Sweating the Details: Emotion Recognition and the Influence of Physical Exertion in Virtual Reality Exergaming

Dominic Potts
dmp59@bath.ac.uk
REVEAL, University of Bath
Bath, United Kingdom

Zoe Broad
zab26@bath.ac.uk
REVEAL, University of Bath
Bath, United Kingdom

Tarini Sehgal
ts2492@bath.ac.uk
REVEAL, University of Bath
Bath, United Kingdom

Joseph Hartley
jh3968@bath.ac.uk
REVEAL, University of Bath
Bath, United Kingdom

Eamonn O'Neill
maseon@bath.ac.uk
REVEAL, University of Bath
Bath, United Kingdom

Crescent Jicol
cj406@bath.ac.uk
REVEAL, University of Bath
Bath, United Kingdom

Christopher Clarke
cjc234@bath.ac.uk
REVEAL, University of Bath
Bath, United Kingdom

Christof Lutteroth
c.lutteroth@bath.ac.uk
REVEAL, University of Bath
Bath, United Kingdom

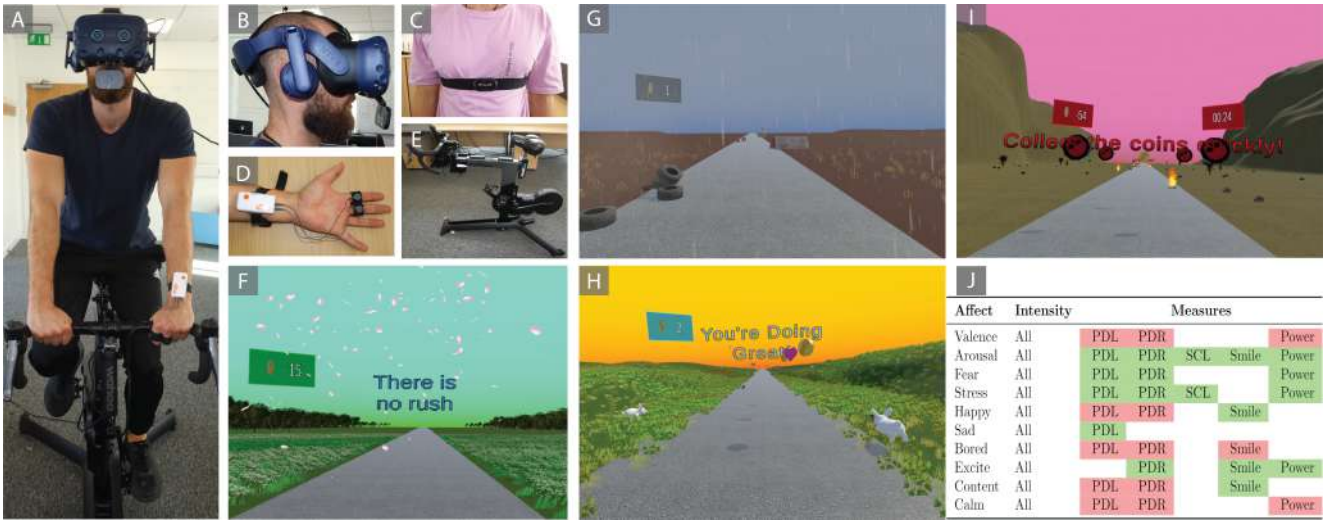

Figure 1: We systematically explored affect recognition in a VR cycling exergame (A) across three exercise intensity levels (low, medium, and high) by collecting affect ratings and physiological measures including gaze and facial gestures from a VR headset (B), heart rate (C), skin conductance (D), and power output from an exercise bike. Four VR exergaming environments were designed to elicit Calmness (F), Sadness (G), Happiness (H), and Stress (I). Linear regression models grounded in hypothesis testing reveal that Pupil Dilation Level (PDL), Pupil Dilation Response (PDR), Skin Conductance Level (SCL), smiling, and power output are all significant positive (green) or negative (red) predictors of different affective states (J).

ABSTRACT

There is great potential for adapting Virtual Reality (VR) exergames based on a user's affective state. However, physical activity and VR interfere with physiological sensors, making affect recognition challenging. We conducted a study (n=72) in which users experienced four emotion inducing VR exergaming environments (happiness, sadness, stress and calmness) at three different levels of exertion (low, medium, high). We collected physiological measures through pupillometry, electrodermal activity, heart rate, and facial tracking, as well as subjective affect ratings. Our validated virtual environments, data, and analyses are openly available. We found that the level of exertion influences the way affect can be recognised, as well as affect itself. Furthermore, our results highlight the importance of data cleaning to account for environmental and interpersonal factors interfering with physiological measures. The results shed light on the relationships between physiological measures and affective states and inform design choices about sensors and data cleaning approaches for affective VR.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642611

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

virtual reality, exergaming, emotion recognition, affect recognition, physiological sensing, psychophysiological correlates, high-intensity exercise

ACM Reference Format:

Dominic Potts, Zoe Broad, Tarini Sehgal, Joseph Hartley, Eamonn O'Neill, Crescent Jicol, Christopher Clarke, and Christof Lutteroth. 2024. *Sweating the Details:* Emotion Recognition and the Influence of Physical Exertion in Virtual Reality Exergaming. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3613904.3642611

1 INTRODUCTION

Regular physical activity helps to maintain a healthy weight, protects against chronic conditions, improves mental health, and increases quality of life [19, 160, 198]. Exergaming, the combination of physical exercise with gaming, holds great promise for incentivising physical activity [68]. Exergames can increase enjoyment and performance compared with conventional exercise by distracting users from uncomfortable sensations when nearing or exceeding the ventilatory threshold [122, 141, 147, 178, 197]. Virtual reality (VR) offers a unique platform for exergaming which can further distract users from the aversive elements of exercise by immersing them in engaging virtual environments [14, 15, 23, 24, 61, 84, 151].

The challenge exergames pose must be commensurate with a user's abilities to realise the benefits of increased enjoyment, immersion, and performance [44, 115, 170]. Adapting the difficulty of an exergame to the user helps them achieve a flow state, i.e. a psychologically optimal state in which they are focused and engaged [45, 56, 85]. For example, exergame difficulty can be adjusted in real time based on a user's heart rate, which can improve flow, enjoyment, and motivation [118] as well as exercise performance [117]. A more advanced method to control exergame adaptations is to estimate a user's emotional state during gameplay based on physiological sensor measures, known as *affect recognition* [139]. In addition to difficulty adjustment, affect recognition can be used to adapt exergames in unique ways such as interactive storytelling [35, 127], as well as having the potential to help us better understand the player experience [126, 128].

Exergaming presents key challenges for affect recognition due to the influences that physical exercise and interpersonal differences have on physiological measures and experienced emotions. First, emotion-inducing exergaming environments are needed to develop and validate affect recognition approaches. Some researchers have proposed such emotion-inducing environments [11, 15, 123, 129]; however, they have focused primarily on valence (i.e. how pleasurable it feels) or flow, considering only fragments of the emotional spectrum, and have not been validated across different levels of exercise intensity. Second, physical exertion influences many physiological measures, e.g., through increased cardiovascular activity, perspiration, and movement [21, 55, 67, 142, 143, 184]. However, there has been no rigorous, systematic comparison of affect recognition in VR exergames across different levels of exertion. Affect recognition has been explored in non-VR exergames only at moderate exercise intensities [31, 124, 126, 127] while research on high-intensity VR exergames has focused only on valence [15]. Third, when analysing physiological sensor data for affect recognition in VR exergames, we need to account for the influences of exercise, interpersonal differences, as well as environmental factors such as the stimuli from the VR exergame. For example, the changing luminance in virtual environments influences pupillary affect measures [146]. Removing these influences from sensor data can increase the robustness, predictive power, and generalisability of affect recognition models, which is crucial in 'noisy' contexts such as VR exergaming. However, it has been unclear how to do this for VR exergames, especially when considering different levels of exercise intensity. Finally, the study of affect in VR exergaming raises questions about the relationship between physical exertion and affect in a VR exergaming context. This paper extends our understanding of affect recognition in VR exergaming by investigating the following research questions:

RQ1 How can we manipulate affect in a VR exergame?
RQ2 How well do physiological measures predict affect during VR exergaming?
RQ3 How can environmental and interpersonal factors influencing physiological sensor data be accounted for?
RQ4 What is the relationship between physical exertion and affect during VR exergaming?

To address RQ1, we designed four virtual environments (VEs) for a VR cycling exergame to induce specific emotions (Happiness, Sadness, Stress, and Calmness), with each emotion representing a different quadrant of Russell's circumplex model of emotion [153, 154]. We then validated the VEs empirically and used them to elicit emotions in a user study (n=72), where participants cycled through the VEs at three different exercise intensities (low, medium, and high). To address RQ2, we analysed the relationships of 10 physiological measures and 10 self-reported ground-truth affect ratings for each VE at each level of intensity. To enhance our understanding of these relationships analytically and transparently, we used multi-level linear regression models grounded in hypothesis testing. These models are used to predict affect from physiological data, which have been found to bear linear relationships [108, 174, 175]. In contrast to machine-learning (ML) approaches, which are often "black box", regression models increase our fundamental understanding and inform other work on affect recognition. For example, Bota et al.'s review of ML-based emotion recognition finds that "*there is still no clear evidence of which feature combination of which physiological signals are the most relevant*" [26], and our regression models shed light on this in the context of VR exergaming. To address RQ3, we compared three different levels of sensor data cleaning: raw data, accounting for environmental factors, and also accounting for interpersonal differences. Finally, we addressed RQ4 by testing the relationships between physical exertion measures and affect, including intrinsic motivation, with linear regression models and analyses of variance. In summary, we make the following contributions:

(1) An openly available set of validated virtual exergaming environments to elicit four different emotions [144].
(2) Validated regression models describing how 10 sensor measures predict 10 types of affect across three levels of exercise

intensity, further evidencing the linear relationships between physiological responses and affect.
(3) Validated approaches for removing the influence of environmental and personal factors from physiological sensor data, which improve the predictive power of affect recognition and prevent model overfitting.
(4) Validated regression models describing the relations between physical exertion and affect.
(5) An open real-time data set (n=72) including physiological measurements and subjective affect ratings for the four VEs across the three levels of exercise intensity [145].
(6) The open source `EmoSense` framework [144] for the collection, cleaning and analysis of real time physiological sensor data, which we developed for this study.

2 RELATED WORK

2.1 Modelling and Measuring Affect

Models of affect attempt to categorise and typify feelings and emotions. The most widely accepted models are categorical, dimensional and appraisal based approaches [34, 75, 207], with categorical and dimensional approaches the most commonly used for automatic analysis and prediction of affect [75, 77]. Categorical approaches assume that there is a small number of fundamental and universally experienced emotions, whereas dimensional models, such as the commonly used circumplex model [153], assume that basic emotions constitute a broader bipolar emotional continuum based on the two dimensions *Valence* (pleasant vs unpleasant) and *Arousal* (sleepy vs alert). The circumplex model has particular advantages over categorical models, such as relativising discrete emotions and representing their intensity [75, 153, 154].

In practice, affect is modelled and measured using several subjective, 'ground truth' measures that rely on users describing their emotional state or rating discrete emotions, valence and arousal through psychometric scales such as Experience Sampling [47], Pleasure-Arousal-Dominance scale [120], Self-Assessment Manikin [27] and Affect Slider [17]. However, there is a number of well established and understood disadvantages to these subjective measures. For instance, it is difficult to retrieve high resolution data of a participant's evolving emotional states, measures are influenced by participant openness and experimenter rapport, real time measurements are subject to the observer effect, measuring retrospectively is limited by participant recall, and reported emotions are influenced by social desirability [75, 80]. These limitations have motivated a wide body of research exploring automatic affect recognition through observing the physiological changes of the body in response to stimuli and correlating these with emotional states.

2.2 Physiological Affect Measures

In the affect recognition literature, physiological patterns have been shown to change in response to stimuli as a consequence of sympathetic nerve activation of the autonomic nervous system (ANS) and, in principle, are indicative of changes in the underlying affect of users [75, 77]. These changes in physiological signals are broadly categorised into *phasic* and *tonic* changes [9, 200]. Phasic activation refers to fluctuations of a signal in a time window occurring either spontaneously or in response to external stimuli, whereas tonic activation refers to a gradual shift in the overall baseline activity of a user [200]. The most common measures used are the cardiovascular system (heart rate and respiration), electrodermal activity (skin conductance and galvanic skin response), muscular activity (EMG), eye activity (pupillometry and eye tracking) and brain activity (EEG) [28, 89, 150, 169].

