mary University of London
Mile End
London

17th August 2018

To Whom It May Concern:

Re: QMREC1935 - Modelling Player Preferences in AR mobile games.

I can confirm that Vivek Warriar has completed a Research Ethics Questionnaire with regard to the above research.

The result of which was the conclusion that his proposed work does not present any ethical concerns; is extremely low risk; and thus does not require the scrutiny of the full Research Ethics Committee.

Yours faithfully

Biddle

Mr Jack Biddle – Research Approvals Advisor

Patron: Her Majesty the Queen
Incorporated by Royal Charter as Queen Mary and Westfield College, University of London

A.2 Participant Information Sheet

Pro forma information sheet and consent form

Queen Mary
University of London

<u>Information sheet</u>

<u>User Driven Virtual Overlays: information for participants</u>

We would like to invite you to be part of this research project, if you would like to. You should only agree to take part if you want to, it is entirely up to you. If you choose not to take part there won't be any disadvantages for you and you will hear no more about it. [*If appropriate:* Choosing not to take part will not affect your access to treatment or services in any way].

Please read the following information carefully before you decide to take part; this will tell you why the research is being done and what you will be asked to do if you take part. Please ask if there is anything that is not clear or if you would like more information.

If you decide to take part you will be asked to sign the attached form to say that you agree.

You are still free to withdraw at any time and without giving a reason.

This study explores Augmented Reality games exploring the interactions of a mixed reality treasure hunt game.The experiments will take from 60-90 min and you will be asked to play a number of rounds of the game and responding to a few experience related questionnaires. These games involve exploring a physical space and solving simple puzzles. Any personal data recorded about you will be kept anonymous.

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep and be asked to sign a consent form.

If you have any questions or concerns about the manner in which the study was conducted please, in the first instance, contact the researcher responsible for the study. If this is unsuccessful, or not appropriate, please contact the Secretary at the Queen Mary Ethics of Research Committee, Room W104, Queen's Building, Mile End Campus, Mile End Road, London or research-ethics@qmul.ac.uk.

A.3 Participant Consent Form

Queen Mary
University of London

Consent form

Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research.

Title of Study: Experience Driven Procedural Content Generation for Exploration in Augmented Reality
Queen Mary Ethics of Research Committee Ref: <REF>

- Thank you for considering taking part in this research. The person organizing the research must explain the project to you before you agree to take part.
- If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time.
- *I understand that if I decide at any other time during the research that I no longer wish to participate in this project, I can notify the researchers involved and be withdrawn from it immediately.*
- *I consent to the processing of my personal information for the purposes of this research study. I understand that such information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 1998.*

Participant's Statement:

I ____ agree that the research project named above has been explained to me to my satisfaction and I agree to take part in the study. I have read both the notes written above and the Information Sheet about the project, and understand what the research study involves.

Signed: Date:

Investigator's Statement:

I ____ confirm that I have carefully explained the nature, demands and any foreseeable risks (where applicable) of the proposed research to the volunteer

A.4 Pre-Study Questionnaire

Demographic Information

1. Email: ______________________________________

2. Gender:*(mark only one option)*
   [ ] Female
   [ ] Male
   [ ] Not Listed:________________________________
   [ ] Prefer not to say

3. Age: *(mark only one option)*
   [ ] 18-24
   [ ] 25-29
   [ ] 30-34
   [ ] 35-39
   [ ] 40-44
   [ ] 45-49
   [ ] 50-54
   [ ] 55-59
   [ ] 60 and above
   [ ] Prefer not to say

4. Have you player Augmented Reality games before? *(mark only one option)*
   [ ] Never
   [ ] I have played less than 3 AR games before
   [ ] I have played 3-6 AR games before
   [ ] I have played more than 6 AR games before

5. How often do you play Augmented Reality games? *(mark only one option)*
   [ ] Never
   [ ] Once in the past
   [ ] A few times in the past
   [ ] A few times a month
   [ ] A few times a week

6. Can you name a few Augmented Reality games you have played in the past? (if applicable)

   ____

   ____

A.5 Post-Session Questionnaire

AR - Treasure Hunt

Please fill up your participant information

1. **Participant ID**

2. **Session ID**

Fun

3. **Please select 1 of the following options**
Mark only one oval.
   - Game 1 felt more FUN than Game 2
   - Game 2 felt more FUN than Game 1
   - Game 1 and Game 2 felt equally FUN
   - Neither of the two games felt FUN

Please rate your level of agreement with the following statements

4. **Game 1 felt FUN**
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree						Strongly Agree

5. **Game 2 felt FUN**
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree						Strongly Agree

Frustration

6. **Please select 1 of the following options**
Mark only one oval.

