Emotion in Motion: Introducing real-time emotional responsiveness to VR exergames

Zoe Broad
zab26@bath.ac.uk
Department of Computer Science
University of Bath

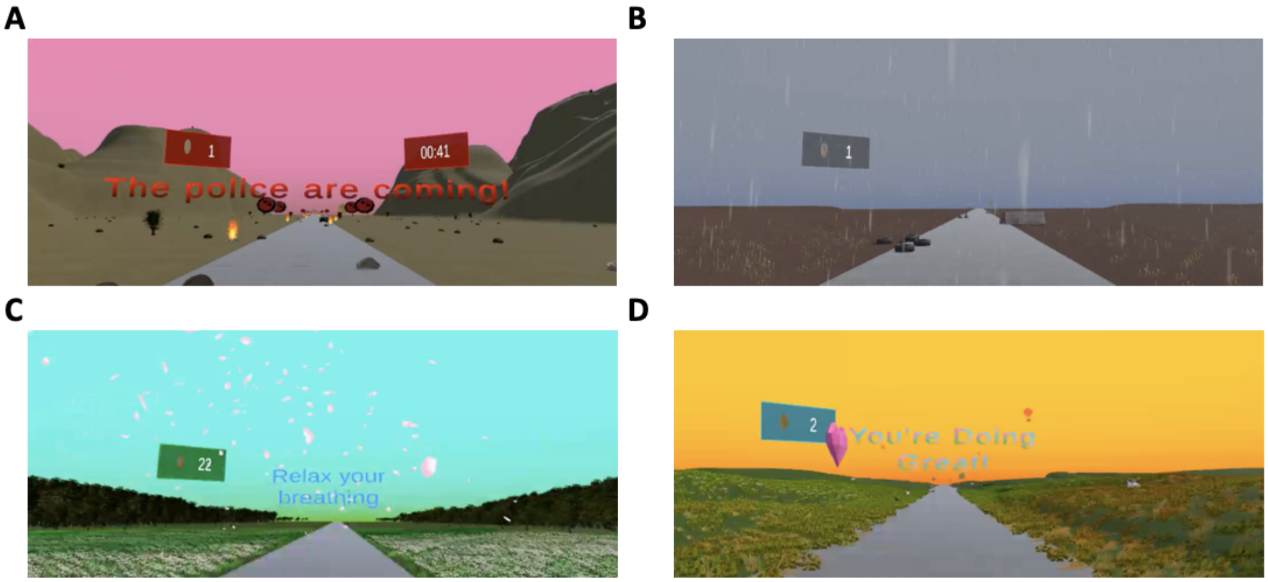

Figure 1: Emotion scenes from developed Virtual Reality cycling exergame: (A) Stress Scene; (B) Sadness Scene; (C) Serenity Scene; (D) Joy Scene

ABSTRACT

This study aimed to determine whether psychophysiological measures could accurately monitor emotional states in VR exergaming at low, moderate, and vigorous exercise intensities. This is a challenge which needs to be addressed for adaptive VR exergames in order for them to become emotionally responsive and therefore more engaging. Emotional states have previously been determined using fMRI, facial EMG, EEG, and electrocardiography (ECG). Virtual Reality exergames pose challenges as the effects of exercise and VR headsets interfere with those measurements. We present an experiment that investigates the use of four different psychophysiological measures for emotion recognition in a VR cycling exergame at low, moderate, and vigorous exercise intensities. This exergame was designed and developed for the study, with four scenes, each targeting a separate emotion from Russell's circumplex model. We identify pupil dilation, blink duration, and skin conductance as suitable measures for monitoring emotional states in VR exergaming and provide a set of new Generalised Least Squares regression formulae for future work. These findings are of great importance for future developers as the implementation of real-time emotional responsiveness opens up a whole new dimension to the VR exergaming world.

KEYWORDS

Emotion recognition, Exergaming, Virtual Reality, Psychophysiological measures, Emotion-sensitive adaptive VR exergames

1 INTRODUCTION

Studies have shown that exergames (a combination of "exercise and gaming", where physical effort is required to progress the game) often result in low-to-moderate intensity activities and therefore do not produce equivalent physiological impact as traditional exercise such as cycling or running. Dynamic Difficulty Adjustment ("DDA") aimed to resolve this, whereby an exergame's difficulty adjusts based on players' readings e.g., score.

The majority of DDA literature focuses on physiological adaptation. For example, Ketelhut et al. [24] developed an exergame

MSc Dissertation, Sept 09–09, 2022, Bath, UK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

which implemented DDA using heart rate ("HR"), where the exergame difficulty would increase when the player's HR dropped below a set threshold. Although the exergame qualified as vigorous physical exercise, a user experience study by Darzi et al. [12] did not support the inclusion of physiological DDA as it was not found to improve user experience. The Dual Flow Model ("DFM") helps to mitigate this lack of improved user experience by balancing the game's challenge with the player's skill (attractiveness), and the game's intensity with the player's fitness (effectiveness).

However, one element that the DFM does not address is emotion recognition. There is theoretical and empirical evidence in the field of educational psychology showing that emotions have a significant influence on learning processes (e.g., directing attention), and as a result performance [55]. Furthermore, being able to adapt games to players' emotions through activating and deactivating their positive and negative emotions can improve task engagement and persistence [55]. However, being able to adapt games based on players' emotions relies upon accurately recognising emotions. This is a challenge which needs to be addressed for adaptive VR exergames in order for them to become emotionally responsive and therefore more engaging.

Müller et al. [40] researched emotion responses (e.g., fear/surprise) to jump scare events in a cycling exergame using facial expression detection and electrodermal activity measurements. This method of emotion recognition was successful, however it has an inherent limitation in that it cannot be used in Virtual Reality ("VR") gaming due to the VR headsets impairing facial expression detection. There is currently no evidence in the literature of emotion recognition in VR exergaming. This study therefore aimed to determine whether psychophysiological measures could provide accurate emotion recognition for low, moderate, and vigorous intensity VR exergaming.

To achieve this, a VR cycling exergame was developed for this study. The 4 distinct emotion scenes were: stress, sadness, serenity, and joy. Each of these emotions represents a different quadrant of the circumplex model which differentiates between emotions using two dimensions: valence and arousal. For example, Stress is of negative valence and high arousal, whilst Serenity is of positive valence and low arousal. The exergame required participants to cycle on a stationary exercise bike (Wahoo KICKR) whilst wearing a Vive Pro Eye VR headset with eye tracking, a Polar HR chest strap, and a Consensys GSR Development Kit shimmer device to measure skin conductance. Participants completed one low, one moderate, and one vigorous intensity exercise bout. Each exercise bout followed the format: 2-minute warm-up (to reduce injury risk and reach target HR) followed by four 1-minute periods of exercise, each with a 1-minute recovery period in between. During each recovery period, participants were asked to verbally rate the intensity at which they felt 8 different emotions during the 1-minute exercise period on a scale of 1-10. This created a ground truth database. The 8 different emotions consisted of the four emotions targeted in the exergame scenes (stress, sadness, serenity, joy) and one additional emotion from each quadrant of the circumplex model (fear, negative quiet, positive quiet, excitation).

The collected psychophysiological data and ground truth database were then analysed in order to answer the following research questions:

RQ1: Can psychophysiological measurements be used to predict emotions in VR exergaming?
RQ2: Can psychophysiological measurements predict emotions at low, moderate, and vigorous exercise intensities in VR exergaming?

The first key finding of the study was that the developed VR exergame environment was effective at targeting emotions. We conducted across-scene analysis (i.e., did the stress scene elicit significantly greater stress than all other scenes?) and within-scene analysis (i.e., within the stress scene, was stress elicited significantly more than all other emotions?) using repeated-measures ANOVAs. All results from these analyses were significant ($p < .001^{***}$). The second key finding of the study was that pupil dilation, blink duration, and skin conductance were significant predictors of emotion ($p < .001^{***}, p < .05^{*}, and p < .05^{*}$ respectively).

From these findings, we make the following contributions:
1. VR exergaming environments can be created to effectively target emotions,
2. Psychophysiological measures can be used to monitor emotions at low, moderate, and vigorous exercise intensities in VR exergaming, and
3. Provision of emotion specific Generalised Least Squares regression formulae.

These findings are of great importance for future developers as the implementation of real-time emotional responsiveness opens up a whole new dimension to the VR exergaming world.

2 RELATED WORK

2.1 Exergame Development

Exergaming is an umbrella term that is being used in a relatively new branch of research. It is composed of two main elements: physical exertion and games. The main concept of exertion is achieving physical fatigue through exercise. The idea of game play is that players voluntarily engage in an activity that has man-made obstacles. Mueller et al. [36] combined these two concepts and defined the overarching term of exergames as "digital games where the outcome of the game is predominantly determined by physical effort".