As Jerritta et al. [89] describe, high quality data is essential to affect recognition systems — ensuring that emotions are elicited 'naturally' and that interpersonal and environmental artefacts are removed. In the context of VR exergaming, certain physiological measures will be more or less appropriate in building an affect recognition model. To motivate our choices of physiological sensors, we describe how a measure has been utilised in the wider literature, what emotions a particular measure has been shown to correlate with, and what the challenges are for using these measures in the context of VR exergaming.

2.2.1 Pupillometry. Measuring changes in pupil diameter is a well established method for measuring activity in the brain and ANS response [152]. Modern eye trackers provide a robust and precise measure of pupil activity and provide additional eye metrics useful for affect recognition such as blinks, fixations and saccadic movements. For non-VR affect recognition, both Pupil Dilation Level (PDL – tonic) and Pupil Dilation Response (PDR – phasic) has been shown to correlate negatively with valence [2, 10, 29, 36, 92, 125, 133, 206] while positively correlating with arousal [29, 113, 146, 173, 194]. Pupil dilation has also been positively correlated with specific emotions that are typically low valence and high arousal, such as fear [37, 113, 173] and stress [132, 136].

However, in the context of VR exergaming, Barathi et al. [15] found conflicting results with pupil dilation weakly correlating with valence. This is an interesting finding that could be attributed to a genuine difference in pupillary affect response under high physical exertion or an artefact induced by the exergame virtual environment — dilation as a reflex to luminosity. Raiturkar et al. [146] proposed a method for decoupling the pupillary light reflex from emotional arousal in a desktop setup by sampling pixel luminance in the user's foveal region (a visual angle of 2°). A similar approach can be employed within VR for a cleaner measure of pupil dilation, providing a more robust predictor of affect as opposed to a predictor of the VE.

Blink metrics such as rate and duration have also been studied and correlated to affect. However, the literature is somewhat conflicted on how blink behaviour correlates with emotional response [15, 116, 172] and it appears to be highly dependent on the stimuli used [116]. Despite this, blink information has been used in multimodal ML approaches for predicting affect [3, 176], but it remains unclear what the exact relationship is between blink measures and emotional response and whether blinks are a significant predictor, especially in the context of VR exergames.

2.2.2 Heart Rate. Measuring the activity of the heart using electrocardiography (ECG) has been a common means of differentiating between positive and negative emotions [76, 97, 130, 167]. Specifically, heart rate, contractions of the heart per minute (BPM), is an indicator of emotional arousal [199], and heart rate variability (HRV), the oscillation between two consecutive heartbeats (interbeat or RR interval), is an indicator of ANS response [6, 166]. HRV, in

particular, has wide applicability in affect recognition [89, 100, 169] as well as affective gaming [150] and is broken down into *frequency domain* measures, the distribution of absolute or relative power into different frequency bands, and *time domain* measures, quantifying the amount of variability in measurements of the RR interval [166].

For affect recognition, time domain measures are typically employed, with related work looking at a variety of metrics including the standard deviation of normal to normal RR intervals (SDNN) [41, 73, 76, 97, 130, 167, 171], root mean square of successive RR interval differences (RMSSD) [41, 130], and percentage of successive RR intervals that differ by more than 50ms (PNN50) [41, 130]. In these studies, HRV has often been utilised in ML approaches to predict affect outside of VR and not during exercise to predict valence and arousal [41, 76, 130] and discrete emotions such as fear [73, 171], stress [87] and happiness [73, 171]. Beyond ML-based affect recognition, HRV has been shown to be positively correlated with valence [167] and negatively correlated with negative emotions such as fear [64, 135].

However, measuring heart rate in exergaming can be challenging due to noise in the signal induced by physical exertion, and it is generally accepted that HRV is a better measure of emotional response [191, 205]. For HRV in the context of exercise, it is established that every energy component of HRV decreases as exercise intensity increases [42]. Moreover, the variance of the RR interval significantly reduces during exercise compared to rest [140]. Shaffer et al. [166] also describe the limitations of short term (≥5 mins) and ultra-short term (<5 mins) measures of HRV compared to 24 hour measures. As a result, it is unclear whether HRV is a robust predictor of affect despite the limitations of measurement duration and the effects of exercise intensity in VR exergaming.

2.2.3 Electrodermal Activity (EDA). EDA describes the changes in the skin's ability to conduct electricity and can be used to understand the overall arousal of the sympathetic nervous system [9, 49, 63, 188]. Sometimes referred to as Galvanic-Skin Response (GSR), although this term is no longer recommended [9, 63], EDA is typically measured through electrodes on the surface of the skin on active areas of the body (e.g. the palm). The metrics acquired from EDA sensors typically used in affect recognition are Skin Conductance (SC), measured in micro-siemens (μS), and Skin Resistance (SR), measured in kiloohms (kohms) [63]. According to Babaei et al., most papers in the HCI literature that utilise EDA either use SC directly or transform their signal into SC (e.g. from SR) [9]. As with the previously discussed physiological signals, there are two types of measurements for Skin Conductance — *phasic* referred to as Skin Conductance Response (SCR), and *tonic* referred to as Skin Conductance Level (SCL).

These EDA measures are widely used in affect recognition, especially for detecting emotional arousal, in which SCL and SCR have both been shown to positively correlate with arousal [15, 29, 155]. Moreover, SC has been observed to correlate with specific emotions: for example, both SCL and SCR positively correlate with fear [65, 103, 203] and stress [22, 179], whereas SCL negatively correlates [210] and SCR positively correlates [95, 210] with happiness. SCR has also been found to parallel other physiological measures such as pupil dilation for both high and low valence stimuli [29]. Additionally, EDA is often incorporated into multimodal ML models for affect recognition [8, 72–74, 89, 97, 98, 105, 149].

However, measuring tonic and phasic EDA in VR exergaming, especially for high intensity exercise, poses significant challenges. For example, both exercise intensity and duration have a large impact on the amount of sweat the body produces and therefore can significantly influence both phasic and tonic EDA [21, 142, 143]. The exercise activity within a VR exergame can also induce motion artefacts and noise in the EDA signal, resulting in EDA becoming a less robust measure of affect. Another compounding factor is posed by the perceptible and sometimes imperceptible effects of using a VR HMD, such as motion sickness, which can influence SCR [69]. Despite these challenges, Barathi et al. [15] demonstrated EDA positively correlating with arousal in high intensity VR exergaming. Yet, it remains unclear how EDA correlates with valence and discrete emotions. With this in mind, it is also unclear how robust EDA is as an affect predictor across different exercise intensities and what data cleaning steps may be necessary to maintain predictive power.

2.2.4 Facial Tracking. Recent studies have also explored facial muscle activation (facial expressions) as measured by electromyography (fEMG) as an indication of emotional response [210]. Specifically, activation of the zygomaticus major, the muscle that controls smiling, is an indication of positive valence, and the corrugator supercilii, the muscle that controls frowning, is an indication of negative valence [33, 210]. Importantly, facial gestures and fEMG response follow the same tonic and phasic activation as previously discussed physiological metrics [70, 94]; however, both zygomatic and corrugator activity can exhibit more or less phasic modulation depending on the stimuli [70].

In non-VR and non-exergaming contexts, fEMG has been explored widely for affect recognition, with zygomatic activation positively correlating with valence [109, 159, 182] and arousal in the presence of high valence [148, 209]. As with the previously discussed physiological measures, fEMG has been incorporated into multimodal ML approaches for predicting affect [86, 171, 181] and has even been used to predict discrete emotions such as fear, happiness and sadness [171].

In VR exergaming, there are practical challenges to incorporating facial tracking. A VR HMD typically obscures a user's face, especially the corrugator supercilii muscle, and may also inhibit muscle activation. Additionally, fEMG electrodes may be subject to mechanical interference and electrical noise from the HMD [189]. However, commercially available face and lip trackers designed for VR HMDs [1] [2] provide blend shapes and gesture estimations of part of a user's face and, importantly, the zygomatic major muscles. While promising, these tracking techniques are primarily designed for conventional VR experiences and it is unclear how different levels of physical movement and exertion will influence the predictive power of facial tracking in affect recognition.

2.2.5 Other Measures. Other physiological signals have been used in affect recognition, such as brain activity [25, 88, 180], the respiratory system and skin temperature [89, 169]. However, for high intensity exercise and VR exergaming we have chosen to exclude

[1] https://www.vive.com/uk/accessory/facial-tracker/
[2] https://business.vive.com/us/product/vive-focus-3-facial-tracker/

these measures. For brain activity measures such as EEG, motion artefacts and electrical interference pose a significant challenge when used for affect recognition, especially when used in the context of exercise [55] and VR [184]. For both respiration and skin temperature measures, the influence of exercise and environmental factors [67] also make it challenging to decipher affective response, especially in the context of high intensity exergaming.

3 AFFECTIVE VIRTUAL ENVIRONMENT DESIGN

To address RQ1 we designed four distinct VR exergame Virtual Environments (VEs), each designed to target different quadrants of Russell's circumplex model [153, 154]. We refer to these different VEs by the emotions they target — **Happy**, **Calm**, **Stress** and **Sadness**. The exergame was designed in the Unity engine[3] and allows users to cycle through the different emotion VEs using an exercise bike while they gain points by collecting coins, avoiding obstacles or simply pedalling. Each VE simulates a different virtual bike ride and game experience, with the specific design choices for each guided by existing literature on emotion elicitation and stimuli [77, 97, 149, 161, 177, 208], as well as affective game design [32, 150] and gamification theory [99, 165]. All four VEs vary by game mechanics, terrain, environmental objects, lighting and colour scheme, and sound design. We composed music soundtracks based on research by Fernández-Sotos et al. [60] and Liu et al. [114] which mapped music tempo and note length to the circumplex model. We used a tempo of 150 beats per minute (bpm) and sixteenth notes for high arousal emotions, and 90 bpm and whole/half notes for low arousal emotions. The work of Ng and Nesbitt [131] informed the design of sound effects and audio feedback within each emotion VE.

In order to achieve a robust dataset for correlating physiological response to affect [89], the exergame VEs should be validated across different exercise intensities to ensure that: (i) the VEs elicit the correct target emotions (e.g. feeling stressed in the stress VE), and (ii) the dominant emotion in each VE aligns with the targeted emotion (e.g. within the stress VE, stress is elicited significantly more than dissimilar emotions). Through extensive pilot testing, described in subsection 4.5, we were able to validate the efficacy of the environments before conducting the main study. We then further validated the virtual environments as part of the main study, as described in section 5.

3.1 Negative Valence Virtual Environments

Geslin et al. [66] describe how environmental colour schemes, lighting and game objects can elicit emotion. To target negative valence (stress/sadness), the negative VEs contained the following features: desaturated colours, darkness, and dirt. For the VR environment colour schemes, we primarily manipulated the skybox with the specific choice of colour informed by Dharmapriya et al. [50] who mapped Itten's colour system [40] to Russell's circumplex model of affect [153]. In this case, we used a gradient of pink for the **Stress** VE and blue for the **Sadness** VE. These colour choices were also supported by research on how colours can be used in constructing emotions by interactive digital narratives [185]. The feature of dirt [66] was also incorporated into both the stress and sadness VEs. A mostly mud landscape was used for the **Sadness** VE with a sparse distribution of dead grass and obstructing dirty road objects. The **Stress** VE landscape was also designed to look barren but incorporated more claustrophobic and stress inducing elements such as surrounding steep rocky cliffs, burnt trees, boulders, fire, pressuring text (e.g. "hurry!", "collect the coins quickly!"), and a timer. Additional features were added to both negative valence VEs that were incorporated in previous VR exergames [13, 15] including a chasing police car and barking dogs in the **Stress** VE and heavy rain in the **Sadness** VE.