- Game 1 felt more FRUSTRATING than Game 2
- Game 2 felt more FRUSTRATING than Game 1
- Game 1 and Game 2 felt equally FRUSTRATING
- Neither of the two games felt FRUSTRATING

Please rate your level of agreement with the following statements

7. **Game 1 felt FRUSTRATING**
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree						Strongly Agree

8. **Game 2 felt FRUSTRATING**
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree						Strongly Agree

Excitement

9. **Please select 1 of the following options**
Mark only one oval.

- Game 1 was more EXCITING than Game 2
- Game 2 was more EXCITING than Game 1
- Game 1 and Game 2 were equally EXCITING
- Neither of the two games were EXCITING

Please rate your level of agreement with the following statements

10. **Game 1 was EXCITING**
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree						Strongly Agree

11. **Game 2 was EXCITING**
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

Boredom

12. **Please select 1 of the following options**
Mark only one oval.

- ◯ Game 1 felt more BORING than Game 2
- ◯ Game 2 felt more BORING than Game 1
- ◯ Game 1 and Game 2 felt equally BORING
- ◯ Neither of the two games felt BORING

Please rate your level of agreement with the following statements

13. **Game 1 felt BORING**
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

14. **Game 2 felt BORING**
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

Challenge

15. **Please select 1 of the following options**
Mark only one oval.

- ◯ Game 1 felt more CHALLENGING than Game 2
- ◯ Game 2 felt more CHALLENGING than Game 1
- ◯ Game 1 and Game 2 felt equally CHALLENGING
- ◯ Neither of the two games felt CHALLENGING

Please rate your level of agreement with the following statements

16. **Game 1 felt CHALLENGING**
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	⬭	⬭	⬭	⬭	⬭	Strongly Agree

17. **Game 2 felt CHALLENGING**
Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	⬭	⬭	⬭	⬭	⬭	Strongly Agree

Appendix B

Experiment-2 materiel

This appendix contains materials used for experiment-2 reported in chapter 4.

B.1 Ethics Approval

Queen Mary
University of London

Queen Mary, University of London
Room W117
Queen's Building
Queen Mary University of London
Mile End Road
London E1 4NS

Queen Mary Ethics of Research Committee
Hazel Covill
Research Ethics Administrator
Tel: +44 (0) 20 7882 7915
Email: h.covill@qmul.ac.uk

c/o Dr Laurissa Tokarchuk
CS412
Department of Computer Science
Queen Mary University of London
Mile End Road
London

24th October 2019

To Whom It May Concern:

Re: QMREC2204 – Modelling Player Experience in Augmented Reality Mobile Games.

I can confirm that Mr Vivek Warriar has completed a Research Ethics Questionnaire with regard to the above research.

The result of which was the conclusion that his proposed work does not present any ethical concerns; is extremely low risk; and thus does not require the scrutiny of the full Research Ethics Committee.

Yours faithfully

Ms Hazel Covill – Research Ethics Facilitator

Patron: Her Majesty the Queen
Incorporated by Royal Charter as Queen Mary and Westfield College, University of London

B.2 Participant Information Sheet

Pro forma information sheet and consent form

Queen Mary
University of London

Information sheet

Modelling Player Experience in AR Mobile Games: information for participants

We would like to invite you to be part of this research project, if you would like to. You should only agree to take part if you want to, it is entirely up to you. If you choose not to take part there won't be any disadvantages for you and you will hear no more about it.

Please read the following information carefully before you decide to take part; this will tell you why the research is being done and what you will be asked to do if you take part. Please ask if there is anything that is not clear or if you would like more information.

If you decide to take part you will be asked to sign the attached form to say that you agree.