The use of immersive exergames to motivate players to exercise was first researched by Finkelstein et al. [15], who introduced the game Astrojumper. This game required players to move their body to avoid incoming asteroids which were displayed on an immersive three-wall stereoscopic projected display. However, moving exergames to VR, Charoensook et al. [7] showed that using a head-mounted display could also increase players' HR during exergame sessions. More recently, Yoo et al. [59] found that VR exergames motivated sedentary workers to be physically active over longer periods of time than non-VR exergames.

2.1.1 Framework. The literature regarding how exergames can be developed most effectively has been inconclusive. This is primarily due to the training intensity achieved by exergames, with many only reaching light-moderate levels of physical activity [36]. However, the framework devised by Mueller et al. [36] has now been widely accepted. Mueller provides a comprehensive framework which was previously missing in the literature, and was based upon splitting exergames into their two composite categories: physical exertion and digital games. Two pre-existing frameworks were then used to support each of these categories: Salen and Zimmerman's [45] schema on the perspectives on gaming was used to support the digital game aspect of Mueller's framework, whilst the physical exertion arm was supported by utilising the 4 perspectives of the body proposed by Jacob et al. [22]. The wide acceptance of Mueller's framework is likely due to its provision of a comprehensive summary as to how exertion games should be considered, designed, and analysed. As such Mueller's framework provides a strong foundation for the current study to adopt when developing the exergame.

2.1.2 Taxonomy. Mueller et al. [37] identified a taxonomy to breakdown and categorise exergames. In the broadest sense, games were considered to be exertion or non-exertion, with exergames being in the exertion category. Exergames were then broken down into those that were non-competitive and those that were competitive. Non-competitive games do not involve an opponent whilst competitive games do. In this instance, the opponent may be a player, or a computer representing a player, who pursues the goal of the game and provides an obstacle preventing the player from achieving their goal. It is important to note that this taxonomy demonstrates that not all exergames fall within the same category and may explain the variation in definitions and results seen within the literature.

2.2 Dynamic Difficulty Adjustment (DDA)

2.2.1 History. The concept of video gaming is continuously changing. Early games such as Pong or Computer Space from c. 1970 were limited to commercial arcades, but are now available on mobile phones, tablets, and computers. In addition, difficulty levels of traditional games were increased linearly throughout the duration of the game. For example, the speed or opponent difficulty could only be selected before starting the game by choosing the difficulty level. This method of selecting gaming difficulty was reported to result in negative user experiences due to the selected level not matching user's ability [25]. Furthermore, the frequent need for players to manually adjust the game's difficulty level was considered to impair concentration and flow of the game [8]. These issues were attempted to be solved through DDA which adjusts the difficulty of an exergame based on players' readings.

Czikszentmihalyi [11] first proposed that players, when kept away from the states of boredom (too easy) or frustration (too hard), travel through a "flow channel". Successfully implementing DDA can help players to reach and maintain this state of flow, which has been reported to induce feelings of a deep sense of enjoyment [9, 10]. For exergames, DDA is primarily achieved by tuning specific ingame parameters such as the speed of objects or success rates of hits. The difficulty of the exercises can also be adjusted to ensure efficient physical training is provided without overburdening the player [59].

2.2.2 Approaches. Between 2009 and 2017 the field of DDA increased in popularity, with the number of works increasing from 4 to 17 per year [60]. Many of these works aimed to develop the methods of DDA implementation. The concept of inventory (i.e., the store of items a player gathers and takes throughout the game) is commonly seen in games [48], with user experience being impacted by the relative abundance or lack of items in inventory. Hamlet, a DDA system built by Hunicke and Chapman [21], uses methods based on inventory management to manipulate the game difficulty. This has also been seen by Stein et al. [52], who measured the excitement of players and adjusted the difficulty when the player's level of excitement went below the threshold value. This is relevant to the current study, where target HR zones will be pre-defined as a percentage of the player's maximum HR ("HRmax") for each of the low, moderate, and vigorous intensity exercise bouts.

In contrast to using pre-defined statistical metrics, Xue et al. [57] used a probabilistic graph to model the player's progression, where engagement was maximised as a well-defined objective function. However, the results from this study focused on core engagement metrics such as game duration, which is not the objective of the current study. Similarly, Jennings-Teats et al. [23] used a complex two-layer approach for DDA. The player's skill and ability levels were first analysed using machine learning and level generation. This information was then fed into a machine-learned difficulty model in a 2D platformer game. Although the use of machine-learning models chose suitable difficulty levels for the existing performance of the player, the complexity of such an approach is not within the scope of the current project.

2.2.3 Physiological DDA Measures. Literature focusing on the use of players' HR for DDA is widespread within exergame-specific literature. For example, Mueller et al. [39] developed an urban jogging exergame where HR data and spatialized sound were used to balance the differing ability levels and geographical locations of joggers. Mueller demonstrated that HR-based DDA positively affected user experience due to each participant performing in their own heart rate training zone whilst also being able to engage with other participants who may have different physical abilities.

HR-based DDA has been shown to be a feasible approach to balancing a player's physical abilities with the exergame difficulty level. This consequently enhanced the user experience and training effects of the exergame [50, 51, 39, 34]. These results are supported by the thorough, pre-existing research of HR within the field of sport and exercise science [27]. These findings can inform the design of the current study, as HR is a relatively straight-forward and cost-effective way to implement exercise intensities within an exergame. This would be achieved through players wearing a Bluetooth wearable device that can send their HR data to the game engine in real-time.

2.2.4 Cognitive DDA Measures. The implementation of cognitive measures for DDA has been studied in the literature less than physiological measures. However, increasing or decreasing the complexity of a sequence is a common method to adjust cognitive difficulty. For example, the difficulty levels within the popular exergame Beat Saber, although not dynamic, function primarily on this principal.

A study by Huber et al. [20] explored the use of Procedural Content Generation (PCG) for cognitive DDA in exergames. This appears to be the first attempt in the literature to implement reinforcement learning methods for DDA in a VR-based exergame. The authors developed a maze environment where the cognitive challenge derives from the players' restricted vision over the walls of the maze. This therefore forces the players to explore the maze whilst trying to remember the directions from the path they have taken. The implemented reinforcement learning methods aim to control the PCG and create varying maze structures that match the player's cognitive ability. The user study design gave players three attempts in the game and players could comment after each attempt whether they were satisfied with the difficulty. The results of the study show that those players who were unsatisfied in the first maze were then satisfied in the second, showing the ability of their agents to learn and use PCG effectively. However, one of the main problems reported in the study was that the generated levels took too long to allow for fast adaption. In exergaming, the need for fast adaption is even greater due to the increased speed of players' movement during physical exercise. This is therefore unlikely to be an appropriate method for the DDA of the current study.

The use of cognitive DDA was also observed in educational computer games. Sampayo-Vargas et al. [46] reported that the difficulty of their game was adjusted through increasing or decreasing the difficulty of the educational content which, as in Beat Saber and the maze by Huber et al. [20], can be categorised by its complexity. Within the educational context, Sampayo-Vargas stated that decreasing the difficulty would allow players more time to learn the content. As the current study is not focusing on educational material, this does not carry across, however the approach of a simple cognitive measure is one that appears successful and could be implemeted in the VR exergame.

2.3 Dual Flow Model (DFM)

Martin-Niedecken et al. [31] introduced the DFM which uses both physiological and cognitive measures (as discussed above) to adjust the game difficulty. Based on the DFM model, Martin-Niedecken and Götz [33] designed and evaluated the DDA exergame 'Plunder Planet' for children. This game used player HR (implemented using a HR chest strap) as the physiological measure and the player's performance (game accuracy) for the cognitive measure. The in-game adjustments based on these measures included increasing/decreasing game tempo and/or the speed of virtual obstacles for the physiological side, and flattening/curving the track and/or the difficulty of overcoming obstacles for the cognitive side of the DFM. A potential disadvantage of these cognitive adjustments is that they may impact a player's physiological measurements. However, there are many advantages to Martin-Niedecken and Götz' exergame design. These were demonstrated in their advanced study which showed that the DDA exergame Plunder Planet was significantly better than the non-adaptive version for game flow, dual flow, motivation, enjoyment, spatial presence, and players' physiological responses.

An additional exergame development that implemented the DFM was the ExerCube [35]. This aimed to challenge players by using whole body exercises (e.g., squats, lunges, and burpees) as well as the additional cognitive challenge for the player of quickly processing and reacting to track information (i.e., coordination challenges). As with Plunder Planet, the ExerCube DFM DDA design used continuous HR tracking (implemented with a HR chest strap) for the physiological measurements. This was defined as measuring the training intensity of the exergame. A pre-defined target heart rate zone (i.e., a percentage of HRmax) was used to increase or decrease the exercise difficulty, gaming speed, and frequency of obstacles in real-time. The players' cognitive measurements were measured in the ExerCube through game accuracy (reacting to visual stimuli at the right time). The display timing of the next movement direction was increased or decreased based on this performance. Martin-Niedecken et al. [30] found an early-stage prototype of the ExerCube to match the results of personal training in relation to levels of immersion, motivation, and flow. These results were supported by a later version of the ExerCube being shown to be a feasible tool for inducing HIIT (High Intensity Interval Training) intensity [42]. Nes et al. [42] reported that although average HR was significantly lower in the ExerCube condition than in the conventional functional HIIT condition, maximal HR reached equivalent medium-high levels for both conditions. Most importantly for the current study, the ExerCube session resulted in significantly better results for flow, enjoyment, and motivation.