In both the **Stress** and **Sadness** VEs, we incorporated a coin collecting game mechanic in which users can accrue points by leaning left or right and intersecting their heads with a coin. For each coin collected, the user typically gains a point and hears a positive reward sound effect [131]. However, in the **Stress** VE skull coins were added to introduce the feature of 'loss' [66] and 'consequence' [99]. When collected, these coins deduct ten points and a harsh buzzer sound effect is emitted [131]. This mechanic parallels the VR exergame by Barathi et al. [13, 15] where points are deducted when colliding with traffic. In the **Sadness** VE, 'loss' was implemented differently, whereby instead of deducting points the number of coins available was greatly reduced. This was intended to create a larger feeling of 'loss' [66] with sparse 'rewards' [99, 165] in comparison to the other VEs. The coins in the **Sadness** VE also had a rusted appearance and produced a less satisfying sound effect when collected to further decrease the positive feedback compared to other VEs [32, 99, 131]. The soundtrack for the **Stress** VE included discordant high pitched notes, while the **Sadness** VE contained distant melancholic sound effects [60, 114].

3.2 Positive Valence Virtual Environments

Informed by the same literature as the negative valence VEs [50, 66, 185], the **Happy** and **Calm** VEs' skybox colour schemes were orange and turquoise, respectively. Wildflowers were added to their landscapes and, using directional light, shadows passing over them gave the effect of natural light [66]. Heads up display text played a different role in the positive valence VEs to feature interaction and positively reinforcing feedback in the environment [32, 66, 99, 165]. For the **Happy** VE, the text included motivational and positive messages, whereas the **Calm** VE included guided breathing exercise instructions and meditative messages such as 'Calm your mind'.

For the **Happy** VE, coins were far more abundant and heart shaped gems appeared that offered 10 bonus points, vastly increasing the 'reward schedule' [99] and increasing 'earnings' [66]. Additionally, a wide variety of animated game objects with sound effects were included such as rabbits running through the fields, colourful birds singing and hot air balloons [131]. The abundance of environmental objects encouraged users to look outwards and upwards to the expansive landscape and skybox, contributing to the features of wide shots and open spaces [66]. For the **Calm** VE, coin collection was replaced with the instruction to 'Gently pedal for points'. This allowed the user to be rewarded with one point for every 2 seconds of cycling. A fundamental concept of calmness and serenity is minimising distractions [46], and users are encouraged

[3]https://unity.com/products/unity-engine

to focus solely on experiencing the environment while still maintaining a positive 'reward schedule' [66, 99]. The soundtrack for the **Happy** VE included upbeat vocal and chanting sound effects, whereas the **Calm** VE soundtrack maintained a consistent rhythm and was purely instrumental [60, 114].

3.3 Neutral Virtual Environment

This VE aimed to elicit no particular emotion and was used as a transition between each of the emotion VEs in order to reduce any carryover effects. To achieve this, the skybox was set to white as this shade is at the centre of Ittin's colour system as mapped by Geslin et al. [66]. The landscape had a flat elevation profile, plain green terrain and no objects of interest (see Figure 2).

4 METHODOLOGY

To address our research questions, we designed an experiment using the aforementioned exergame, in which users cycled through the four VEs at different exercise intensities while physiological data was recorded. The experiment followed a within-participants design with exercise intensity (three levels) and Emotion VE (four levels) as independent variables. Each 90-minute experimental session included one low, one medium, and one high-intensity exercise bout in which the four emotion VEs each were experienced for 60 seconds each ($3 \times 4 = 12$ Emotion VE exposures).

Each exercise intensity was scaled as a percentage Heart Rate Reserve (HRR), the difference between a participant's age-predicted maximal heart rate (HRMAX) [168] and resting heart rate, which is often used in calculating exercise training capacity. Low exercise intensity was defined as 50-60% of participant HRR, medium as 60-70%, and high as 70-80%. These HRR ranges are typical for each exercise intensity and were also validated through pilot testing. The orders of exercise intensity and Emotion VE were counterbalanced using a balanced Latin square design.

4.1 Apparatus

The VR exergame developed for this study, described in Section 3, required participants to cycle on a stationary Wahoo KICKR exercise bike while wearing a Vive Pro Eye VR headset. Physiological measures were collected using the eye tracker in the VR headset (pupillometry), a Shimmer3 GSR+ tethered to a participant's middle and ring finger (EDA) [48], a Polar H10 HR monitor chest strap (HR and HRV) [162], and a Vive face tracker (facial tracking). All physiological measures were sent to a PC (Intel 13900K, Nvidia GTX 4090 and 64GB of DDR5 RAM) running the Unity VR exergame over Bluetooth (BLE protocol), which recorded all measures at a sample rate of 40-50 Hz using the `EmoSense` SDK [144].

4.2 Measures

4.2.1 Ground Truth Measures of Affect. We collected affect ground truth ratings using a combination of experience sampling (ESM) [47] and Pleasure-Arousal (PAD) sampling [120] administered within the VEs. To measure discrete categorical emotions, we used 11-point rating scales (0-10) for Fear, Excitement, Stress, Happiness, Sadness, Calmness, Boredom and Contentedness (0 being the least amount of that emotion possible and 10 being the most). To measure valence and arousal, we used the Affective Slider [17], a validated scale that builds on the Self-Assessment Manikin [27]. The Affective Slider questions were also administered on an 11-point rating scale (0-10). Using 11-point rating scales is a validated approach for collecting data that can be analysed on an interval scale; this is supported by both theory and simulation [79, 204] and has previously been validated for measuring affect in an exercise context [78, 187]. Such interval-scaled data can be analysed with parametric statistical techniques such as repeated measures ANOVA and linear regression, given all their assumptions are sufficiently met [4, 102].

Russel's circumplex model describes the linear relationships between categorical emotions and its two dimensions valence and arousal [153, 154], e.g. Stress indicates low valence and high arousal. This means multi-item measures of valence and arousal can be derived by weighting the points corresponding to each categorical emotions in the circumplex model by their respective emotion rating, yielding weighted averages based on all eight categorical emotion ratings. Such multi-item measures outperform single items with regard to predictive validity [51, 156], therefore we use them as our primary measures of Valence and Arousal. During our analyses, we confirmed that the multi-item measures were significantly correlated with the single items of the Affect Slider but were more robust in avoiding assumption violations and predicting affect from physiological measures. More details about the multi-item measures can be found in the Supplementary Material.

4.2.2 Physiological Sensor Measures. We collected *phasic* and *tonic* physiological metrics known to be associated with affect. For pupillometry, we measured pupil dilation level (PDL) as the mean pupil size in millimetres (mm) and the dilation response (PDR) as the standard deviation of the pupil size. The standard deviation has previously been found a useful measure for quantifying series of phasic dilation responses during prolonged and continuous exposure to stimuli [5, 164], as is the case in our exergame. Blink rate (BR) was measured as blinks per minute, and blink duration (BD) in milliseconds (ms). For EDA we measured Skin Conductance Level (SCL) and Skin Conductance Response (SCR) in microsiemens (μS). For SCR we specifically calculate the 'EDA positive change' [112], an approach for measuring SCR during continuous exposure to stimuli such as our exergaming VEs. For facial tracking, we measured the movement of the zygomaticus major (Smile) using the Vive facial tracker blend shape weightings for *Mouth_Smile* [193]. For HR and HRV, we measured beats per minute (BPM) and inter-beat (RR) intervals (ms) which were then used to compute SDNN and RMSSD [166]. The power output (watts) of cycling was measured through the exercise bike.

4.2.3 Other Measures. We recorded participants' overall enjoyment of the VR exergame after each exercise bout at a given intensity using the Intrinsic Motivation Inventory (IMI) [13, 119], a 7-point Likert scale (1="not at all true", 4="somewhat true", 7="very true") that measures different aspects of intrinsic motivation. Specifically, we considered the *Interest/Enjoyment* (7 items), *Pressure/Tension* (5 items), and *Perceived Competence* (6 items) subscales, which are well-validated for use in an exercise context and have been used in prior VR exergaming studies [13, 15].

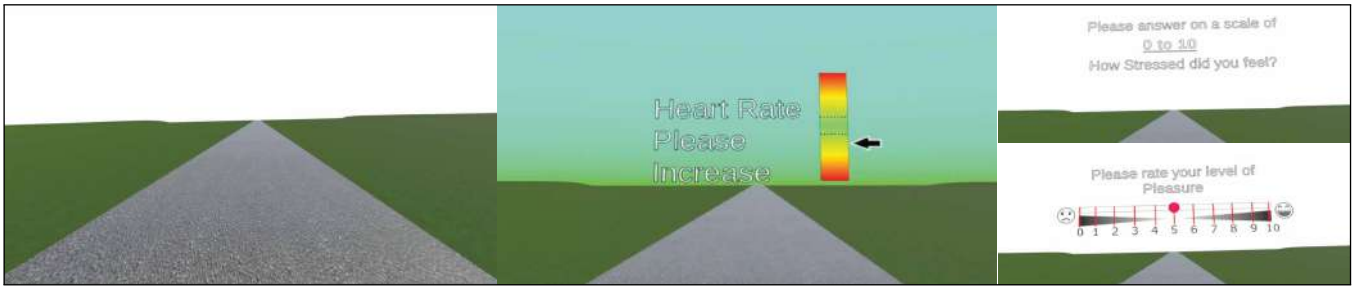

Figure 2: Left: The Neutral virtual environment participants cycled in during warm up and cool down. Middle: The real time HR visualiser to guide participants into the right HR range for a given exercise intensity. Right: The ground truth measures administered to participants in VR using experience sampling and Affective Slider.

4.3 Data Cleaning

We considered three levels of data cleaning, **raw**, **env**, and **pers**, which build on one another and increasingly remove the influence of factors unrelated to affect.

*4.3.1 No Cleaning (**Raw**).* This level does not change the measures as provided by the sensors, i.e. does not remove effects unrelated to affect. It serves as a baseline to compare the next levels against.

*4.3.2 Environmental Cleaning (**Env**).* This level aims to removes outliers, e.g. caused by body movements or the HMD, by removing values that were clearly erroneous or fell far outside the typical ranges reported in the literature. For pupil dilation, values were filtered out when the pupils were not tracked such as during blinks. For calculating blink duration and blink rate, we defined a successful blink as both eyes being closed for ≥50 ms and <700 ms [58, 106]. For EDA we removed negative skin conductance values. For HR, values that had an RR-interval of less than 200ms and more than 2000ms were removed [91, 93, 166]. For HRV we additionally applied the age based filtering algorithm for RR-intervals proposed by Karlsson et al. [91], which is uses recursive filtering to remove changes in RR-intervals that are unlikely given a participant's age [201, 202]. For facial tracking, blank cells where tracking and pose estimation were lost as determined by the Vive Sranipal SDK [193] were removed.

Furthermore, we removed artefacts induced by the VEs and exercise, which is important to avoid a model overfitting VE stimuli or exercise-related effects rather than predicting a user's affective response. As environmental luminosity can interfere with pupil measures, we applied the approach proposed by Raiturkar et al. [146] to remove the influence of the pupillary light reflex triggered by light changes in the VEs. This involved taking baseline measures of each pupil at 16 different luminosity levels with no emotional stimuli present in VR (pilot testing showed the eight luminosity levels recommended by Raiturkar et al. did not provide enough granularity when applied in VR). Pupil dilation values were then corrected in real time based on the observed foveal luminosity (2° visual angle of the VE at the point of gaze estimated by the HMD eye gaze tracker) by subtracting the baseline. For EDA, we removed the influence of pre-existing sweat by calculating the log-transformed ratio of the current skin conductance and the baseline PDL measured immediately before each VE exposure [16, 30, 192].

*4.3.3 Personalised Cleaning (**Pers**).* Here we combined the prior cleaning methods with methods to remove interpersonal differences for each physiological measure and rating of affect. EDA is influenced by individual differences in eccrine activity [157, 158], PD and BR by different pupillary sensitivities [81], HR/HRV by different parasympathetic and sympathetic stimulation of the heart [38, 101], and Smile by differences in zygomaticus major activity [163]. We account for a participant's natural baseline and spread in physiological measures by applying z-score transforms, subtracting the participant's mean and dividing by their standard deviation [196] as estimated from the participant's collected data, which can improve the predictive power of measures [7, 15, 16]. Similarly, we also apply z-score transforms to all affect rating measures to correct for personal response biases [62, 121].