You are still free to withdraw at any time and without giving a reason.

This study explores Augmented Reality games by investigating the interactions of a mixed reality target acquisition. The study will take up-to 60 min and you will be asked to play a number of rounds of the game and respond to a few experience related questionnaires. These games involve moving through a physical space and capturing creatures within a time limit. Any personal data recorded about you will be kept anonymous. It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep and be asked to sign a consent form.

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep and be asked to sign a consent form.

Please read Queen Mary's privacy notice for research participants[1] for important information about your personal data and your rights in this respect.

If you have any questions or concerns about the manner in which the study was conducted please, in the first instance, contact Vivek Warriar (v.r.warriar@qmul.ac.uk). If. If this is unsuccessful, or not appropriate, please contact the Secretary at the Queen Mary Ethics of Research Committee, Room W104, Queens' Building, Mile End Campus, Mile End Road, London, E1 4NS or research-ethics@qmul.ac.uk. If you have any questions relating to data protection, please contact Data Protection Officer, Queens' Building, Mile End Road, London, E1 4NS or data-protection@qmul.ac.uk

[1] This is found at: http://www.arcs.qmul.ac.uk/media/arcs/policyzone/Privacy-Notice-for-Research-Participants.pdf

B.3 Participant Consent Form

Queen Mary
University of London

Consent form

Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research.

Title of Study: Modelling Player Experience in AR Mobile Games
Queen Mary Ethics of Research Committee Ref: ___________________

Thank you for considering taking part in this research. The person organizing the research must explain the project to you before you agree to take part.

If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time. If you are willing to participate in this study, please circle the appropriate responses and sign and date the declaration underneath.

Statement	**Circle a response**
I agree that the research project named above has been explained to me to my satisfaction in verbal and/or written form	YES / NO
I understand that if I decide at any other time during the research that I no longer wish to participate in this project, I can notify the researchers involved and be withdrawn from it immediately	YES / NO
I have read both the notes written above and the Information Sheet about the project, and understand what the research study involves	YES / NO
I agree to take part in the study, which will include use of my personal data	YES / NO

Participant's Signature:__________________________ Date: ___________________

Investigator's Statement:

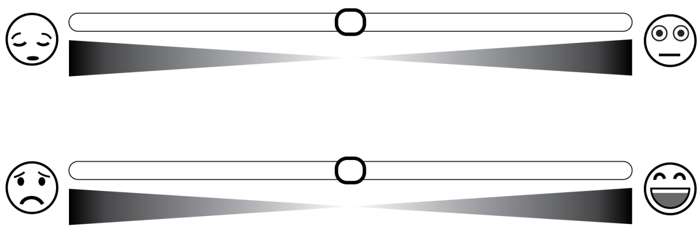
I ____ confirm that I have carefully explained the nature, demands and any foreseeable risks (where applicable) of the proposed research to the volunteer and provided a copy of this form

B.4 Pre-Study Questionnaire

The same pre-study questionnaire as the previous study (A.4) was was used in this study.

B.5 Post-Session Questionnaire

Affective Slider



Game Experience Questionnaire

Please rate the following statements according on a scale of 1-7

1. I was interested in the game's objective

Strongly Disagree			Neutral			Strongly Agree

2. I felt successful

Strongly Disagree			Neutral			Strongly Agree

3. I felt bored

Strongly Disagree			Neutral			Strongly Agree

4. I found it impressive

Strongly Disagree			Neutral			Strongly Agree

5. I forgot everything around me

Strongly Disagree			Neutral			Strongly Agree

6. I felt frustrated

Strongly Disagree			Neutral			Strongly Agree

7. I found it tiresome

Strongly Disagree			Neutral			Strongly Agree

8. I felt irritable

Strongly Disagree			Neutral			Strongly Agree

9. I felt skilful

Strongly Disagree			Neutral			Strongly Agree

10. I felt completely absorbed

Strongly Disagree			Neutral			Strongly Agree

11. I felt content

Strongly Disagree			Neutral			Strongly Agree

12. I felt challenged

Strongly Disagree			Neutral			Strongly Agree

13. I had to put a lot of effort into it

Strongly Disagree			Neutral			Strongly Agree

14. I felt good

Strongly Disagree			Neutral			Strongly Agree

Appendix C

Experiment-3 materiel

This appendix contains materials used for experiment-3 reported in chapter 5.