2.4 Emotion Recognition

One element that the DFM does not cover is emotion recognition. There is theoretical and empirical evidence in the field of educational psychology showing that emotions determine learning processes (e.g., directing attention), and as a result performance [55]. Furthermore, being able to adapt games to players' emotions has been shown to improve task engagement and persistence [55]. The challenge facing emotion-sensitive adaptive games lies in the accuracy of monitoring emotional states.

Müller et al. [40] researched emotion responses (e.g., fear/surprise) to jump scare events in a cycling exergame using facial expression detection and electrodermal activity measurements. In Virtual Reality gaming, VR headsets impair facial expression detection, so psychophysiological measures must be used to recognise emotions in this field of gaming.

Other previous studies which have determined players' emotional state based on psychophysiological measures used fMRI [2], facial EMG [6] , EEG [61] and electrocardiography (ECG) [41]. However, most of these sensors are too sensitive to be used in an exergaming environment due to motion artefacts, and therefore shall not be considered in the current study. Instead, the eye tracking (pupil dilation, blink rate, blink duration) and GSR (skin conductance) will be used.

There is currently no evidence in the literature of emotion recognition in VR exergaming. This study therefore aims to determine whether psychophysiological measures can provide accurate emotion recognition for low, moderate, and vigorous intensity VR exergaming.

3 EXERGAME DEVELOPMENT

3.1 Software Requirements

As stated in RQ1, the primary objective of the current study is to try and map psychophysiological data to the emotions that have been elicited by the exergame. The exergame therefore needs to be able to isolate emotions or ensure that an emotion is dominant within a scene. To achieve this, the exergame will have four separate scenes, with each targeting an emotion from a separate quadrant of the circumplex model. These emotions will be stress (high arousal, negative valence), sadness (low arousal, negative valence), serenity (low arousal, positive valence), and joy (high arousal, positive valence).

To allow participants to transition between emotions and reduce carry-over effects, the exergame should have a neutral scene played between each emotion scene. This neutral scene should elicit no particular emotion. In addition, to ensure consistency of the study protocol, both the emotion and neutral scenes should be 1-minute in length.

Furthermore, to answer RQ2, the study protocol should use three different exercise intensities: low, moderate, and vigorous. These will be determined for each participant as a percentage of their HRmax. The exergame must therefore be able to increase and decrease exercise intensity to maintain the specified HR. The exercise intensity will be controlled through the Power (Watts) set on the exercise bike and will be constant during each exercise bout (using Erg Mode, Wahoo KICKR bike).

To obtain the psychophysiological data required for this study, the exergame will use the Vive Pro Eye VR headset, which measures eye tracking (pupil dilation, blink rate, blink duration), and a Consensys GSR Development Kit shimmer device, which measures skin conductance. A Polar HR chest strap connected via Bluetooth to the Android HR and HRV logger application (v1.3.5) will be used to measure HR.

3.2 Design of Software System and Implementation

The four exergame scenes (stress, sadness, serenity, joy) were extensively researched and user tested during the development phase. Visual changes to the environment were used such as different terrains, objects (e.g., colourful birds, distressed trees), the proximity of objects to the user, and light exposure/colours. These features were combined with auditory effects such as emotion-inducing soundtracks including music, siren noises, and barking dogs, depending on the scene. The abstractions of the features developed for each scene can be found below.

3.2.1 Negative Valence Quadrants. The research by Geslin et al. [17] on how colour and game properties can be used to elicit emotions in virtual games informed the scene development in the current study. Geslin states that the game environment for the negative valence quadrants of the circumplex model (fear/stress and sadness/negative quiet) should contain the following features: desaturated colours, darkness, dirt, no trade, and loss. To achieve desaturation, less saturated, darker colours were used in the stress and sadness scenes of the VR exergame compared to the joy and serenity scenes. The colour choices were informed by Dharmapriya et al. [13] who mapped Itten's colour system to Russell's circumplex model of affect. Dharmapriya identified the main colour that should be selected per quadrant of the circumplex model; with pink being used for the stress scene and blue for the sadness scene [13]. These colour choices were also supported by research on how colours can be used in constructing emotions by interactive digital narratives [54].

The feature of dirt [17] was also incorporated into both the stress and sadness scenes. The sadness scene was designed with a mud landscape with sparse distribution of dead grass and distressed, dirty road objects. The stress scene landscape was also designed to look barren, with a more sand-like appearance and the use of burnt trees, rocks, and fire.

To create the game aspect of the exergame, coins were added to the stress, sadness, and joy scenes. The user could collect the coins by tilting their head to the left or right depending on the coin location. For each coin collected, the user would typically gain one point and hear a positive reward sound effect. However, for the stress scene, skull coins were added to introduce the feature of loss [17]. The user was instructed to avoid these coins as they would lose a point for each one collected. This holds similarities to the affect recognition study by Barathi et al. [1] where points were deducted if a user collided with traffic. The sound effect for the collection of a skull coin differed to the gold coins with a harsh horn sound effect being used. For the sadness scene, a different interpretation of loss was implemented, whereby users didn't lose points but instead the frequency of coins available to collect was greatly reduced. This was intended to create a feeling of loss in comparison to the other scenes where higher scores could be achieved. The coins in the sadness scene were also adjusted to have a rusted appearance, as well as having the sound effect of a pebble dropping to the floor when collected.

Due to developing a non-competitive exergame (i.e., with no opponent) [37], the feature of no trade [17] was not incorporated. However, four of the five features which Geslin reported for the negative valence quadrants were implemented, so the lack of a no trade feature is unlikely to undermine the construction of the targeted emotions. Furthermore, additional features were added to both the stress and sadness scenes following their successful implementation in a previous exergaming study [1]. These included a police car with siren sound effects and a chasing barking dog in the stress scene, and heavy rain and mist in the sadness scene.

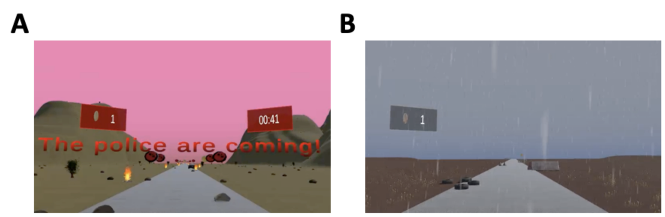

Figure 2: Scenes of developed VR exergame: (A) Stress Scene; (B) Sadness Scene

3.2.2 Positive Valence Quadrants. The research by Geslin et al. [17] states that for the positive valence quadrants of the circumplex model (excitation/joy and serenity/positive quiet), the game environment should contain the following features: strong colours, lights, interactions, sharing, creation, and earnings. As with the negative valence scenes (i.e., stress and sadness), the colour choices were informed by Dharmapriya et al. [13] and Taveter & Taveter [54], with turquoise being used for the serenity scene and orange for the joy scene. To achieve strong colours, greater saturation was used for the joy and serenity scenes compared to the stress and sadness scenes. Wildflowers were also added to the joy and serenity scene landscapes. Using directional light, shadows gently passed over these flowers to achieve the feature of natural light [17].

The feature of interactions [17] was achieved through the use of text in both the joy and serenity scenes. The text was set to change every 10-seconds throughout the scene to maintain the interaction with the user and was colour matched using the same informed approach as the sky [13, 54]. For the joy scene, the text included motivational and positive messages such as ’10 Bonus Points!’ or ’Keep Up The Good Work!’. For the serenity scene, the text interactions were more meditative including breathing exercises or text such as ’Calm your mind’.

As with the no trade feature [17] in the negative valence scenes, the sharing and creation features [17] of the positive valence scenes were restricted due to the style of exergame developed in this study. For example, with the exergame being non-competitive, the user did not have an opponent to share their coins or experience with. Furthermore, creation features commonly involve use of handheld controllers to pick up and move objects [23]. As the game play of this exergame is performed through head tilts due to the user’s hands holding the handlebars of the exercise bike, this reduced the ability to incorporate creation features.