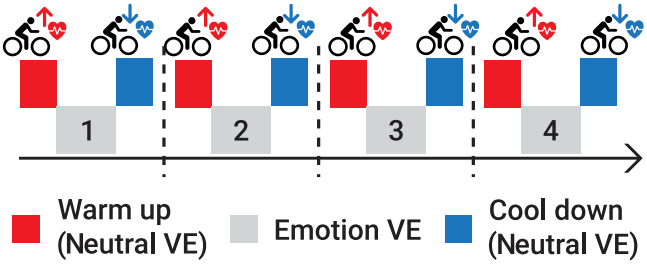

Figure 3: Overview of the procedure for one exercise bout at a given exercise intensity. Red cells denote a warm up phases and blue cells denote cool down phases, both in the Neutral VE. Grey squares denote an exposure phase of 60-seconds in an Emotion VE.

4.4 Procedure

Participants were screened with the Physical-Activity Readiness Questionnaire (PARQ) [186] and a custom VR screening questionnaire, which excluded participants who were susceptible to health risks of high-intensity exercise and using VR technology. We provide both screening questionnaires in the Supplementary Material. The remaining participants gave informed consent and completed a demographics questionnaire. Participants were then familiarised with the exercise bike, setting a comfortable initial pedal resistance and position. The experimenter then fitted the physiological sensors, adjusted the Inter-Pupillary Distance (IPD) of the HMD, calibrated the eye gaze tracker using a standard 5 point calibration, recorded baseline pupil diameter measures under different levels of luminosity [146], and performed a basic eye test to ensure participants could read any text in the VEs.

When ready, participants started the exergaming exposure protocol illustrated in Figure 3. During each warm up phase in the **Neutral** VE, a visualisation of actual and target HR to enable participants to reach the desired level of exercise intensity (see Figure 2). Participants were free to increase or decrease the bike resistance, staying in the warm up phase until they maintained the target HR for 10 seconds. Participants were then exposed to one of the **Emotion** VEs for 60 seconds. During pilot testing we found 60 seconds sufficient to induce emotions and obtain meaningful physiological measures while avoiding confounding factors such as physical exhaustion, especially for the high-intensity condition. Prior work has shown such short exposures are sufficient to elicit emotions and measure affect [15, 52].

Afterwards participants were transitioned back to the **Neutral** VE for a cool down phase, in which they answered the affect rating questions verbally (see Figure 2). Once ready, participants transitioned once again to the warm up phase for another **Emotion** VE. Warm up, exposure, and cool down were repeated four times in each bout of exercise. After completing an exercise bout for a given intensity, participants exited VR and completed the IMI questionnaire. Participants were able to take a break during this period before re-entering VR and beginning the next exercise bout.

4.5 Pilot Study

To validate the VEs and methodology for the main study, and inform our hypotheses, we conducted a pilot study with 29 participants (16 male, 13 female, age 19-33 $M = 25, SD = 3$). The methodology was very similar to that of the main study, but limited in the number of physiological sensors (SCL, PDL, HR, BD, and BR), metrics, and cleaning approaches used. The detailed results and analysis R scripts can be found in Supplementary Material.

4.6 Hypotheses

For manipulating affect (RQ1), we have the following four families of *a-priori* hypotheses for each of the four VEs. They are based on the differences in valence and arousal between the four quadrants of Russell's circumplex model [153, 154] that are targeted by our four VEs, and are also corroborated by the pilot data. We distinguish *comparisons between VEs* (H1-H3), which compare the effects different VEs have on the same affect measure, and *comparisons within VEs* (H4), which compare the effects a single VE has on different affect measures:

H1: Each VE elicits more of its target emotion than other VEs (e.g. the Happy VE elicits more Happiness than the other VEs).

H2: The high valence VEs elicit higher valence than the low valence VEs (i.e. the Happy and Calm VEs elicit higher valence than the Stress and Sad VEs).

H3: The high arousal VEs elicit higher arousal than the low arousal VEs (i.e. the Happy and Stress VEs elicit higher arousal than the Calm and Sad VEs).

H4: Each VE elicits more of its target emotion than of the emotions targeted by the other VEs (e.g. the Happy VE elicits more Happiness than eliciting Calmness, Stress, and Sadness).

Table 1: Hypotheses about the physiological measures that predict specific affect ratings during VR exergaming (highlighted cells). Each cell summarises the evidence for a positive (+), negative (-), or unclear (?) relationship.

Affect	PDL / PDR	SCL / SCR	HRV	Smile	Power
Valence	**+ :** [15] **- :** Pilot, [2, 29] [36, 92, 125] [133, 206] **? :** [104, 134]		**+ :** [167] **? :** [76, 130]	**+ :** [109, 159] [182] **- :** [187, 190] **? :** [86]	**+ :** [15] **- :** [53, 137] [18, 187]
Arousal	**+ :** Pilot, [113] [146, 173, 194]	**+ :** [15, 29] [155] **? :** [8]			**+ :** [90, 107] [183]
Fear	**+ :** Pilot, [37] [173]				**+ :** [18, 53] [137, 183]
Stress	**+ :** Pilot, [132] [136] **? :** [12]	**+ :** [22, 179] **? :** Pilot, [136]			**+ :** [18, 53] [137, 183]
Happiness	**- :** Pilot	**- :** Pilot [210]		**+ :** [54, 210] **? :** [171]	
Sadness	**+ :** Pilot, [37]				
Boredom	**+ :** Pilot **- :** [195]				
Contentedness	**- :** Pilot			**+ :** [109, 159] [182] **- :** [187, 190] **? :** [86]	
Calmness	**- :** Pilot				**+ :** [90, 107] [183]

Details about all the hypotheses in each family are provided in the Extended Analysis Report in Supplementary Material. For predicting affect in VR exergaming (RQ2), Table 1 provides an overview of affect rating measures together with their hypothesised physiological predictors (in grey). The hypotheses are undirected and are based on the pilot study results and affect recognition literature. Each table cell summarises evidence for a positive (+) and negative (-) relationship, as well as citing works where the relationship is unclear (?). Excitement has no hypothesised physiological predictors due to a lack of clear results in the pilot study and wider literature and, as a result, is excluded from the table. For the same reason, we excluded columns for blink rate, blink duration, and heart rate.

4.7 Participants

We recruited 72 participants ($Male = 43, Female = 27, Non\text{-}Binary = 1, Other = 1$) who were predominantly staff and students of the University of Bath. Participants were aged 18-60 ($M = 32.542, SD = 11.334$) and, according to the results of the International Physical Activity Questionnaire (IPAQ) [43, 111], most participants had high physical activity ($High = 42, Moderate = 28, Low = 2$). Most participants had used VR occasionally ($Occasionally = 49, Never = 20, Weekly = 2, Daily = 1$) and had played video games occasionally ($Occasionally = 44, Daily = 15, Weekly = 10, Never = 3$). A total necessary sample size of 72 participants was

calculated using G*Power 3.1.9.7 [59] analysis for multi-level linear regression with 3 predictors; which would be able to detect small effects (Effect Size: 0.15, Power: 0.85, alpha: 0.05).

5 RESULTS

This section provides an overview of the analysis strategy and summarises the main study results for each research question. All analyses were carried out using R 3.1 and JASP 0.18.1. The detailed R scripts and JASP files used to perform the analyses and create the results can be found in the Extended Analysis Report in Supplementary Material.

5.1 RQ1: Affect Manipulation

Figure 4 shows boxplots of the affect ratings of the four VEs across all participants, providing an overview of how different types of emotions were elicited in each environment.

5.1.1 Comparisons Between VEs (H1-H3). We tested the normality assumption of Analysis of Variance (ANOVA) using Shapiro-Wilk tests and by inspecting QQ-plots, and decided to use non-parametric test alternatives to address any concerns about violations of normality. We tested the overall effects of our four VEs on valence, arousal, and the four target emotions using Friedman tests, followed by pairwise Wilcoxon signed-rank tests with Holm–Bonferroni correction using the `coin` R package [83]. The main effects of the VEs on valence, arousal, and the four target emotions were all significant ($\chi^2 \geq 89.891$, $W \geq 0.416$, $p < .001^{***}$).

Table 2 summarises the results of all pairwise comparisons based on the median affect ratings across the three exercise intensity levels. The results hypothesised by H1-H3, which are highlighted in yellow and blue, are all highly significant ($p < .001^{***}$). Our more detailed results in the Extended Analysis Report in Supplementary Material test each exercise intensity level separately and confirm these results except that the Happy VE did not elicit significantly more Happiness than the Calm VE during high-intensity exercise. This is likely due to the similar and non-exclusive nature of happiness and calmness. Overall, our results support H1-H3, indicating that the VEs elicit the emotions they are targeting and achieve the right levels of valence and arousal in relation to one another. For more details on the test methodology and results, please refer to the Extended Analysis Report in Supplementary Material.

5.1.2 Comparisons Within VEs (H4). In order to make measures for different emotions comparable (e.g. happiness and stress), personalised cleaning with z-score transforms was applied to each emotion measure [1, 196]. This removes response biases, which affect the different emotion measures differently and hence hamper comparisons between them [62, 121]. We tested the normality assumption of ANOVA using Shapiro-Wilk tests and by inspecting QQ-plots, and decided to use non-parametric test alternatives to address any concerns about violations of normality. We tested the overall effects of the type of target emotion (Happy, Stress, Calm, Sad) on the four target emotion measures using Friedman tests, followed by pairwise Wilcoxon signed-rank tests with Holm–Bonferroni correction between the different target emotion measures using the `coin` R package [83].

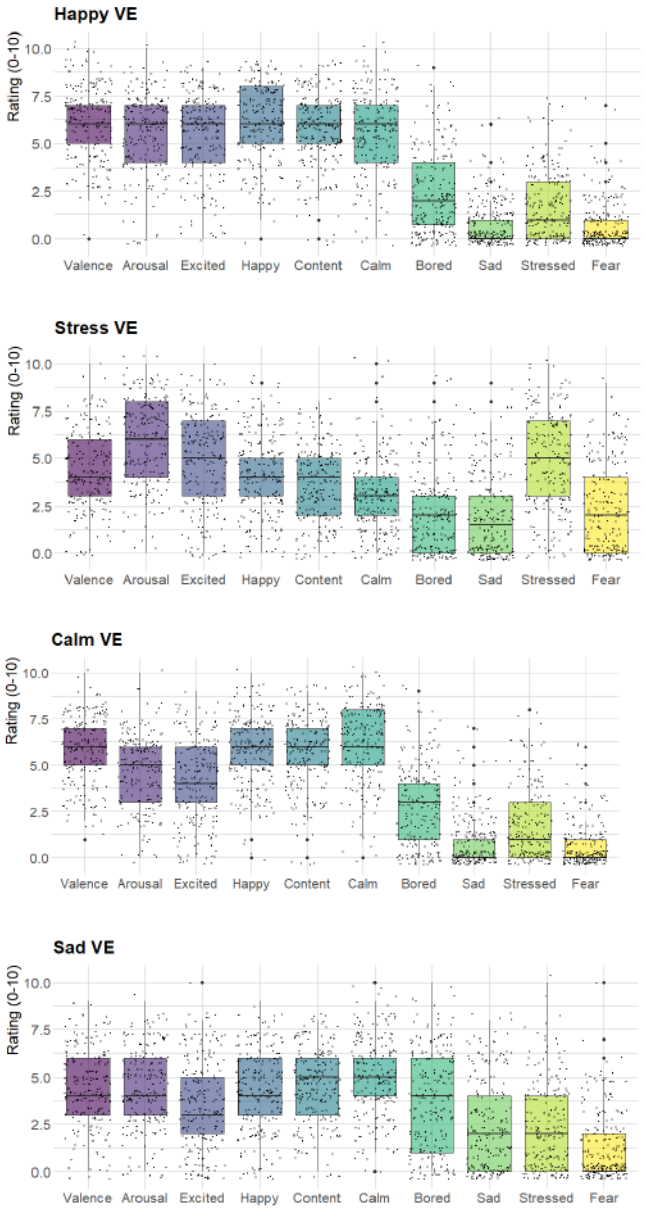

Figure 4: Boxplots showing all 10 affect ratings from all participants across the four VEs: Happy, Stress, Calm, & Sad.