C.1 Ethics Approval

The same ethics approval as the previous study (B.1) was extended for this one.

C.2 Participant Information Sheet

The same participant information sheet as the previous study (B.2) was used for this study.

C.3 Participant Consent Form

The same participant consent form as the previous study (B.3) was used for this study.

C.4 Pre-Study Questionnaire

The same pre-study questionnaire as the first study (A.4) was was used in this study.

C.5 Post-Session Questionnaire

Post-Session Questionnaire

Please rate the following statements according on a scale of 1-7

1. This level was fun to play.

Strongly Disagree			Neutral			Strongly Agree

2. I was no longer aware of my surroundings while I was playing this level.

Strongly Disagree			Neutral			Strongly Agree

3. I thought this level was quite enjoyable.

Strongly Disagree			Neutral			Strongly Agree

4. I was immersed in this level.

Strongly Disagree			Neutral			Strongly Agree

5. I felt a sense of freedom about how I wanted to play this level.

Strongly Disagree			Neutral			Strongly Agree

6. I thought this was a boring level.

Strongly Disagree			Neutral			Strongly Agree

7. I felt I was good at playing this level.

Strongly Disagree			Neutral			Strongly Agree

8. I would describe this level as very interesting.

Strongly Disagree			Neutral			Strongly Agree

9. I felt free to play this level in my own way.

Strongly Disagree			Neutral			Strongly Agree

10. I enjoyed playing this level very much.

Strongly Disagree			Neutral			Strongly Agree

11. I was fully focused on this level.

Strongly Disagree			Neutral			Strongly Agree

12. I felt a sense of mastery playing this level.

Strongly Disagree			Neutral			Strongly Agree

13. I felt like I had choices regarding how I wanted to play this level.

Strongly Disagree			Neutral			Strongly Agree

14. While I was playing this level, I was thinking about how much I enjoyed it.

Strongly Disagree			Neutral			Strongly Agree

15. I felt capable while playing the level.

Strongly Disagree			Neutral			Strongly Agree

Appendix D

Experiment-4 materiel

This appendix contains materials used for experiment-4 reported in chapter 5.

D.1 Ethics Approval

Queen Mary
University of London

Queen Mary, University of London
Room W117
Queen's Building
Queen Mary University of London
Mile End Road
London E1 4NS

Queen Mary Ethics of Research Committee
Hazel Covill
Research Ethics Facilitator
Tel: +44 (0) 20 7882 7915
Email: research-ethics@qmul.ac.uk

c/o Dr Laurissa Tokarchuk
School of Electronic Engineering and
Computer Science
Queen Mary University of London
Mile End Road
London
E1 4NS
United Kingdom

18 November 2020

To Whom It May Concern:

Re: QMERC20.040 - Dynamic Difficulty Adaptation for Augmented Reality Mobile Games.

I can confirm that Vivek Ramesh Warriar has completed a Research Ethics Application with regard to the above study.

The result of which was the conclusion that the proposed work does not present any ethical concerns; is low risk; and thus does not require the scrutiny of the full Research Ethics Committee.

Yours faithfully

Mantalena Sotiriadou – Research Ethics Facilitator

D.2 Participant Information Sheet

Insert DATE and VERSION NUMBER

Queen Mary
University of London

Participant Information Sheet

Study title

Player Experience in Augmented Reality Reality Mobile Games

Version number and date

Version 0.1: 09.11.2020

Researcher's name

Vivek Warriar supervised by Dr. Laurissa Tokarchuk

Queen Mary Ethics of Research Committee reference number:

[Insert reference number allocated to your study by the Research Ethics Facilitators]

Invitation paragraph

You are being invited to participate in a research study. Before you decide whether or not you wish to participate in this study, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us questions if there is anything that is not clear or if you would like more information.

What is the purpose of the study and what would taking part involve?