The feature of earnings [17] was achieved in the joy scene through the addition of heart shaped gems and the occurrence of multiple coins/gems in a row. The gems rewarded the user with 10 bonus points when collected compared to the 1 point per gold coin collected. Providing multiple coins/gems in a row allowed the user to greatly increase their earnings, reaching 25 times more than in the sadness scene. A strong pink colour was also used for the heart-shaped gems to further strengthen the strong colours feature [17]. For the serenity scene, coin collection was replaced with the instruction to ’Gently pedal for points’. This allowed the user to be rewarded with one point every 2-seconds they were cycling. A fundamental concept of serenity is minimising distractions and to focus on one thing [11]. It was therefore important to allow the user to focus solely on experiencing the environment without the distraction of coin collection, whilst maintaining their ability to earn points.

Figure 3: Scenes of developed VR exergame: (A) Joy Scene; (B) Serenity Scene

3.2.3 Soundtrack Development. The soundtracks used in each of the emotion scenes were specifically developed for this study using Garageband (Apple, MacBook Pro). The soundtracks were composed following the research by Fernández-Sotos et al. [14] and Liu et al. [26] which mapped music tempo and note length to the circumplex model. This identified that the soundtracks for the high arousal quadrants of the circumplex model (i.e., stress and joy) should have a tempo of 150 beats per minute ("bpm") and use sixteenth notes, whilst the low arousal quadrants (i.e., sadness and serenity) should use 90bpm and whole/half notes. These tempos are similar to those used in the optimal condition of the affect recognition study by Barathi et al. [1]. Appropriate sound samples were selected for each emotion and adjusted to the respective tempos for their scene. Percussive backing tracks were then created at the relevant tempo and focussed on emphasising the note lengths required for that scene. Additional elements were added to each soundtrack to thicken the texture. These included upbeat vocal chants/sound effects for the joy soundtrack, discordant high-pitched notes for the stress soundtrack, and distant melancholic sound effects for the sadness soundtrack.

Developing the soundtracks specifically for this study also allowed us to manage the transitions between each soundtrack. As 120bpm is exactly halfway between 150bpm (stress/joy) and 90bpm (sadness/serenity), each track started and ended with this tempo and increased or decreased to 150bpm/90bpm respectively within 5-seconds. This allowed the soundtracks of each emotion scene to transition smoothly.

3.2.4 Additional Features. The research by Geslin et al. [17] also stated that movement speed should increase between scenes of low arousal and high arousal. To achieve this in the current study, three environment speeds were used: (i) Fastest (for the stress and joy scenes), (ii) Middle (for the neutral scene), and (iii) Slowest (for the sadness and serenity scenes). The use of a slower environment speed for the sadness scene was also supported by Geslin’s research which reported several game features that help to distinguish between scenes of high arousal, positive valence (i.e., joy) and low arousal, negative valence (i.e., sadness). Of these features, slow movements were noted to enhance the low arousal, negative valence scenes.

The slow movements feature was strengthened in the sadness scene through the reduced number of coins to collect. This reduced both the amount of head movements for the user during the scene as well as the movement speed required to collect each coin as there was no significant time pressure for the user to quickly tilt their head to the other side to collect/avoid another coin. Additional features that also enhanced low arousal, negative valence scenes were enclosed spaces, close-ups, no communication, and obedience/docility [17]. Close-ups were implemented by strategically placing the distressed road objects on the road to ensure that they were the dominant object in the user’s vision as they cycled by. Furthermore, heavy rain and mist was added surrounding the user resulting in regular close-ups as the rain falls. This is in contrast to the serenity scene for example, where there was an open plane of grass between the user and the forest. The no communication feature was achieved by not including the central in-game text as seen in the joy and serenity scenes. The final feature of obedience/docility was achieved by removing the possibility for the user to be disobedient. For example, in the stress scene the user could intentionally collect the skull coins despite these causing the user to lose points. By only having coins that would reward the user in the sadness scene, they only had the option to be obedient/docile.

The game features that enhance scenes of high arousal, positive valence (i.e., joy) were reported by Geslin et al. to be rapid movements, wide open spaces, wide shots, social-communication, and action [17]. Rapid movements were encouraged in the developed exergame through increasing the number of coins/gems to collect in the joy scene. If the user were to collect all of these objects, it would require fast head movements from the left- to right-tilts. Furthermore, a wide variety of game objects were included in the joy scene such as white rabbits, a range of colourful of birds, hot air balloons, a wildflower meadow, gold coins, and heart-shaped gems. This depth of game objects encouraged the user to look around the scene, increasing the speed of their head movements. The design of the low-profile rolling wildflower meadow achieved the features of wide open spaces and wide shots. The inclusion of the hot air balloons also contributed to the achievement of these two features as it helped to draw the user's view upwards, expanding their horizons. Several of the game objects also contributed to the action feature through the use of animations. For example, the white rabbits had both idle and hopping animations, and the colourful birds had pecking animations and sounds of birdsong. Finally, despite the exergame being non-competitive [37], the social communication feature was attempted to be implemented through the use of encouraging on-screen text.

User feedback was obtained for the four emotion scenes and the neutral scene to ensure that the requirements specification was met, and that each scene effectively elicited the targeted emotion. Adjustments following this feedback included the addition of stressful text and a countdown timer to the stress scene. The red text colour in the stress scene was chosen using the same informed approach as the sky colours [13, 54].

3.2.5 Neutral Scene. The neutral scene was used as both a transition between each of the emotion scenes (in order reduce any carry-over effects of emotion from the previous scene), as well as a period to verbally collect the ground truth data. As stated in the software requirements (section 3.1), the neutral scene aimed to elicit no particular emotion. To achieve this, the sky colour was set to white due to this shade being at the centre of Ittin's colour system and of Russell's circumplex model (i.e., no emotion) when mapped together by Geslin et al. [17]. Additional design elements of the neutral scene included a flat elevation profile, plain green terrain, and no objects of interest. To transition smoothly between scenes, a colour lerping function was used.

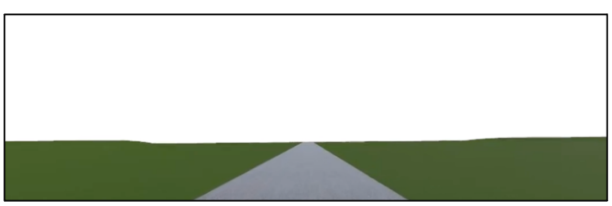

Figure 4: Neutral Scene in developed VR exergame

4 METHODOLOGY

29 participants were recruited for this study (16 male, 13 female, age 19-33, mean 25±3), who were students and employees of the University of Bath. We screened the participants with the Physical Activity Readiness Questionnaire (PAR-Q) and the VR questionnaire (from VR Health and Safety' SOP, University of Bath Department of Psychology). Participants were excluded if they answered 'Yes' to any questions without doctor review. The remaining participants were informed about all experimental conditions and asked to complete a demographics questionnaire. This sample size allowed us to counterbalance the order of the 3 x exercise intensities and 4 x emotion scenes using the Balanced Latin Square design.

The laboratory setup for user testing required the following apparatus: 1 x HTC Vive Pro Eye VR headset (inbuilt eye tracking), 2 x Lighthouses and tripods, 1 x HTC Vive hand controller, 1 x Wahoo KICKR stationary exercise bike, 1 x Polar HR chest strap, 1 x Android phone with HR & HRV logger application installed (real-time Bluetooth connection), 1 x Consensys GSR Development Kit shimmer device, 1 x University of Bath computer with Unity engine, SteamVR, Qualtrics, and Consensys access.

4.1 Design of Experiment

For user testing, a within subject design was used. The exergame developed for this study required participants to cycle on a stationary Wahoo KICKR exercise bike whilst wearing a Vive Pro Eye VR headset with eye tracking, a Polar HR chest strap, and a Consensys GSR Development Kit shimmer device to measure skin conductance. Each participant attended one 60-minute session. Each session included one low, one moderate, and one vigorous intensity exercise bout.

The exercise intensities were calculated for each participant prior to testing as percentages of their HRmax. To do this, we used Fox et al.'s age-predicted maximal heart rate ("APMHR") formula of $220 - age$. Although the literature offers arguments for and against APMHR formulae, this was appropriate for our study due to its simplicity. For example, the APMHR formulae does not require maximal testing such as graded treadmill exercise tests to exhaustion. Using this kind of maximal testing would have significantly narrowed our participant pool to those of higher trained fitness levels which was not desired.

This study used the Fox equation based upon research by Shookster et al. [47] which concluded that the Fox equation was the strongest APMHR option for a general population as it is less likely to under or overestimate based on individual HRmax. Heart rate thresholds were then calculated for the low, moderate, and vigorous exercise intensity thresholds as a percentage of HRmax for each participant. We used 57% of HRmax for low intensity exercise, 64% for moderate, and 77% for vigorous.

Each exercise bout followed the format: 2-minute warm-up followed by four 1-minute periods of exercise (one for each emotion scene), with a 1-minute recovery period between each (see Figure 5).