The main effects of the type of target emotion on the target emotion measures were all significant ($\chi^2 \geq 78.632$, $W \geq 0.364$, $p < .001^{***}$). Table 3 summarises the results of all pairwise comparisons based on the median affect ratings across the three exercise intensity levels. The non-empty table cells show the results for all the hypotheses of family H4; all comparisons are highly significant ($p < .001^{***}$). Our more detailed results in the Extended Analysis Report in Supplementary Material test each exercise intensity level separately and confirm these results except that the Calm VE did not elicit significantly more Calmness than Happiness during high-intensity exercise. This is likely due to the similar and non-exclusive nature of happiness and calmness. Overall, our results

Table 2: RQ1 results: Comparisons between VEs of the elicited valence, arousal, and target emotions (H1-H3) across the three levels of exercise intensity based on median affect ratings. Each row reports the comparisons between the two VEs *a* and *b* listed in the first two columns. The cells with the hypothesised results are highlighted either in yellow meaning the hypothesis is "VE *a* elicits more than VE *b*", or in blue meaning "VE *b* elicits more than VE *a*". Cells show the means $\bar{a}$ and $\bar{b}$ and standard deviations σ for the two VEs, and non-parametric Wilcoxon signed-rank test results with effect size r ($r < 0.3$ for 'small', $0.3 \geq r < 0.5$ for 'moderate', and $r \geq 0.5$ for 'large'). For separate Wilcoxon signed-rank tests for each exercise intensity level please refer to the Extended Analysis Report in Supplementary Material.

VE Comparison		Valence	Arousal	Happy	Stress	Sad	Calm
a) Happy VE	*b*) Stress VE	$\bar{a}$= 0.438, σ= 0.133 $\bar{b}$= 0.153, σ= 0.199 Z = 12.694, r= .864 p<.001***	$\bar{a}$= -0.072, σ= 0.096 $\bar{b}$= 0.052, σ= 0.157 Z = -10.609, r= -.722 p<.001***	$\bar{a}$= 6.083, σ= 1.734 $\bar{b}$= 4.125, σ= 1.919 Z = 11.692 r= .796 p<.001***	$\bar{a}$= 1.500, σ= 1.427 $\bar{b}$= 4.681, σ= 2.320 Z = -12.304, r= -.837 p<.001***	$\bar{a}$= 0.681, σ= 0.986 $\bar{b}$= 1.931, σ= 2.041 Z = -10.174, r= -.692 p<.001***	$\bar{a}$= 5.639, σ= 1.809 $\bar{b}$= 3.028, σ= 1.830 Z = 12.359, r= .841 p<.001***
a) Happy VE	*b*) Sad VE	$\bar{a}$= 0.438, σ= 0.133 $\bar{b}$= 0.236, σ= 0.220 Z = 11.157, r= .759 p<.001***	$\bar{a}$= -0.072, σ= 0.096 $\bar{b}$= -0.147, σ= 0.149 Z = 9.807, r= .667 p<.001***	$\bar{a}$= 6.083, σ= 1.734 $\bar{b}$= 4.139, σ= 1.922 Z = 11.402, r= .776 p<.001***	$\bar{a}$= 1.500, σ= 1.427 $\bar{b}$= 2.264, σ= 2.133 Z = -6.353, r= -.432 p<.001***	$\bar{a}$= 0.681, σ= 0.986 $\bar{b}$= 2.417, σ= 2.271 Z = -10.821, r= -.736 p<.001***	$\bar{a}$= 5.639, σ= 1.809 $\bar{b}$= 5.125, σ= 1.653 Z = 5.302, r= .361 p<.001***
a) Happy VE	*b*) Calm VE	$\bar{a}$= 0.438, σ= 0.133 $\bar{b}$= 0.422, σ= 0.134 Z = 1.269, r= .086 p=.204	$\bar{a}$= -0.072, σ= 0.096 $\bar{b}$= -0.148, σ= 0.101 Z = 10.404, r= .708 p<.001***	$\bar{a}$= 6.083, σ= 1.734 $\bar{b}$= 5.722, σ= 1.630 Z = 5.0467, r= .343 p<.001***	$\bar{a}$= 1.500, σ= 1.427 $\bar{b}$= 1.444, σ= 1.619 Z = 2.570, r= .175 p=.010*	$\bar{a}$= 0.681, σ= 0.986 $\bar{b}$= 0.833, σ= 1.056 Z = -3.187, r= -.217 p=.001**	$\bar{a}$= 5.639, σ= 1.809 $\bar{b}$= 6.292, σ= 1.732 Z = -5.145, r= -.350 p<.001**
a) Calm VE	*b*) Stress VE	$\bar{a}$= 0.422, σ= 0.134 $\bar{b}$= 0.153, σ= 0.199 Z = 12.671, r= .862 p<.001***	$\bar{a}$= -0.148, σ= 0.101 $\bar{b}$= 0.052, σ= 0.157 Z = -12.022, r= -.818 p<.001***	$\bar{a}$= 5.722, σ= 1.630 $\bar{b}$= 4.125, σ= 1.919 Z= 10.049, r= .684 p<.001***	$\bar{a}$= 1.444, σ= 1.619 $\bar{b}$= 4.681, σ= 2.320 Z = -12.449, r= -.847 p<.001***	$\bar{a}$= 0.833, σ= 1.057 $\bar{b}$= 1.931, σ= 2.041 Z = -8.695, r= -.592 p<.001***	$\bar{a}$= 6.292, σ= 1.732 $\bar{b}$= 3.028, σ= 1.830 Z = 12.716, r= .866 p<.001***
a) Calm VE	*b*) Sad VE	$\bar{a}$= 0.422, σ= 0.134 $\bar{b}$= 0.236, σ= 0.220 Z = 11.428, r= .778 p<.001***	$\bar{a}$= -0.148, σ= 0.101 $\bar{b}$= -0.165, σ= 0.134 Z = 1.148, r=.078 p=.251	$\bar{a}$= 5.722, σ= 1.630 $\bar{b}$= 4.185, σ= 2.098 Z = 10.298, r= .701 p<.001***	$\bar{a}$= 1.444, σ= 1.619 $\bar{b}$= 2.394, σ= 2.224 Z = -5.997, r= -.408 p<.001***	$\bar{a}$= 0.833, σ= 1.057 $\bar{b}$= 2.417, σ= 2.271 Z = -10.093, r= -.687 p<.001***	$\bar{a}$= 6.292, σ= 1.732 $\bar{b}$= 5.125, σ= 1.653 Z = 9.231, r= .628 p<.001***
a) Stress VE	*b*) Sad VE	$\bar{a}$= 0.153, σ= 0.199 $\bar{b}$= 0.236, σ= 0.220 Z= -6.234, r= -.424 p<.001***	$\bar{a}$= 0.052, σ= 0.157 $\bar{b}$= -0.165, σ= 0.134 Z = 12.541, r= .853 p<.001***	$\bar{a}$= 4.125, σ= 1.919 $\bar{b}$= 4.139, σ= 1.922 Z = -0.164, r= -.011 p=.870	$\bar{a}$= 4.681, σ= 2.320 $\bar{b}$= 2.264, σ= 2.133 Z = 11.673, r= .794 p<.001***	$\bar{a}$= 1.931, σ= 2.041 $\bar{b}$= 2.417, σ= 2.271 Z = -4.343, r= -.296 p<.001***	$\bar{a}$= 3.028, σ= 1.830 $\bar{b}$= 5.125, σ= 1.653 Z = -11.729, r= -.798 p<.001***

support H4 that each VE elicits more of its target emotion than of the emotions targeted by the other VEs. For more details on the test methodology and results, please refer to the Extended Analysis Report in Supplementary Material.

5.2 RQ2: Affect Recognition

We used multi-level linear regression models from the R nlme package [20] to test the hypothesised physiological predictors (Table 1) for each affect variable, because of the power of such models when comparing repeated measures data [39, 138]. We confirmed the assumptions for linear regression by inspecting residual plots [4, 102]. In our regression tables (Table 4 and Table 5), coefficients are marked with $\not{L}$ if they violate linearity and marked with $\not{H}$ if they violate heteroskedasticity. Coefficients of determination R^2 are marked with $\not{N}$ if the normality of residuals is violated. Note that violations do not render a model useless – such a model can still have a high R^2 and be useful in practice. However, violations render the p-values of coefficients inaccurate, so they need to be considered with care. We also mark coefficients with unbalanced residual plots with U. While this does not violate the assumptions of linear regression, it indicates that the model can likely be improved by transforming the data, e.g. by applying a further level of cleaning.

When validating a regression model, we first included only those physiological sensor measures in the model that were hypothesised to predict an affect variable. The regression coefficients, which represent the linear effects of sensor measures on the affect variable, were tested with two-tailed tests at $\alpha = .05$. We then used Chow tests to detect discontinuities between different levels of exercise intensity [110], e.g. to detect whether the regression coefficients changed sufficiently between low-intensity and medium-intensity exercise to warrant separate regression models for them. That is, we merge the regression results of two exercise levels only if their coefficients are sufficiently similar. We conducted two such Chow tests, to determine whether low and/or high exercise intensity should be modeled separately. We adjusted the p-values of the two Chow tests with Holm-Bonferroni correction. As a result, our analysis yielded up to three regression models, to describe predictions at different exercise intensities. Finally, we added the sensor measures that were not hypothesised to predict an affect variable to the regression models and tested them with two-tailed tests at $\alpha = .05$, adjusting their p-values with Holm-Bonferroni correction. We repeated the whole procedure for each of the three levels of data cleaning, using raw, cleaned, or personalised sensor measures and affect variables, respectively. All regression models are presented

Table 3: RQ1 results: Comparisons of emotions elicited within each VE (H4) across the three levels of exercise intensity based on median affect ratings. The first column lists the VE and the following columns compare measures of the VE's target emotion *a* with measures of a non-targeted emotion *b*, respectively. Cells show the means $\bar{a}$ and $\bar{b}$ and standard deviations σ for the two emotions, and non-parametric Wilcoxon signed-rank test results with effect size r ($r < 0.3$ for 'small', $0.3 \geq r < 0.5$ for 'moderate', and $r \geq 0.5$ for 'large'). For separate Wilcoxon signed-rank tests for each exercise intensity level please refer to the Extended Analysis Report in Supplementary Material.

	a) Target Emotion Vs. *b*) Happy Rating	*a*) Target Emotion Vs. *b*) Stress Rating	*a*) Target Emotion Vs. *b*) Calm Rating	*a*) Target Emotion Vs. *b*) Sad Rating
Happy VE	—	$\bar{a}$= 0.638, σ= 0.447 $\bar{b}$= -0.492, σ= 0.500 Z = 11.712, r= .797 p<.001***	$\bar{a}$= 0.638, σ= 0.447 $\bar{b}$= 0.382, σ= 0.550 Z = 5.670, r= .386 p<.001***	$\bar{a}$= 0.638, σ= 0.447 $\bar{b}$= -0.541, σ= 0.387 Z = 12.642, r= .860 p<.001***
Stress VE	$\bar{a}$= 0.951, σ= 0.605 $\bar{b}$= -0.481, σ= 0.673 Z = 12.009, r= .817 p<.001***	—	$\bar{a}$= 0.951, σ= 0.605, $\bar{b}$= -0.930, σ= 0.458 Z = 12.671, r= .862 p<.001***	$\bar{a}$= 0.951, σ= 0.605, $\bar{b}$= 0.195, σ= 0.713 Z = 9.686, r= .659 p<.001***
Calm VE	$\bar{a}$= 0.675, σ= 0.511 $\bar{b}$= 0.381, σ= 0.498 Z = 6.528, r= .444 p<.001***	$\bar{a}$= 0.675, σ= 0.511 $\bar{b}$= -0.589, σ= 0.494 Z = 11.813, r= .804 p<.001***	—	$\bar{a}$= 0.675, σ= 0.511 $\bar{b}$= -0.440, σ= 0.425 Z = 12.361, r= .841 p<.001***
Sad VE	$\bar{a}$= 0.496, σ= 0.765 $\bar{b}$= -0.543, σ= 0.606 Z= 10.205, r= .694 p<.001***	$\bar{a}$= 0.496, σ= 0.765 $\bar{b}$= -0.190, σ= 0.579 Z = 9.895, r= .673 p<.001***	$\bar{a}$= 0.496, σ= 0.765 $\bar{b}$= 0.044, σ= 0.532 Z = 5.637, r= .384 p<.001***	—

with standardised coefficients as they indicate effect sizes and can be compared against one another.