The purpose of the study is evaluate the experience of playing 2 versions of an Augmented Reality Game. If you agree to take part in the study you will be asked to fill in a pre-study survey where you will select one (out of two) site where the study will take place. The research sites are two parks where you will need to travel to in order to play these games. For your reference the sites selected are:

1. **Carlton Square and Gardens**
   Bethnal Green, London
   This is a park behind the computer science building. Google maps: https://goo.gl/maps/61QSDh8y9hW6WqZu6

2. **Russell Square Park**
   Bloomsbury, London WC1B 5BG
   This is a park next to Russell Square Tube station. Google maps: https://goo.gl/maps/74qku1YbhfyyejRD9

As part of this study, you will be asked to play a research game titled “Running Chickens” in the park you have selected to travel to. In this game, you will use your mobile device to move around the park and capture augmented reality chickens. In each level you will be presented with a number of chickens, the goal of the game is to collect all the chickens in the level or as many as you can before the time limit of the level runs out.

You will be sent 3 games (1 training game and 2 experiment games) to download onto your mobile device and the order in which you will need to play them. You will also be sent a tutorial video on how to play the game. The game has been designed such that all game play is contained within the park boundaries. Do not leave the boundaries of the park to play this game. If it seems that you would need to leave the park to play the game, terminate the game immediately and contact the researcher. Once you arrive at the site, you should make your way to the approximate centre of the park. Where you will play the games.

At this point please play the training game to familiarise yourself with how the game works. Once the training is done you will take a 5 min break. After this, you will play the first experiment game for 10 levels, after which the game will present you with a post-session questionnaire where you will rate you experience playing the game across a number of factors of interest.

After the first session is completed, you will take a 10 min break. You will then play the second experiment game for 10 levels. You will then be presented with a similar questionnaire while will evaluate your experience playing the second game.

Once you have played both experiment games and filled in both questionnaires, please notify the researcher by email: v.r.warriar@qmul.ac.uk.

The researcher will schedule a post-study call with you for a short post-experiment interview. You will then be debriefed and the study will be concluded. You will then be paid £15.00 in Amazon vouchers for your participation in the study. The study should take approximately 1 hour to complete once you have arrived at the study site.

Why am I being invited?

You are being invited to participate in this research study because you are over the age of 16 and have normal ability to walk/jog.

You should not take part in this study if you are under the age of 16 and if you are not able to jog in an outdoor park.

Do I have to take part?

This participant information sheet has been written to help you decide if you would like to take part. It is up to you whether you wish to take part. If you do decide to take part you will be free to withdraw at any time without needing to provide a reason, and with no penalties or detrimental effects.

What are the possible benefits of taking part?

Taking part in the study will support this PhD research that aims to further our understanding of human experiences in movement-based Augmented reality games. The outcome of this study will help game designers make more engaging movement-based games which would lead to the promotion of physical activity.

What are the possible disadvantages and risks of taking part?

Keep in mind that this study involves moving around a park to play the games being evaluated in the study. Please keep your health and safety is the highest priority and do not attempt the study if it is raining or the ground is unsafe for mild jogging in the park.

Expenses and payments

Participants will be paid £15.00 in Amazon vouchers for their participation in the study.

What information about me will you be collecting?

Data about your age, gender and your previous experience previous experience playing Augmented reality mobile games. In addition, you experience playing each version of the game will be collected in the form of surveys. Finally, data about your performance in the game (eg: game score, gameplay time) will also be collected.

How will my data be stored and who will have access to it?

Your data will be stored in fully anonymised format in London, UK, and only the PhD researcher Vivek Warriar will be able to access it. Your data information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 2018 and comply with the data processing and storage policies of Queen Mary University of London

Data Protection Policy

Information/Data Governance Policy – DG14 – Storage of information

When and how will my data be destroyed?

Your data will be saved electronically for 5 years after collection and erased through hard drive formatting.

How will my data be used and shared?

The results of this study will be part of the PhD thesis that is connected to this research project. Results will be mentioned in a future conference or journal paper publication. All data is stored locally in an anonymised form and will not be accessible for or shared with others.

Research Data Access and Management Policy

Under what legal basis are you collecting this information?

Queen Mary University of London processes personal data for research purposes in accordance with the lawful basis of 'public task'.