The 2-minute cycling warm-up was conducted at the beginning of each exercise bout to reduce injury risk and reach the participant's target HR. The resistance of the Wahoo KICKR was incrementally increased until the participant reached their target HR for the current exercise intensity. The HR logger was started at the beginning of the warm-up for each exercise bout (see Figure 5, Time point 1). The neutral scene was played in the Vive Pro Eye headset during the warm-up to minimise the emotion felt prior to the emotion scenes starting.

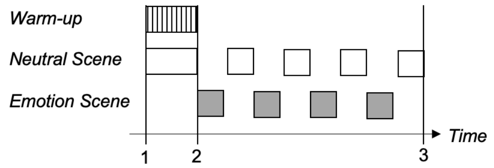

Figure 5: Experimental protocol. Scan QR code for YouTube video of VR exergame experience

The order of the emotion scenes (stress, sadness, serenity, joy) and exercise intensities (low, moderate, vigorous) were counterbalanced using the Latin square design to minimise the impact of any carry over or familiarisation effects on the results. For each exercise intensity, the recordings for the Consensys GSR Development Kit shimmer device and the HTC Vive Pro Eye eye tracking were started at the beginning of the first emotion scene (see Figure 5, Time point 2), and stopped along with the HR logger after the last (fourth) recovery period (see Figure 5, Time point 3).

During the recovery periods of each exercise bout, the resistance of the Wahoo KICKR was lowered before participants were asked to verbally rate the intensity at which they felt 8 different emotions and presence during the previous 1-minute exercise period on a scale of 1-10. The 8 emotions included the target emotions of each scene (stress, sadness, serenity, joy) and one other emotion from each quadrant of the circumplex model (fear, negative quiet, positive quiet, excitation). The verbally reported values for these 8 emotions created the ground truth database. Prior to the next emotion scene, the Wahoo KICKR resistance was increased again to bring the participant's HR back to their target for that exercise intensity. A 10-minute break was provided between each exercise bout to avoid fatigue and minimise any negative effects of the VR headset. The Intrinsic Motivation Inventory and flow experience questionnaires were completed during the break between each exercise bout. A post questionnaire was done upon completion of all three exercise intensities. This included a 10-part Big Five Personality Inventory and an opportunity for qualitative feedback.

4.1.1 Participant Equipment Setup. Prior to each exercise bout, the participant was provided with a Polar HR chest strap and was informed that this required skin contact and should be placed just below the chest. The placement of the chest strap, and its successful Bluetooth connection, was checked by the researcher for each participant. The Consensys GSR Development Kit shimmer device was then fitted to the right-hand ring- and middle-finger for each participant (see Figure 6A). The participant was then asked to mount the Wahoo KICKR exercise bike and the saddle height was adjusted. This required the participant to place their feet on the pedals, with one leg at full extension. The saddle height was then adjusted until an approximate angle of 30-40 degrees was achieved at the knee and no hip rocking was felt. The fore/aft position of the saddle was adjusted until the participants knee was vertically above the pedal axle when the crank was in the foremost horizontal position.

The HTC Vive Pro Eye headset was then put on the participant and the tension adjusted until the headset was both stable and comfortable. The participant was then asked to read text of gradually decreasing font sizes in the headset. If the text was unclear, the inter-pupillary distance of the headset was adjusted until the participant could read all text clearly. Figure 6B demonstrates the full laboratory setup for the participant.

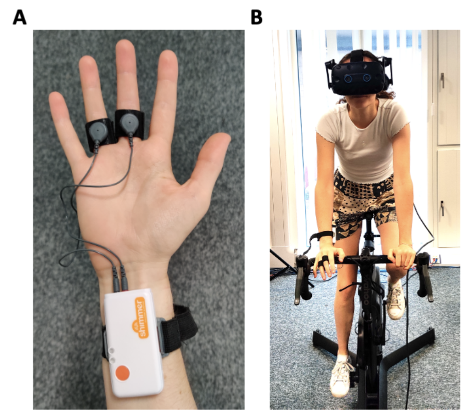

Figure 6: Participant setup for (A) Consensys GSR device; (B) Full VR cycling exergame session

An experiment was judged successful if the participant completed the full 8-minute exercise bout for all three exercise intensities (low, moderate, and vigorous).

4.2 Statistical Analysis

The psychophysiological data and the ground truth database were analysed both across-scene (i.e., did the stress scene elicit significantly greater stress than all other scenes?) and within-scene (within the stress scene, was stress elicited significantly more than all other emotions?) using repeated-measures ANOVAs. The η^2 measure was used for effect sizes and two-tailed t-tests with Holm correction for pairwise comparisons.

A Generalised Least Squares ("GLS") regression was then performed for each of the 8 measured emotions (fear, stress, sadness, negative quiet, serenity, positive quiet, excitation, joy) using the nlme package in R Studio [RStudio Team (2020), R v4.2.1]. This was to show whether the four psychophysiological variables (pupil dilation, blink rate, blink duration, and skin conductance) could significantly predict the emotion in question. Each GLS regression used the following formula, with 'Emotion' replaced with the emotion in question:

$gls(Emotion \sim Pupil\_Dilation + Blink\_Rate + Blink\_Duration + Skin\_Conductance + Average\_HR, data = dl, na.action = na.exclude, corr = corCompSymm(form = 1|Participant))$

Because of our within-participants design, we accounted for the individual differences between participants in both psychophysiological and ground truth measurements using the corr argument. This increased the statistical power as no aggregation of measurements was necessary when testing intra-individual hypotheses [3]. The level of significance used for all tests was $\alpha = .05$. Plots and tables show 95% confidence intervals of the means.

5 RESULTS

5.1 Environment Effectiveness

5.1.1 Across-Scene Analysis. The ANOVAs and t-tests were able to detect large-sized effects (η^2 >.14 and $Cohen's\_d > .8$ respectively).

For stress, the main effect of scene was significant ($F(2, 56) = 121.15, p < .001^{***}$) with a large effect size ($\eta^2 = 0.693$). Post Hoc comparisons were then conducted and revealed that the stress scene elicited significantly greater stress than all other scenes ($p < .001^{***}$), with a large effect size.

For joy, the main effect of scene was significant ($F(2, 64) = 124.86, p < .001^{***}$) with a large effect size ($\eta^2 = 0.692$). Post Hoc comparisons were then conducted and revealed that the joy scene elicited significantly greater joy than all other scenes ($p < .001^{***}$), with a large effect size.

For sadness, the main effect of scene was significant ($F(2, 47) = 86.90, p < .001^{***}$) with a large effect size ($\eta^2 = 0.663$). Post Hoc comparisons were then conducted and revealed that the sadness scene elicited significantly greater sadness than all other scenes ($p < .001^{***}$), with a large effect size.

For serenity, the main effect of scene was significant ($F(3, 81) = 100.94, p < .001^{***}$) with a large effect size ($\eta^2 = 0.658$). Post Hoc comparisons were then conducted and revealed that the serenity scene elicited significantly greater serenity than all other scenes ($p < .001^{***}$), with a large effect size.

These results are represented below in Figure 7, which illustrates which scene elicited the strongest stress, sadness, serenity, or joy respectively.

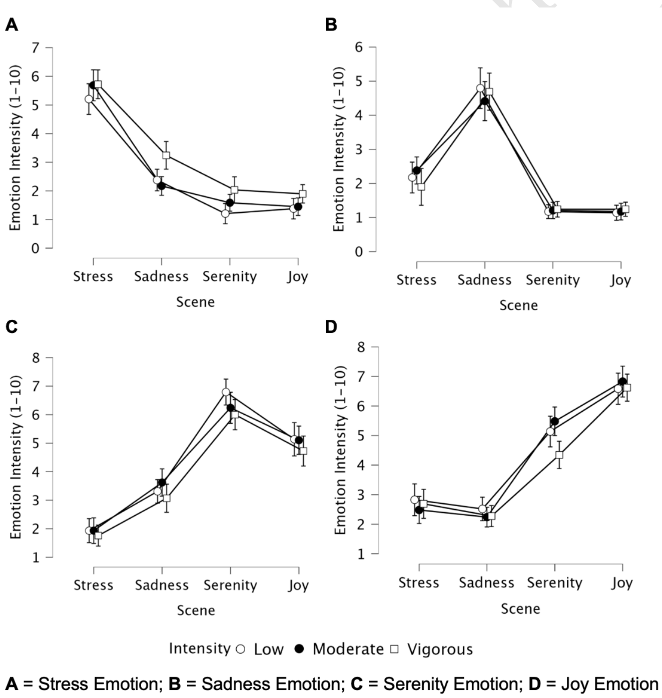

Figure 7: Across-scene emotion analysis with 95% CI

5.1.2 Within-Scene Analysis. For the stress scene, the main effect of emotion was significant ($F(3, 97) = 33.34, p < .001^{***}$) with a large effect size ($\eta^2 = 0.430$). Post Hoc comparisons were then conducted and revealed that stress was elicited significantly more than all other emotions in the stress scene ($p < .001^{***}$), with a large effect size.