The results for the regression models are found in Table 4, showing the standardised coefficients of predictors, asterisks to indicate their level of significance, coefficients of determination R^2, and assumption violations ($\not{L}$, $\not{H}$ and $\not{N}$). Both tables indicate for each predictor whether its regression coefficient is positive (green) or negative (red), and show separate regression models for the same affect rating for different exercise intensities if coefficients differ significantly across exercise intensities. Table 4 highlights coefficients that were hypothesised to be significant in bold, revealing that most hypotheses were accepted, but some were rejected (for those coefficients that are bold but not coloured). According to Falk and Miller [57], models with an $R^2 \geq 0.1$ can be considered as 'adequate'. The results in Table 4 show that regression models with the highest 'personalised' level of cleaning outperform all others in terms of model fit and can adequately predict Arousal (R^2 = .321), Calm (R^2 = .291), Stress (R^2 = .256), Valence (R^2 = .170), Fear (R^2 = .169), Excited (R^2 = .128), and Content (R^2 = .128). Having adequate models for both Valence and Arousal shows that regression models can in principle be used to predict a wide range of emotions on the circumplex model. However, we did not find adequate models for Sad (R^2 = .040), Bored (R^2 = .054) and Happy (R^2 = .085).

Table 4 reveals the best physiological measures for predicting each affect variable based on the magnitude of the standardised coefficients (reported in brackets in this paragraph). The significant predictors are also summarised in Figure 1-J. The only physiological measures that we found to be significant predictors were PDL, PDR, Power, Smile, and SCL. For all of the adequate models the pupillometry measures are the best predictors in all cases with one exception: Excited (Power = .182, PDR = .138, Smile = .112), which is also the only adequate model where PDR but not PDL is a significant predictor. In all other adequate models, PDL is a better predictor than PDR with the exception of Stress (PDR = .236, PDL = .231, Power = .171, SCL = .078). Apart from PDR and PDL, we see Power as a significant predictor in all adequate models, e.g. for Fear (PDL = .276, PDR = .158, Power = .105) and Calm (PDL = -.302, PDR = -.222, Power = -.167), with the exception of Content (PDL = -.224, PDR = -.129, Smile = .062). Valence (PDL = -.246, PDR = -.158, and Power = -.103) and Arousal (PDL = .250, PDR = .250, Power = .237, SCL = .111, Smile = .082), which can be used for predicting other emotions, have similar predictors but with opposite relationships for PDL, PDR, and Power. However, Arousal is also predicted by SCL and unexpectedly by Smile, which contribute to the better model fit compared to Valence along with PDR and Power being better predictors. Finally, despite their inadequate models, Happiness (PDL = -.184, Smile = .125, PDR = -0.081), Boredom (Smile = -.116, PDR = -.108, PDL = -.077), and Sadness (PDL = .150) still have significant predictors, which all include pupillometry measures.

5.3 RQ3: Data Cleaning

We compared the three levels of cleaning by inspecting their respective regression models, i.e. their standardised coefficients, significance of coefficients, and coefficient of determination R^2. In

Table 4: RQ2/RQ3 results: Overview of all affect models with standardised coefficients and overall coefficients of determination R^2. Green and red highlighting is used to denote significant positive and negative predictors respectively. Affect measures are predicted by pupil dilation level (PDL) and response (PDR), blink rate (BR) and duration (BD), skin conductance level (SCL), skin conductance response (SCR), heart rate (HR), heart rate variability (HRV), zygomaticus major activity (Smile), and bike power output (Power). Violations of regression assumptions are denoted as $\not{L}$ (Linearity), $\not{H}$ (Heteroskedasticity), and $\not{N}$ (Normality). Unbalanced residual plots are denoted with U.

DV	Cleaning	Intensity	R^2	PDL	PDR	BR	BD	SCL	SCR	HR	HRV	Smile	Power
	Pers	All	0.17	**-0.246*****	**-0.158*****	-0.019	-0.071	-0.013	-0.041	-0.063	**0.004**	**0.044**	**-0.103*****
Valence	Env	All	0.1	**-0.279*****	**-0.228*****	-0.01 U	-0.031 U	-0.016 U	-0.064 U	-0.087	**-0.075** U	**-0.04** U	**-0.118*****
	Raw	Low	0.155	**-0.552*****	**0.014**	0.068 U	-0.042 $\not{H}$	-0.012 U	0.048 U	-0.004	**-0.049** U	**0.042** U	**-0.017**
		Med	0.132	**-0.507*****	**0.065**	-0.022 U	-0.073 $\not{H}$	0.009 U	-0.138 U	0.111	**0.003** U	**-0.040** U	**-0.129**
		High	0.122	**-0.405*****	**-0.083**	-0.035 U	0.029 $\not{H}$	0.032 U	-0.050 U	0.046	**0.044** U	**-0.105** U	**-0.167**
	Pers	All	0.321	**0.250*****	**0.250*****	-0.011	0.024	**0.111*****	**-0.020**	-0.028	0.024	0.082*	**0.237*****
Arousal	Env	All	0.166	**0.346*****	**0.289*****	0.011 U	-0.002 U	**0.076****	**-0.043** U	0.074	0.014	0.112* U	**0.221*****
	Raw	Low/Med	0.174	**0.397*****	**0.209*****	0.019 U	-0.015 U	**-0.182**** U	**0.003** U	0.042	0.036 U	0.105 U	**0.227*****
		High	0.116	**0.267*****	**0.155**	-0.032 U	-0.012 U	**-0.244** U	**0.126** U	-0.026	-0.075 U	0.039 U	**0.236***
	Pers	All	0.169	**0.276*****	**0.158*****	0.057	0.038	0.017	-0.038	-0.027	-0.014	0.055	**0.105****
Fear	Env	Low/Med	0.106 $\not{N}$	**0.301***** $\not{L}$	**0.214*****	0.099* U	0.001 $\not{H}$	-0.004	0.036 U	0.008	0.005 U	0.086 U	**0.060**
		High	0.040 $\not{N}$	**0.158*** $\not{L}$	**0.214****	0.053 $\not{H}$	-0.021 $\not{H}$	0.050 $\not{H}$	-0.025 U	0.072	-0.049 U	-0.019 U	**0.125** $\not{H}$
	Raw	Low	0.146 $\not{N}$	**0.369***** $\not{L}$	**0.057**	0.110 $\not{H}$	-0.051 $\not{H}$	-0.058 U	0.064 U	0.127	0.046 U	-0.009 U	**0.052**
		Med	0.082 $\not{N}$	**0.487***** $\not{L}$	**-0.037**	0.085 U	0.089 $\not{H}$	-0.073 U	0.096 U	0.138	-0.078 U	0.124 U	**0.054**
		High	0.043 $\not{N}$	**0.245***** $\not{L}$	**0.104** $\not{L}$	0.034 U	-0.015 $\not{H}$	-0.237 U	0.107 U	0.079	0.038 U	-0.035 U	**0.122**
	Pers	All	0.256	**0.231*****	**0.236*****	0.004	0.051	**0.078***	**-0.002**	0.057	0.058	0.042	**0.171*****
Stress	Env	All	0.152	**0.296*****	**0.325*****	0.017 $\not{H}$	0.009 $\not{H}$	**0.070***	**-0.010** U	0.039	0.055 U	0.082 U	**0.231*****
	Raw	Low	0.224	**0.475***** $\not{L}$	**0.111**	0.003 $\not{H}$	0.005 $\not{H}$	**-0.131** U	**0.101** U	0.132	0.063 U	0.033 U	**0.147**
		Med/High	0.175	**0.492***** $\not{L}$	**0.077**	0.004 $\not{H}$	0.027 $\not{H}$	**-0.056** U	**0.021** U	0.139 $\not{H}$	0.021 $\not{H}$	0.084 U	**0.229*****
	Pers	All	0.085	**-0.184*****	**-0.081***	-0.041	-0.058	**0.041**	**-0.039**	-0.121	0.036	**0.125*****	0.054
Happy	Env	Low	0.036	**-0.178***	**-0.059**	0.075 $\not{H}$	-0.094 $\not{H}$	**-0.014** U	**-0.096** U	0.056 $\not{H}$	0.104 $\not{H}$	**0.082** U	0.054 $\not{H}$
		Med/High	0.065	**-0.166*****	**-0.084**	-0.060 $\not{H}$	-0.029 $\not{H}$	**0.018** U	**-0.101*** $\not{H}$	0.000 $\not{H}$	-0.137 U	**0.098** U	-0.105 $\not{H}$
	Raw	Low	0.057	**-0.390***** $\not{L}$	**0.081**	0.080 $\not{H}$	-0.079 $\not{H}$	**-0.062** U	**-0.016** U	0.026 $\not{H}$	0.000 U	**0.117** U	0.079 $\not{H}$
		Med	0.039	**-0.308*****	**0.096**	-0.002 $\not{H}$	-0.080 $\not{H}$	**0.051** U	**-0.206** $\not{H}$	0.043 $\not{H}$	0.056 U	**0.150** $\not{H}$	-0.046 $\not{H}$
		High	0.082	**-0.297*****	**0.008** $\not{H}$	-0.110 $\not{H}$	0.010 $\not{H}$	**-0.052** $\not{H}$	**0.004** $\not{H}$	-0.044 $\not{H}$	-0.058 U	**-0.025** U	-0.138 $\not{H}$
	Pers	All	0.040 $\not{N}$	**0.150***** $\not{L}$	**0.034**	-0.013 $\not{H}$	-0.004 $\not{H}$	-0.083	0.014	-0.035	-0.047	-0.020	0.063
Sad	Env	All	0.013 $\not{N}$	**0.140***** $\not{L}$	**0.099*** $\not{H}$	-0.026 $\not{H}$	-0.033 $\not{H}$	-0.046 U	0.013 $\not{H}$	0.004 $\not{H}$	-0.073 $\not{H}$	0.061 U	0.043 $\not{H}$
	Raw	All	0.026 $\not{N}$	**0.314***** $\not{L}$	**-0.026** $\not{L}$	-0.036 $\not{H}$	-0.028 $\not{H}$	0.028 $\not{H}$	-0.019 $\not{H}$	-0.017 $\not{H}$	-0.076 U	0.055 U	-0.007 $\not{H}$
	Pers	All	0.054	**-0.077****	**-0.108*****	0.081	0.036	-0.042	0.060	-0.055	-0.031	-0.116* $\not{H}$	0.015
Bored	Env	All	0.041 $\not{N}$	**-0.099*****	**-0.048**	-0.002 $\not{H}$	0.049 $\not{H}$	-0.074 U	0.092 U	0.025 $\not{H}$	0.008 $\not{H}$	-0.130* U	-0.073 $\not{H}$
	Raw	All	0.032 $\not{N}$	**0.001** $\not{H}$	**-0.110**** $\not{H}$	-0.002 $\not{H}$	0.049 $\not{H}$	0.204* U	-0.040 U	0.004 $\not{H}$	-0.005 U	-0.117 U	-0.100 $\not{H}$
	Pers	All	0.128	0.104	0.138**	-0.052	-0.039	-0.096	-0.067	-0.125	0.028	0.112*	0.182**
Excited	Env	Low	0.049	0.094	0.103	-0.022 U	-0.052 $\not{H}$	0.052 U	-0.027 U	0.126 $\not{H}$	0.039 U	0.054 U	0.162 $\not{H}$
		Med/High	0.067	0.105	0.123	-0.001 U	-0.031 $\not{H}$	0.015 U	-0.088 $\not{H}$	-0.006 $\not{H}$	-0.111 U	0.142* U	0.150 $\not{H}$
	Raw	Low	0.067	-0.072 $\not{H}$	0.217* $\not{H}$	-0.021 $\not{H}$	-0.060 $\not{H}$	-0.066 U	0.010 U	0.154 $\not{H}$	-0.039 U	0.036 U	0.197 $\not{H}$
		Med	0.055	-0.001	0.131	0.023 $\not{H}$	-0.048	0.022 $\not{H}$	-0.139 $\not{H}$	0.102 $\not{H}$	0.086 U	0.172 U	0.177 $\not{H}$
		High	0.032	-0.005	0.038 $\not{H}$	-0.032 U	-0.033 $\not{H}$	-0.053 $\not{H}$	-0.035 $\not{H}$	0.021	-0.095 U	0.058 U	0.131 $\not{H}$
	Pers	All	0.128	**-0.224*****	**-0.129*****	-0.026	-0.039	-0.014	-0.046	-0.118	0.041	**0.062***	0.000
Content	Env	Low	0.055	**-0.260*****	**-0.141***	0.088 $\not{H}$	0.016 $\not{H}$	-0.017 U	-0.046 U	-0.032 $\not{H}$	0.001 U	**-0.019** U	-0.001 $\not{H}$
		Med/High	0.056	**-0.213***** $\not{H}$	**-0.139**** $\not{H}$	-0.010 U	-0.016 $\not{H}$	0.011 U	-0.102 $\not{H}$	-0.078 $\not{H}$	-0.143 U	**0.048** U	-0.140 $\not{H}$
	Raw	LowMed	0.046	**-0.388*****	**-0.076** $\not{H}$	0.051 $\not{H}$	-0.034 $\not{H}$	-0.038 U	-0.018 U	-0.065 $\not{H}$	0.040 U	**0.023** U	-0.061 $\not{H}$
		High	0.069	**-0.295***** $\not{H}$	**-0.116** $\not{H}$	-0.019 U	0.003 $\not{H}$	-0.009 U	-0.031 $\not{H}$	-0.002 $\not{H}$	-0.018 U	**-0.050** U	-0.136 $\not{H}$
	Pers	All	0.291	**-0.302*****	**-0.222*****	-0.006	-0.038	-0.013	-0.047	-0.060	0.033	0.029 $\not{H}$	**-0.167*****
Calm	Env	Low	0.095	**-0.335*****	**-0.254*****	0.014 $\not{H}$	-0.099 $\not{H}$	-0.021 U	0.021 U	-0.117	-0.010 U	-0.032 U	**-0.188**** $\not{H}$
		Med/High	0.135	**-0.344*****	**-0.262*****	-0.029 $\not{H}$	-0.038 $\not{H}$	-0.040 U	-0.093 $\not{H}$	-0.047 $\not{H}$	-0.066 U	-0.074 U	**-0.285***** $\not{H}$
	Raw	Low	0.162	**-0.463***** $\not{L}$	**-0.227**** $\not{L}$	0.035 $\not{H}$	-0.075 $\not{H}$	0.031 U	0.079 U	-0.175 $\not{H}$	-0.024 U	0.029 $\not{H}$	**-0.154*** $\not{H}$
		Med	0.183	**-0.501*****	**-0.032**	-0.039 $\not{H}$	-0.056 $\not{H}$	0.277 U	-0.263 $\not{H}$	-0.036 $\not{H}$	0.123 U	-0.138 U	**-0.257**** $\not{H}$
		High	0.16	**-0.462***** $\not{L}$	**-0.071** $\not{H}$	-0.010 U	-0.044 $\not{H}$	0.102 U	-0.141 $\not{H}$	-0.069 $\not{H}$	0.059 U	-0.051 $\not{H}$	**-0.270***** $\not{H}$