Please read Queen Mary's privacy notice for research participants containing important information about your personal data and your rights in this respect. If you have any questions relating to data protection, please contact Queen Mary's Data Protection Officer, Queens' Building, Mile End Road, London, E1 4NS or data-protection@qmul.ac.uk or 020 7882 7596.

What will happen if I want to withdraw from this study?

You can withdraw from this study at any time without providing a reason. Withdrawing will have no disadvantage for you, and you will hear no more about this study. Your data will only be submitted if you complete the study.

Your data will be saved entirely anonymised and is not possible to link the data to a particular person. For this reason, it is however not possible to delete the data entry of a specific person.

What should I do if I have any concerns about this study?

If you have any concerns about the manner in which the study was conducted, in the first instance, please contact the researcher(s) responsible for the study Dr. Laurissa Tokarchuk: laurissa.tokarchuk@qmul.ac.uk. If you have a complaint which you feel you cannot discuss with the researchers then you should contact the Research Ethics Facilitators by e-mail: research-ethics@qmul.ac.uk. When contacting the Research Ethics Facilitators, please provide details of the study title, description of the study and QMERC reference number (where possible), the researcher(s) involved, and details of the complaint you wish to make.

Who can I contact if I have any questions about this study?

Vivek Warriar

v.r.warriar@qmul.ac.uk

07852812239

D.3 Participant Consent Form

[Insert DATE and VERSION NUMBER]

Queen Mary
University of London

Consent Form

Title of Research Study: Player Experience in Augmented Reality Mobile Games

Principal Investigator: Vivek Warriar supervised by Dr. Laurissa Tokarchuk

Queen Mary Ethics of Research Committee Ref: [Insert the reference number allocated to your research ethics application by the Research Ethics Facilitator].

Thank you for your interest in this research.

Should you wish to participate in the study, please consider the following statements. Before signing the consent form, you should initial all or any of the statements that you agree with. Your signature confirms that you are willing to participate in this research, however you are reminded that you are free to withdraw your participation at any time.

Statement	**Please initial box**
1. I confirm that I have read the Participant Information Sheet dated 9/11/2020 version 0.1 for the above study; or it has been read to me. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.	
2. I understand that my participation is voluntary and that I am free to stop taking part in the study at any time without giving any reason and without my rights being affected.	
3. I understand that my data will be accessed by the investigator	

4. I understand that my data will be securely stored in London, UK and in accordance with the data protection guidelines of the Queen Mary University of London 5 years in a fully anonymised form.	
5. I understand that collected data is completely anonymized and information that I have provided can therefore not be withdrawn after submission.	
6. I agree to the post-experiment interview being audio recorded.	
7. I understand that the information collected about me will be used to support other research in the future, and it may be shared in anonymised form with other researchers.	
8. I agree to take part in the above study.	

Participants should read Queen Mary's privacy notice for research participants which contains important information about your personal data and your rights in this respect. If you have any questions relating to data protection, please contact Data Protection Officer, Queens' Building, Mile End Road, London, E1 4NS or data-protection@qmul.ac.uk or 020 7882 7596.

Agree button (consent form will be provided online)

Principal Investigator (or Supervisor for student projects)

Dr. Laurissa Tokarchuk

Laurissa.tokarchuk@qmul.ac.uk

Student Investigator (if applicable)

Vivek Warriar

v.r.warriar@qmul.ac.uk

07852812239

D.4 Pre-Study Questionnaire

Demographic Information

1. Email: ____________________________________

2. Gender: *(mark only one option)*
   [ ] Female
   [ ] Male
   [ ] Not Listed:_______________________________
   [ ] Prefer not to say

3. Age: *(mark only one option)*
   [ ] 18-24
   [ ] 25-29
   [ ] 30-34
   [ ] 35-39
   [ ] 40-44
   [ ] 45-49
   [ ] 50-54
   [ ] 55-59
   [ ] 60 and above
   [ ] Prefer not to say

4. Have you player Augmented Reality games before? *(mark only one option)*
   [ ] Never
   [ ] I have played less than 3 AR games before
   [ ] I have played 3-6 AR games before
   [ ] I have played more than 6 AR games before