For the joy scene, the main effect of emotion was significant ($F(3, 79) = 144.98, p < .001^{***}$) with a large effect size ($\eta^2 = 0.771$). Post Hoc comparisons were then conducted and revealed that joy was elicited significantly more than all other emotions in the joy scene ($p < .001^{***}$), with a large effect size.

For the sadness scene, the main effect of emotion was significant ($F(3, 83) = 17.64, p < .001^{***}$) with a large effect size ($\eta^2 = 0.312$). Post Hoc comparisons were then conducted and revealed that sadness was elicited significantly more than fear, stress, positive quiet, excitation, and joy ($p < .001^{***}$), with a large effect size, and significantly more than serenity (p < .01**), with a medium effect size ($d = 0.752$) in the sadness scene. There was no significant difference found between levels of sadness and negative quiet elicited in the sadness scene.

For the serenity scene, the main effect of emotion was significant ($F(3, 79) = 104.74, p < .001^{***}$) with a large effect size ($\eta^2 = 0.708$). Post Hoc comparisons were then conducted and revealed that serenity was elicited significantly more than all other emotions except positive quiet in the joy scene ($p < .001^{***}$), with a large effect size.

These results are represented below in Figure 8, which illustrates which of the eight measured emotions was dominant within each of the four scenes (stress, sadness, serenity, joy).

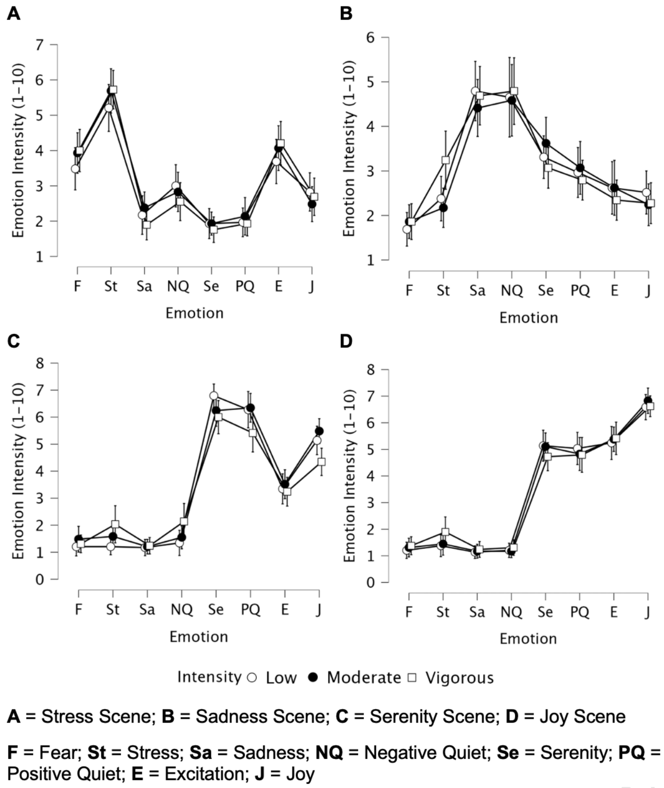

Figure 8: Within-scene emotion analysis with 95% CI

5.2 Psychophysiological Emotion Models

Table 1. Generalised Least Squares regression analyses with 95% CI and 342 DF. (p <.05 = *, p <.01 = **, p <.001 = ***).

	Pupil Dilation	Blink Rate	Blink Duration	Skin Conductance
Fear	1.49***	-0.23	0.04	0.04
	[1.16, 1.81]	[-0.52, 0.05]	[-0.09, 0.18]	[-0.04, 0.13]
Stress	2.36***	-0.33	0.18*	0.13*
	[1.92, 2.80]	[-0.71, 0.05]	[0.00, 0.35]	[0.02, 0.25]
Sadness	0.72***	-0.24	0.16	0.02
	[0.38, 1.07]	[-0.60, 0.13]	[-0.02, 0.33]	[-0.11, 0.06]
Negative Quiet	0.81***	-0.24	0.17	0.05
	[0.39, 1.23]	[-0.66, 0.17]	[-0.02, 0.37]	[-0.06, 0.15]
Serenity	-2.35***	0.39	-0.09	-0.08
	[-2.82, -1.89]	[-0.02, 0.80]	[-0.28, 0.10]	[-0.20, 0.04]
Positive Quiet	-2.35***	0.31	-0.02	-0.12*
	[-2.81, -1.89]	[-0.10, 0.71]	[-0.21, 0.17]	[-0.24, -0.00]
Excitation	-0.25	0.17	-0.18*	-0.06
	[-0.67, 0.18]	[-0.21, 0.55]	[-0.35, -0.00]	[-0.17, 0.05]
Joy	-1.57***	0.42	-0.21	-0.13*
	[-2.02, -1.13]	[-0.02, 0.86]	[-0.42, 0.00]	[-0.24, -0.01]

The effect of pupil dilation was significant for all emotions except excitation ($p < .001^{***}$). The effect of blink duration was significant for stress and excitation ($p < .05^{*}$). The effect of skin conductance was significant for stress, joy, and positive quiet ($p < .05^{*}$).

6 DISCUSSION

To be able to answer RQ1 as to whether psychophysiological measurements can be used to predict emotions in VR exergaming, the developed game environment had to be effective on two grounds: (i) be able to elicit the target emotions in the correct scenes (e.g., feeling stressed in the stress scene), and (ii) ensure that the dominant emotion within each scene was the targeted emotion (e.g., within the stress scene, stress was elicited significantly more than all other emotions). The across-scene analysis shows that the environment successfully achieved point (i), with each target emotion being elicited significantly more than each of the other emotions in the relevant scene ($p < .001^{***}$). This shows that the game features selected for each emotion scene were effective at eliciting the targeted emotion. However, it was critical to the environment's success that point (ii) also be achieved. This would ensure that the dominant emotion within each scene was the targeted emotion. The within-scene analysis shows that this was the case ($p < .001^{***}$). The combination of both the across-scene analysis and the within-scene analysis being significant with large effect sizes shows that the developed VR cycling exergame was effective at targeting specific emotions. This has important design implications for future VR exergame development, whereby this environment and its features can be used as an example of how to effectively target emotions.

The success of the developed exergame also spans across exercise intensity. Figure 7 & 8 demonstrate how emotion intensity was tightly clustered across low, moderate, and vigorous exercise intensities, resulting in no changes to the across-scene or within-scene analyses. This helps to answer RQ2, that psychophysiological measurements can be used to predict emotions at low, moderate, and vigorous exercise intensities in VR exergaming. However, the main challenge of emotion-sensitive adaptive games lies in the accuracy of predicting the emotions. This was shown in the current study through the GLS analyses.

The GLS regression analyses demonstrate that pupil dilation is a strong predictor of emotion, with pupil dilation being able to significantly predict all measured emotions except excitation ($p < .001^{***}$). Table 1 shows that pupil dilation increased for negative valence emotions, and decreased for positive valence emotions. This differs from the literature, where pupil dilation has been reported to increase with arousal due to the association between increased arousal and increased sympathetic activity [5, 19]. The current study results agreed with this literature for (i) high arousal, negative valence emotions (fear, stress) where pupil dilation increased, and (ii) low arousal, positive valence emotions (serenity, positive quiet) where pupil dilation decreased. However, disagreement was observed for (i) low arousal, negative valence emotions (sadness, negative quiet) where pupil dilation increased, and (ii) the high arousal, positive valence emotion (joy) where pupil dilation decreased. This suggests that valence may impact the relationship between pupil dilation and arousal, resulting in the reversed relationship directions seen.

The current study also found blink duration to significantly predict stress and excitation ($p < .05^{*}$). Barathi et al. [1] reported that psychophysiological measures related to blinks are more useful for predicting valence than arousal. On Russell's circumplex model [44], stress and excitation appear equally spaced either side of zero valence, with stress being negative and excitation being positive. Table 1 shows the fixed effect estimates of stress and excitation to be B = 0.18 and B = -0.18 respectively. These results match the positioning of the emotions on the circumplex model and also support the literature indicating that blink duration decreases with increased valence. However, if the relationship were this direct, we may have expected emotions of similar valence to also be significantly predicted by blink duration (e.g., positive quiet (similar positive valence to excitation) and negative quiet (similar negative valence to stress)). However, this was not the case. This suggests that blink duration may only significantly predict high arousal emotions across a narrow valence range. An additional blink variable was also measured in the current study - blink rate. However, the GLS analyses did not find this to be a significant predictor of any of the measured emotions.