particular, we looked for notable changes such as changes in the sign of a coefficient or violations of the assumption that better cleaning improves model fit. Note that the Chow test cannot be applied to compare the regression models as they are based on the same data sets [110].

Table 4 also provides an overview of the effects the three levels of cleaning (*Pers*, *Env* and *Raw*) have on the regression models. For

Table 5: RQ4 results: Overview of regression models describing the relationship between physical exertion and affect, with standardised coefficients for bike power output (Power) and overall coefficient of determination R^2. The highest level of cleaning (Pers) was used for all models. Violations of regression assumptions are denoted as $\not{L}$ (Linearity), $\not{H}$ (Heteroskedasticity), and $\not{N}$ (Normality).

DV	Intensity	R^2	Power
Valence	Low/Med	0.045	-0.190***
	High	0.062	-0.211***
Arousal	All	0.154	0.392***
Fear	All	0.063 $\not{N}$	0.251*** $\not{L}$
Sad	All	0.011 $\not{N}$	0.105** $\not{L}$
Bored	All	0.009	-0.093**
Content	Low	0.008	-0.060
	Med	0.025	-0.140* $\not{H}$
	High	0.025	-0.132*
Calm	Low/Med	0.106	-0.293***
	High	0.120	-0.325***
Happy	Low	0.002	0.080 $\not{H}$
	Med/High	0.007	-0.037
Excited	Low	0.085	0.292***
	Med/High	0.035	0.206***
Stress	Low/Med	0.089	0.286*** $\not{L}$
	High	0.088	0.271***

each affect, a model (or more if coefficients are inconsistent across exercise intensities) is provided for each cleaning level. Models are compared by their standardised coefficients, their significance, and coefficients of determination R^2 as indicators of model fit.

Table 4 demonstrates that regression models with different levels of cleaning for a specific affect variable exhibit significant coefficients that are largely consistent in their sign. However, these models widely differ in fit and violated assumptions. For example, for all affect variables, the coefficient of determination R^2 is always higher for the personalised cleaning level compared to raw and environmental cleaning, indicating a better model fit.

Similarly, the models generally exhibit increased robustness to the effects of exercise with higher levels of data cleaning. Models at the personalised cleaning level are consistently not separated by exercise intensity. Additionally, regression assumptions tend to be violated or unbalanced at the raw level but are typically valid at the personalised level. Notably, the model fit typically worsens when transitioning from raw to environmental cleaning, as shown by a decrease in R^2. This phenomenon is a result of physiological markers correlating with environmental stimuli rather than a user's affective response, such as pupils responding to light rather than emotion. Further discussion on this topic can be found in section 6.

5.4 RQ4: Exertion and Affect

Similar to RQ2, we used multi-level linear regression models and Chow tests to analyse the relationships between affect variables and physical exertion. We first regressed each affect variable onto power output as an indicator of physical exertion, using the highest 'personalised' cleaning level. Table 5 provides an overview of the regression models describing the relationships between physical exertion and affect. The highlighted cells indicate that most regressions were significant, although with 'weak' coefficients of determination, and some of them violate regression assumptions.

Table 6: RQ4 results: Comparisons of IMI subscale scores between different levels of exercise intensity. The first column lists the two compared levels of exercise intensity *a* and *b*. The following columns each compare IMI subscale scores between levels *a* and *b*. Cells show the means $\bar{a}$ and $\bar{b}$ and standard deviations σ for the two levels, and non-parametric Wilcoxon signed-rank test results with effect size r ($r < 0.3$ for 'small', $0.3 \geq r < 0.5$ for 'moderate', and $r \geq 0.5$ for 'large').

Exercise Intensity		IMI Interest	IMI Pressure	IMI Competence
a) Low	*b*) Med	$\bar{a}$= 4.984, σ= 1.149 $\bar{b}$= 4.645, σ= 1.253 Z= 6.473, r= .381 p<.001***	$\bar{a}$= 2.736, σ= 1.069 $\bar{b}$= 2.931 , σ= 1.063 Z= -3.989, r= -.235 p<.001***	$\bar{a}$= 4.493, σ= 1.308 $\bar{b}$= 4.197, σ= 1.260 Z= 4.432, r= .261 p<.001***
a) Low	*b*) High	$\bar{a}$= 4.984, σ= 1.149 $\bar{b}$= 4.623, σ= 1.329 Z= 6.488, r= .382 p<.001***	$\bar{a}$= 2.736, σ= 1.064 $\bar{b}$= 3.389, σ= 1.252 Z= -7.318, r= -.431 p<.001***	$\bar{a}$= 4.493, σ= 1.308 $\bar{b}$= 3.794, σ= 1.302 Z= 8.744, r= .515 p<.001***
a) Med	*b*) High	$\bar{a}$= 4.645, σ= 1.253 $\bar{b}$= 4.623, σ= 1.336 Z= 0.575, r= .034 p= .566	$\bar{a}$= 2.931, σ= 1.063, $\bar{b}$= 3.389, σ= 1.252 Z= -5.422, r= -.320 p<.001***	$\bar{a}$= 4.197, σ= 1.260 $\bar{b}$= 3.794, σ= 1.302 Z = 5.779, r= .341 p<.001***

We then also analysed the relationship between physical exertion and IMI scores, which were only measured once per level of exercise intensity. We did this by regressing each IMI subscale score onto the level of exercise intensity, encoded as 0 for low, 1 for medium and 2 for high-intensity. The encoded level of exercise intensity was treated as interval-scaled predictor because the three levels are equally spaced in terms of their ranges of heart rate reserve (50%-60%, 60%-70% and 70%-80%). Multi-level linear regressions showed that the exercise intensity level significantly decreased IMI Interest/Enjoyment ($B = -0.181, t(214) = -3.113, p = .002$) and Perceived Competence ($B = -0.350, t(214) = -5.612, p < .001$), as well as significantly increasing Pressure/Tension ($B = 0.326, t(214) = 4.847, p < .001$).

In addition to linear regressions, we also performed repeated measures ANOVA for each IMI subscale with exercise intensity level as the independent variable. Similar to our analysis for RQ1, after testing normality using Shapiro-Wilk tests and inspecting QQ-plots we decided to non-parametric test alternatives to address any concerns about violations of normality. We tested the overall effects of exercise intensity level on an IMI subscale score. If Mauchly's tests indicated a violation of sphericity, Huynh-Feldt correction was used. If the main effect of exercise intensity level was significant, we performed pairwise Wilcoxon signed-rank tests with Holm–Bonferroni correction.

The main effect of exercise intensity level was significant for all IMI subscales, i.e. Interest/Enjoyment ($\chi^2(2) = 12.878$, $W = 0.089$, $p = .002^{**}$), Pressure/Tension ($\chi^2(2) = 14.210$, $W = 0.099$,

$p < .001^{***}$), and Perceived Competence ($\chi^2(2) = 14.127$, $W = 0.098$, $p < .001^{***}$). Table 6 summarises the results of all pairwise comparisons, which support the results of the regression models.

6 DISCUSSION

In this section we first discuss our findings for each research question and suggest future work. Then we provide practical recommendations for affect recognition in VR exergames.

6.1 RQ1: Affect Manipulation

Our VEs were consistent with design suggestions from related work on emotion-inducing stimuli [60, 66, 99, 114, 165, 185]. The results validate these design choices, with all VEs eliciting significantly more of their target emotion than the other VEs, and the respective target emotions significantly more dominant in each VE compared to the other target emotions. While it might not always be desirable for exergames to elicit some of these emotions, it is important that affective exergames can *detect* them to optimise the player experience. It is important to elicit these emotions appropriately to build robust affect recognition models [89].

Our results indicate that to elicit Happiness, Stress, Calmness, and Sadness irrespective of exercise intensity level, researchers and exergame designers should consider exergame mechanics and difficulty (e.g. the quantity of 'rewards' and obstacles) [99, 165], communication and feedback to the player (e.g. messages of encouragement or countdown timers) [32], the aesthetics of the exergame environment (e.g. lighting, skybox colours, terrain textures, and game objects) [50, 66], and the sound design (e.g. game object sound effects, ambient sound effects, and soundtracks) [60, 114, 131]. We provide the full Unity implementations of our exergame VEs for other researchers and designers to build upon.

Considering the results in Figure 4, Fear and Sadness ratings were comparatively low. While Fear was not a target emotion for our VEs, this could be explored further in a VR exergaming context, e.g. in a survival-horror exergame and with jump scares similar to Müller et al. [123]. However, Sadness was a target emotion and we made informed design decisions to elicit it appropriately. Müller et al. [123] elicited Sadness through repeated failure, which is similar to our Stress VE with the inclusion of Skull coins. While the Sadness VE did increase Sadness ratings, further steps could be taken to induce it, e.g. by including sad narrative devices or staging sad social situations through non-player characters.

Furthermore, Boredom, which plays an important role in player experience, could be explored further by designing a VE that is repetitive, linear in gameplay, and lacks visual variety, challenge, and interaction [11, 129]. Boredom is likely more easily induced at lower exercise intensities, due to the reduced challenge, and will likely be elicited more reliably over longer gameplay sessions.

6.2 RQ2: Affect Recognition

Our results confirm many physiological measures reported in the affect literature as significant predictors of affect during VR exergaming across different exertion levels. In particular, pupil dilation (PDL and PDR) was a common, strong predictor and the signs of PDL/PDR coefficients agree with the literature and pilot results [2, 29, 36, 37, 92, 125, 133, 173, 206]. Furthermore, we confirmed Smile as a positive predictor of Happiness, Excitement, and Contentness [54, 109, 159, 210]. PDL and PDR significantly predicted Boredom as hypothesised; however, our results agreed with literature on directionality (negative) [195] rather than our pilot results. We also confirmed SCL as a positive predictor of Arousal [15, 29, 155] and Stress [22, 179].