5. How often do you play Augmented Reality games? *(mark only one option)*
   [ ] Never
   [ ] Once in the past
   [ ] A few times in the past
   [ ] A few times a month
   [ ] A few times a week

6. Can you name a few Augmented Reality games you have played in the past? (if applicable)

   ____

   ____

7. Please select which of the study site is the most convenient for you to travel to for this study:
   [ ] **Carlton Square and Gardens**
   Bethnal Green, London
   This is a park behind the computer science building. Google maps:
   https://goo.gl/maps/61QSDh8y9hW6WqZu6

   [] **Russell Square Park**

Bloomsbury, London WC1B 5BG
This is a park next to Russell Square Tube station. Google maps: https://goo.gl/maps/74qku1YbhfyyejRD9

D.5 Post-Game Questionnaire

Post-Game Questionnaire

Please rate the following statements according on a scale of 1-7

1. I felt I was good at playing this level.

Strongly Disagree			Neutral			Strongly Agree

2. I felt a sense of mastery playing this level.

Strongly Disagree			Neutral			Strongly Agree

3. I felt capable while playing the level.

Strongly Disagree			Neutral			Strongly Agree

D.6 Post-Session Questionnaire

Post session Questionnaire

Please rate the following statements according on a scale of 1-7

1. Playing the game was meaningful to me.

Strongly Disagree			Neutral			Strongly Agree

2. The game felt relevant to me.

Strongly Disagree			Neutral			Strongly Agree

3. Playing this game was valuable to me.

Strongly Disagree			Neutral			Strongly Agree

4. I felt capable while playing the game.

Strongly Disagree			Neutral			Strongly Agree

5. I felt I was good at playing this game.

Strongly Disagree			Neutral			Strongly Agree

6. I felt a sense of mastery playing this game.

Strongly			Neutral			Strongly

Disagree Agree

7. I was no longer aware of my surroundings while I was playing.

Strongly Disagree Neutral Strongly Agree

8. I was immersed in the game.

Strongly Disagree Neutral Strongly Agree

9. I was fully focused on the game.

Strongly Disagree Neutral Strongly Agree

10. I felt a sense of freedom about how I wanted to play this game.

Strongly Disagree Neutral Strongly Agree

11. I felt free to play the game in my own way.

Strongly Disagree Neutral Strongly Agree

12. I felt like I had choices regarding how I wanted to play this game.

Strongly Disagree Neutral Strongly Agree

13. I felt eager to discover how the game continued.

Strongly Disagree | Neutral | Strongly Agree

14. I wanted to explore how the game evolved.

Strongly Disagree | Neutral | Strongly Agree

15. I wanted to find out how the game progressed.

Strongly Disagree | Neutral | Strongly Agree

16. I thought the game was easy to control.

Strongly Disagree | Neutral | Strongly Agree

17. The actions to control the game were clear to me.

Strongly Disagree | Neutral | Strongly Agree

18. It was easy to know how to perform actions in the game.

Strongly Disagree | Neutral | Strongly Agree

19. The game was challenging but not too challenging.

Strongly Disagree | Neutral | Strongly Agree

20. The game was not too easy and not too hard to play.

Strongly Disagree — Neutral — Strongly Agree

21. The challenges in the game were at the right level of difficulty for me.

Strongly Disagree — Neutral — Strongly Agree

22. The game gave clear feedback on my progress towards the goals.

Strongly Disagree — Neutral — Strongly Agree

23. I could easily assess how I was performing in the game.

Strongly Disagree — Neutral — Strongly Agree

24. The game informed me of my progress in the game.

Strongly Disagree — Neutral — Strongly Agree

25. I enjoyed the way the game was styled.

Strongly Disagree — Neutral — Strongly Agree

26. I liked the look and feel of the game.

Strongly Disagree — Neutral — Strongly Agree

27. I appreciated the aesthetics of the game.

Strongly Disagree · Neutral · Strongly Agree

28. The goals of the game were clear to me.

Strongly Disagree · Neutral · Strongly Agree

29. I grasped the overall goal of the game.

Strongly Disagree · Neutral · Strongly Agree

30. I understood the objectives of the game.

Strongly Disagree · Neutral · Strongly Agree