The final psychophysiological factor investigated was skin conductivity. This was found to significantly predict stress, positive quiet, and joy ($p < .05^{*}$). Barathi et al. [1] reported that skin conductivity is mainly a measure of arousal. However, the results of the current study do not entirely support this. For instance, as stress and joy are both within the high arousal half of the circumplex model, it would be expected for skin conductivity to respond in the same direction for both emotions, according to Barathi. However, the fixed effects estimates for stress and joy were B = 0.13 and B = -0.13 respectively, which are opposite directions. Instead, joy closely matches the value and direction of the fixed effects estimate for positive quiet (B = -0.12), which is a low arousal emotion. However, joy and positive quiet are both also positive valence emotions, whilst stress is negative valence. This may suggest that the relationship between skin conductivity and emotion is in fact influenced by both valence and arousal.

The GLS regression results show which psychophysiological factors could be used for emotion-sensitive adaptive exergames in future. As blink rate was not a significant predictor of any of the 8 measured emotions, it would not be recommended to use this as an adaptive measure. In contrast, pupil dilation, blink duration, and skin conductance all contributed to significantly predicting emotions. However, these factors were not significant predictors for all of the measured emotions. We have therefore created a set of new GLS formulae using only the factors which were significant for each emotion. Fear, sadness, negative quiet, and serenity should be predicted with the following formula, replacing 'Emotion' with the relevant emotion being measured:

$gls(Emotion \sim Pupil\_Dilation+Average\_HR, data = dl, na.action = na.exclude, corr = corCompSymm(form =\sim 1|Participant))$

Joy and positive quiet should both be predicted with the following formula, again replacing 'Emotion' with the relevant emotion being measured:

$gls(Emotion \sim Pupil\_Dilation+Skin\_Conductance+Average\_HR, data = dl, na.action = na.exclude, corr = corCompSymm(form =\sim 1|Participant))$

Excitation should be predicted with the formula:

$gls(Excitation \sim Blink\_Duration + Average\_HR, data = dl, na.action = na.exclude, corr = corCompSymm(form =\sim 1|Participant))$

And finally, stress should be predicted with the formula:

$gls(Stress \sim Pupil\_Dilation+Blink\_Duration+Skin\_Conductance + Average\_HR, data = dl, na.action = na.exclude, corr = corCompSymm(form =\sim 1|Participant))$

Barathi et al. [1] demonstrated that combining psychophysiological measures, as seen in the new joy, positive quiet, and stress formulae, can build stronger predictors. Furthermore, the formula for joy and positive quiet includes pupil dilation which produced the highest significance level of all psychophysiological measures in the current study ($p < .001^{***}$). It is for these reasons that the joy and positive quiet formula may be the most effective at predicting emotion. However, a benefit mutual to all of these new GLS regression formulae is that the psychophysiological measures used are context agnostic, i.e., they can be used for any exergaming experience. This is in contrast to gaze fixations, which require semantic information about each scene.

6.1 Limitations/Future Work

Several of the current study's results were close to statistical significance ($p < .05^{*}$). Therefore, if a larger sample size had been used, the current study may have found more significant results. Calculating the ground truth database values with respect to presence may also have improved the result set, as it is possible that participants who felt less present may have felt the emotions at lower intensities.

For some participants, particularly those whose first language was not English, emotions such as serenity and negative/positive quiet proved difficult to understand during the initial verbal recall periods. The ground truth database could therefore have been made more reliable by providing participants with standard definitions of each measured emotion at the start of their session.

Although the current study was counterbalanced to account for the familiarisation of scenes and exercise intensities, it was not specifically investigated as to whether emotion intensities decreased as participants became familiar with the scenes. Future studies could test this using longer scene durations, for example 5-minutes/scene compared to 1-minute/scene as used in the current study. If this showed that emotion intensities did decline with scene familiarity, future work could look to counteract this with procedural content generation, using the environment elements which have been shown to be effective from this study.

Future work could also look at calculating an overall measure of emotion by using the circumplex model to create coordinate-like valence/arousal values for each emotion measured. The ground truth database would then act as weightings for these values to produce an overall value for the participant's emotion.

7 CONCLUSION

We developed a VR cycling exergame with four emotion-targeted scenes and proved its holistic effectiveness through significant across-scene and within-scene analysis. Building on this, we identified context-agnostic psychophysiological measures suitable for emotion-sensitive adaptive VR exergames and determined new single- and multi-sensor GLS regression formulae to help predict eight emotions from Russell's circumplex model. These contributions will allow future developers to scientifically build emotion-sensitive adaptive VR exergames that can optimise player experience and engagement.

We have also provided suggestions for future work, to help researchers further design and test VR exergame environments

and emotion predictors. In summary, we come to the following conclusions:
1. VR exergaming environments can be created to effectively target emotions,
2. Pupil dilation, blink duration, and skin conductance can be used to monitor emotional states at low, moderate, and vigorous exercise intensities in VR exergaming, and
3. The combination of the two points above allow for the future development of emotion-sensitive VR adaptive exergames.
These findings are of great importance for future developers as the implementation of real-time emotional responsiveness opens up a whole new dimension to the VR exergaming world.

REFERENCES

[1] SC Barathi, M Proulx, E O'Neill, and C Lutteroth.. *Affect recognition using psychophysiological correlates in high intensity VR exergaming in CHI 2020.* Conference on Human Factors in Computing Systems - Proceedings, Vol. 2020-April. Association for Computing Machinery.

[2] S Anders, M Lotze, M Erb, W Grodd, and N Birbaumer. 2004. Brain activity underlying emotional valence and arousal: A response-related fMRI study. *Human Brain Mapping* 23 (2004), 200–209. Issue 4.

[3] J. Z Bakdash and L. R Marusich. 2017. Repeated measures correlation. *Frontiers in Psychology* 8 (2017), 456.

[4] P Bosch, J Poloni, A Thornton, and J Lynskey. 2012. The Heart Rate Response to Nintendo Wii Boxing in Young Adults. *Cardiopulmonary Physical Therapy Journal* 23, 2 (2012), 13–18.

[5] M. M Bradley, L Miccoli, M. A Escrig, and P. J Lang. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology* 45 (2008), 602–607. Issue 4.

[6] J. T Cacioppo, R. E Petty, M. E Losch, and H. S Kim. 1986. Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of Personality and Social Psychology* 50 (1986), 260. Issue 2.

[7] T Charoensook, M Barlow, and M Lakshika. 2019. Heart rate and breathing variability for virtual reality game play. *in 2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH)* (2019), 1–7.

[8] J Chen. 2007. Flow in games (and everything else). *Commun. ACM* 50 (2007), 31–34. Issue 4.

[9] M Csikszentmihalyi. 2014. Toward a psychology of optimal experience. *in Flow and the foundations of positive psychology. Springer* (2014), 209–226.

[10] M Czikszentmihalyi. 1990. Flow: The psychology of optimal experience. *Harper & Row* (1990).

[11] M Czikszentmihalyi. 2009. Flow: The Psychology of Optimal Experience. *Harper Row* (2009).

[12] A Darzi, S McCrea, and D Novak. 2021. User Experience With Dynamic Difficulty Adjustment Methods for an Affective Exergame: Comparative Laboratory-Based Study. *JMIR Serious Games* 9, 2 (2021), 257–271.

[13] J Dharmapriya, L Dayarathne, T Diasena, S Arunathilake, N Kodikara, and P Wijesekera. 2021. Music emotion visualization through colour. In *2021 International Conference on Electronics, Information, and Communication (ICEIC)*. 1–6.

[14] A Fernández-Sotos, A Fernández-Caballero, and J. M Latorre. 2016. Influence of tempo and rhythmic unit in musical emotion regulation. *Frontiers in computational neuroscience* 10 (2016), 80.

[15] S.L Finkelstein, A Nickel, Z Lipps, T Barnes, Z Wartell, and E.A Suma. 2011. Astrojumper: Motivating exercise with an immersive virtual reality exergame. *Presence Teleoperators Virtual Environ.* 20 (2011), 78–92. Issue 1.

[16] C Garber, B Blissmer, M Deschenes, B Franklin, M Lamonte, I Lee, D Nieman, and D Swain. 2011. Quantity and Quality of Exercise for Developing and Maintaining Cardiorespiratory, Musculoskeletal, and Neuromotor Fitness in Apparently Healthy Adults. *Medicine & Science in Sports & Exercise* 43, 7 (2011), 1334–1359.

[17] E Geslin, L Jegou, and D Beaudoin. 2016. How color properties can be used to elicit emotions in video games. *International Journal of Computer Games Technology* 2016 (2016), 1–9. Issue 3.

[18] L Graves, N Ridgers, K Williams, G Stratton, G Atkinson, and N Cable. 2010. The Physiological Cost and Enjoyment of Wii Fit in Adolescents, Young Adults, and Older Adults. *Journal of Physical Activity and Health* 7, 3 (2010), 393–401.

[19] E. H Hess and J. M Polt. 1960. Pupil size as related to interest value of visual stimuli. *Science* 132 (1960), 349–350. Issue 3423.

[20] T Huber, S Mertes, S Rangelova, S Flutura, and E Andres. 2021. Dynamic Difficulty Adjustment in Virtual Reality Exergames through Experience-driven Procedural Content Generation. *IEEE Symposium Series on Computational Intelligence (SSCI)* (2021), 1–8.

[21] R Hunicke and V Chapman. 2004. AI for Dynamic Difficulty Adjustment in Games. *in Proceedings of the Challenges in Game Artificial Intelligence AAAI Workshop* (2004), 91–96.

[22] R Jacob, A Girouard, L Hirshfield, M Horn, O Shaer, E Solovey, and J Zigelbaum. 2008. Reality-based interaction: a framework for post-WIMP interfaces. *CHI '07 extended abstracts on Human factors in computing systems* (2008), 2465–2470.

[23] M Jennings-Teats, G Smith, and N Wardrip-Fruin. 2010. Polymorph: A model for dynamic level generation. *in Proceedings of the 6th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (2010), 138–143.

[24] S Ketelhut, L Röglin, E Kircher, A Martin-Niedecken, R Ketelhut, K Hottenrott, and K Ketelhut. 2021. The New Way to Exercise? Evaluating an Innovative Heart-rate-controlled Exergame. *International Journal of Sports Medicine* 43, 1 (2021), 77–82.

[25] R Koster. 2014. A theory of fun for game design. *OReilly Media* (2014).

[26] Y Liu, G Liu, D Wei, Q Li, G Yuan, S Wu, G Wang, and X Zhao. 2018. Effects of musical tempo on musicians' and non-musicians' emotional experience when listening to music. *Frontiers in psychology* 9 (2018), 2118.

[27] M Ludwig, K Hoffmann, S Endler, A Asteroth, and J Wiemeyer. 2018. Measurement, prediction, and control of individual heart rate responses to exercise—basics and options for wearable devices. *Front. Physiol.* (2018), 9–12.

[28] J Marshall and C Linehan. 2020. Are exergames exercise? A scoping review of the short term Effects of exertion games. *IEEE Trans. Games* 99 (2020), 1.

[29] A Martin-Niedecken, A Mahrer, K Rogers, E de Bruin, and A Schättin. 2020. "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).

[30] A.L Martin-Niedecken, K Rogers, L.T Vidal, E.D Mekler, and E.M árquez Segura. 2019b. Exercube vs. personal trainer: evaluating a holistic, immersive, and adaptive fitness game setup. *in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019b), 1–15.

[31] A Martin-Niedecken, T Schwarz, and A Schättin. 2021. Comparing the Impact of Heart Rate-Based In-Game Adaptations in an Exergame-Based Functional High-Intensity Interval Training on Training Intensity and Experience in Healthy Young Adults. *Frontiers in Psychology* 12 (2021).

[32] A.L Martin-Niedecken and A Schättin. 2020. Let the body 'n'brain games begin: toward innovative training approaches in esports athletes. *Front. Psychol.* 11 (2020), 1–15. Issue 138.

[33] A. L Martin-Niedecken and U Götz. 2016. Design and evaluation of a dynamically adaptive fitness game environment for children and young adolescents. *in Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts* (2016), 205–212.

[34] A. L Martin-Niedecken and U Götz. 2017. Go with the dual flow: evaluating the psychophysiological adaptive fitness game environment "Plunder Planet". *in Proceedings of the Joint International Conference on Serious Games* (2017), 32–43.

[35] A. L Martin-Niedecken and E. D Mekler. 2018. The exercube: participatory design of an immersive fitness game environment. *in Proceedings of the 4th Joint International Conference on Serious Games* (2018), 263–275.

[36] F Mueller, D Edge, F Vetere, M Gibbs, S Agamanolis, B Bongers, and J Sheridan. 2011. Designing sports: A framework for exertion games. *In Proc. of CHI 2011, ACM* (2011), 2651–2660.

[37] F Mueller, M Gibbs, and F Vetere. 2008. Taxonomy of Exertion Games. *In OzCHI 2008, ACM Press* (2008), 263–266.

[38] F Mueller, R Khot, K Gerling, and R Mandryk. 2016. Exertion Games. *Foundations and Trends® in Human–Computer Interaction* 10, 1 (2016), 1–86.

[39] F Mueller, F Vetere, M Gibbs, D Edge, S Agamanolis, and J Sheridan. 2012. Balancing exertion experiences. *in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), 1853–1862.

[40] L Müller, A Bernin, A Kamenz, S Ghose, K von Luck, C Grecos, Q Wang, and F Vogt. 2017. Emotional journey for an emotion provoking cycling exergame. In *IEEE 4th Intl. Conference on Soft Computing & Machine Intelligence (ISCMI 2017)*.

[41] M Naji, M Firoozabadi, and P Azadfallah. 2014. Classification of music-induced emotions based on information fusion of forehead biosignals and electrocardiogram. *Cognitive Computation* 6 (2014), 241–252. Issue 2.

[42] B Nes, I Janszky, U Wisløff, A Støylen, and T Karlsen. 2013. Age-predicted maximal heart rate in healthy subjects: the hunt fitness study. *Scand. J. Med. Sci. Sports* 23 (2013), 697–704.

[43] W Peng and J Crouse. 2013. Playing in Parallel: The Effects of Multiplayer Modes in Active Video Game on Motivation and Physical Exertion. *Cyberpsychology, Behavior, and Social Networking* 16, 6 (2013), 423–427.

[44] J Posner, J. A Russell, and B. S Peterson. 2005. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17 (2005), 715–734. Issue 2.

[45] K Salen and E Zimmerman. 2011. *Rules of Play: Game Design Fundamentals.* Johanneshov: TPB.

[46] S Sampayo-Vargas, C. J Cope, Z He, and G. J Byrne. 2013. The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Computers & Education* 69 (2013), 452–462.

[47] D Shookster, B Lindsey, N Cortes, and J Martin. 2020. Accuracy of commonly used age-predicted maximal heart rate equations. *International Journal of Exercise Science* 13 (2020), 1242–1250. Issue 7.

[48] Z Simpson. 2000. The In-game Economics of Ultima Online. *in Proceedings of the Game Developers Conference* (2000).

[49] J Sinclair, P Hingston, and M Masek. 2007. Considerations for the design of exergames. *in Proceedings of the 5th international conference on Computer graphics and interactive techniques in Australia and Southeast Asia* (2007), 289–295.

[50] J Sinclair, P Hingston, and M Masek. 2009. Exergame development using the dual flow model. *in Proceedings of the Sixth Australasian Conference on Interactive Entertainment* (2009), 1–7.

[51] T Stach, T Graham, J Yim, and R. E Rhodes. 2009. Heart rate control of exercise video games. *in Proceedings of the Graphics Interface* (2009), 125–132.

[52] A Stein, Y Yotam, R Puzis, G Shani, and M Taieb-Maimon. 2018. EEG-triggered dynamic difficulty adjustment for multiplayer games. *Entertainment Computing* 25 (2018), 14–25.

[53] J Sween, S. F Wallington, V Sheppard, T Taylor, A. A Llanos, and L. L Adams-Campbell. 2014. The role of exergaming in improving physical activity: a review. *Journal of physical activity & health* 11 (2014), 864–870. Issue 4.

[54] K Taveter and E. M Taveter. 2021. Case study on using colours in constructing emotions by interactive digital narratives. , 11 pages.

[55] C Tyng, H Amin, M Saad, and A Malik. 2017. The Influences of Emotion on Learning and Memory. *Frontiers In Psychology* 8 (2017).

[56] M Willems and T Bond. 2009. Comparison of Physiological and Metabolic Responses to Playing Nintendo Wii Sports and Brisk Treadmill Walking. *Journal of Human Kinetics* 22, 2 (2009), 43–49.

[57] S Xue, M Wu, J Kolen, N Aghdaie, and Zamanm K. A. 2017. Dynamic Difficulty Adjustment for Maximized Engagement in Digital Games. *in Proceedings of the the 26th International Conference* (2017), 465–471.

[58] S Yoo, P Gough, and P Kay. 2020. Embedding a VR game studio in a sedentary workplace: Use, experience and exercise benefits. *in CHI '20: CHI Conference on Human Factors in Computing Systems* (2020), 1–14.

[59] S Yoo, C Parker, and J Kay. 2017. Designing a personalized VR exergame. *in Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (2017), 431–435.

[60] M Zohaib. 2018. Dynamic Difficulty Adjustment (DDA) in Computer Games: A Review. *Advances in Human-Computer Interaction* (2018), 1–12.

[61] H Çelikkanat, H Moriya, T Ogawa, J-P Kauppi, M Kawanabe, and A Hyvärinen. 2017. Decoding emotional valence from electroencephalographic rhythmic activity. In *In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* 4143–4146.