Some physiological measures commonly used in non-exercise contexts were rejected as predictors during VR exergaming. HRV and Smile did not predict Valence as was initially hypothesised [109, 159, 167, 182]. For HRV, this could be a consequence of the well-established limitations of below-24-hour measures of HRV as described by Shaffer and Ginsberg [166]. However, HRV should be explored further for affective VR exergaming as it has been shown to be a strong predictor of Valence outside of exergaming contexts. Additional measures (e.g. PNN50) and cleaning approaches could be necessary to increase the predictive power of HRV in exergaming. For Smile, zygomaticus major activity could be influenced by physical responses to exercise such as mouth-agape panting, which could also explain why it was a significant predictor of Arousal. Despite this, Smile was still a significant predictor of discrete high Valence emotions — Happiness, Excitement, and Contentness. To improve Smile as a predictor of Valence, alternative sensing approaches could be explored, such as fEMG integrated into the VR headset to directly sense zygomaticus major and corrugator supercilii activation rather than analysing blend shapes of the mouth provided by a visual lip tracker.

Interestingly, while SCL was a strong predictor of Arousal and Stress, our results reject SCR as a significant predictor despite being hypothesised [15, 22, 29, 155, 179]. Our results also reject SCL and SCR as significant predictors of Happiness [210]. These shortcomings of SCL and especially SCR for affect recognition in exergaming could be due to ceiling effects in EDA while exercising. The EDA induced by exercise may outweigh any activity induced by emotion. Future work could look at varying EDA electrode placement on the user to better measure SCL/SCR responses to affect during VR exergames, such as placement on the plantar fascia (foot) instead of the palm. This placement has been shown to be more robust to motion artefacts during weight lifting [82], but has yet to be explored for cycling. For SCR, alternatives to the 'EDA positive change' measure [112] such as event-related skin conductance responses [96] could also be explored.

The inadequate models for Sadness, Boredom, and Happiness may be a result of these emotions being more difficult to elicit [52], and we observed a large variation in participant responses. Boredom was not directly targeted and the responses indicate that it was quite low across all VEs; this is unsurprising given the generally arousing nature of VR exergaming and the fact that many participants had not experienced VR exergames before. The Happy and Sad VEs may not have elicited deep feelings of happiness and sadness given the relatively abstract nature of the VEs and tasks. Another explanation for the inadequate model for detecting Sadness could be a hidden variable influencing participant's pupil dilation unrelated to sadness. It is the only model that violates assumptions of normality and linearity for PDL at the highest level of cleaning, with the residual plots showing a left-skewed distribution and many more residuals in the positive Sad range. This indicates that the model is overestimating sadness using PDL. A potential influence

could be the luminosity of the environment, despite correcting for this by decoupling light reflex [146]. To improve the model, different temporal lag parameters for the luminosity correction could be explored to account for interpersonal differences in pupillary response time to light [71].

Our affect recognition models yielded some unexpected results. Perhaps the most interesting unexpected result was that the affect model for Excited had an adequate model fit with three significant predictors, despite not being hypothesised at all. This is a particularly novel finding as there is limited related work exploring physiological correlates of excitement, especially in the context of VR exergaming. Feelings of excitement and amazement are key affordances in VR, so it could be important to measure and monitor excitement during these experiences to understand how and when it is formed. For affective exergaming, such an understanding could be operationalised in adaptive environments to optimise the user experience. Another unexpected result that was not hypothesised by prior work was that Smile was a predictor of Boredom, i.e. the more bored participants were the less they smiled. As hypothesised, Smile is a strong predictor of Happy, and participants are more likely to be engaged in the experience if they are enjoying it [85]. Not smiling could therefore relate to lack of engagement and in turn increased boredom. However, it is important to note that both Happy and Bored had 'inadequate' models, and that Smile as a predictor of boredom violated the assumption of Heteroskedasticity and was only barely significant at the highest cleaning level. Therefore, the relationship between Smile and boredom warrants further investigation.

The insights from the regression models, the evidence of how data cleaning increases predictive power and validity, and our open source dataset all provide a springboard for future research and development in affect recognition, including new approaches using ML. Apart from addressing the problem of affect recognition as a whole, ML could also be used to address specific sub-problems such as recognising individual SCR responses in a continuous data stream. Hybrid models combining ML with statistical regressions may improve predictions whilst maintaining transparency and understanding between physiology and affect. For example, ML may improve the predictive power of physiological measures that were not found to be significant or which do not have a linear relationship, such as HRV where the relationship to Valence is well evidenced in related work [167].

6.3 RQ3: Data Cleaning

The differences in Table 4 between the three levels of cleaning (*Raw*, *Env*, and *Pers*) showcase a trend of increasingly significant predictors, decreased violations of assumptions, and increased coefficients of determination R^2. Additionally, the models become more consistent across exercise intensities and are hardly separated by intensity at the *Pers* level, whereas they are almost always separated at the *Raw* level. These general trends in cleaning levels highlight in particular the importance of z-score transforms to account for interpersonal differences, which has also been shown in other affective exergaming research albeit in a more limited scope [13, 15].

A crucial finding is that the coefficient of determination often decreases between *Raw* and *Env*. This is most likely a prime example of the affect models at the *Raw* level overfitting to the emotional stimuli, rather than recognising the affect they elicit. For example, environmental luminance is not accounted for at the *Raw* level and the positive valence VEs were generally brighter than the negative valence VEs, resulting in smaller pupil diameter in positive VEs. At the *Env* level, the model becomes less fitted to the stimuli due to accounting for pupillary light reflexes. With this in mind, we recommend that researchers and developers should consider the physiological byproducts of their stimuli when building affect recognition models both inside and outside of VR exergaming. In our case, we sampled brightness based on the foveal position inferred from the eye tracker (2° visual angle) to sanitise pupil dilation measures; however, sampling the brightness of the entire image in the headset may also be a sufficient cleaning measure in cases where the gaze position is inaccurate or unavailable. This warrants further investigation.

Another example of the importance of data cleaning is highlighted by SCL as a predictor of Arousal. At the *Raw* level for low and medium exercise intensity, SCL is significantly *negatively* correlated with Arousal with a large coefficient and an apparently 'adequate' model fit ($R^2 = 0.174$) — something which directly conflicts with the literature on SCL and Arousal. However, by taking into account existing sweat levels of the user at the *Env* and *Pers* cleaning levels, SCL becomes significantly *positively* correlated, in line with related work.

The relationship between data cleaning and model adequacy is stark, yet there is still room for additional cleaning methods that could increase the predictive power of some physiological measures. For example, our results did not show BR, BD, SCR, HR, or HRV to be significant predictors of affect during VR exergaming, warranting further investigation into how they can be cleaned.

6.4 RQ4: Exertion and Affect

The significant relationships shown in Table 5 and Table 6 can be explained by known effects of exercise [18, 53, 137, 187]. For example, discomfort as a result of intense exercise can reduce valence, especially when exceeding the ventilatory threshold [13, 15, 178]. Exercise is also often used to induce arousal in psychological studies [107]. However, most regressions shown in Table 5 have only weak coefficients of determination and effect sizes in Table 6 are small to moderate, reaffirming the assertion presented in related work that VR exergames can distract users from uncomfortable sensations in exercise [122, 141, 147, 178, 197]. In light of the results, designers should consider the type of emotion(s) they want to elicit in an exergame and match the activity and exertion level to be complementary. Table 5 can be used as a reference to select the right exertion level to support the emotion and experience they are trying to achieve. For example, if an exergame should be stress-inducing or exciting, exercise activities with higher intensity levels should be considered.

6.5 Guidelines for Affect Recognition in VR Exergames

Based on our results, we make the following recommendations for building affect recognition into VR exergames:

(1) Incorporate pupillometry (PDL and PDR) with luminosity correction because it provides the strongest predictors for almost all affect variables.

(2) Incorporate the user's power output because it is a powerful predictor of both Valence and Arousal as well as most other affect variables.

(3) Take the preexisting sweat levels of a user into account when using SCL to predict Arousal and Stress.

(4) Avoid linear regression models for predicting Sadness, Boredom, and Happiness.

(5) Clean sensor data using the personalised approach as this provides the best predictive power and validity by accounting for interpersonal differences.

(6) Do not use raw data without any cleaning as this can lead to overfitting and erroneous predictions.

(7) Use multiple physiological sensors as this will increase predictive power.

(8) Do not use blink measures as they provide little benefit.

6.6 Limitations

Our study used a within-participant design, which meant that our results were influenced by participant familiarity and physical fatigue. However, this was mitigated through counterbalancing and providing participants extended breaks after an exercise bout. Participants were recruited through convenience and snowball sampling, resulting in a small bias towards males, younger participants, and more physically active people. Due to our large sample size our results are still generalisable to women, people from a fairly wide age range (20s to 40s), and people who are only moderately active. Furthermore, participants only had a brief experience playing the VR exergame, with 12 minutes of gameplay total excluding warm ups and cool downs. Typically exergames are played for longer periods and over multiple gameplay sessions. Future work could consider the longitudinal aspects of VR exergames and how players' emotional responses evolve over repeated gameplay sessions.

Additionally, our VR exergame and dataset consider only one type of exercise: cycling. A clear avenue for future work is to apply and validate the same physiological measures, cleaning procedures, and regression models for other types of VR exergames, e.g., other cardiovascular exercises such as running and rowing, and strength exercises such as weight-lifting. Future work could also explore different exergame genres and game mechanics to target emotions beyond what was explored in this paper. For instance, a horror exergame could induce fear, and a multiplayer exergame could introduce social dynamics of communication and competition.

Reflecting on our affect recognition models, we could have used other popular approaches such as ML to recognise affect. However, our goal was to provide transparency on the relationships between emotions and physiological responses in the context of VR exergaming. Our results can inform parameter choice for future ML affect models as well as provide validated approaches for removing environmental and interpersonal artefacts in physiological data.

Future work could consider the game context in affective exergaming, allowing for appraisal-based affect recognition models to be constructed. By considering the context of what a player is currently experiencing in a VR exergame, a model can appraise estimates of core affect [154] in light of the context, e.g., interpret physiological responses in the context of a user colliding with an obstacle or defeating a difficult opponent. While we did not consider context in our affect recognition models, our open dataset provided in Supplementary Materials also contains exergame data (such as coins collected), which we invite researchers to analyse and apply their own models to.

6.7 Impact

Our work advances affective VR exergames by providing guidelines on sensor and parameter choices for affect recognition models. The results can also be used by designers to inform exergame activity and environment design to target specific emotions. Exergaming is a notoriously noisy environment for physiological sensing; our affect recognition models and approaches to data cleaning could be applied, adapted, and validated for other equally noisy contexts such as industrial applications that involve physical labour.

7 CONCLUSION

We developed and validated four virtual environments to induce specific emotions in a VR cycling exergame, which were then used to analyse the relationship between ten physiological measures and ten affect ratings. We constructed affect recognition models across three exercise intensities, and three levels of data cleaning that account for environmental and interpersonal factors. Finally, we tested the relationship between affect and physical exertion. In summary, this led us to the following conclusions:

(1) Emotions can be consistently induced across exercise intensities in VR exergaming.

(2) Despite VR exergaming creating a lot of noise in physiological sensing, we identified several significant predictors of affect, with pupil dilation being the strongest.

(3) Data cleaning of environmental and interpersonal factors is important and not only improves predictive power but also removes violations of assumptions for linear regression models.

(4) There is a significant albeit weak relationship between physical exertion and most measures of affect.

Our findings support the design of adaptive VR exergaming experiences that optimise enjoyment, performance, and adherence.

ACKNOWLEDGMENTS

This work is supported by the European Union's Horizon Europe research and innovation program and Innovate UK under grant agreement *No 101070533*, project EMIL (The European Media and Immersion Lab): https://emil-xr.eu/.

REFERENCES

[1] Hervé Abdi. 2007. Z-scores. *Encyclopedia of measurement and statistics* 3 (2007), 1055–1058.

[2] Yasmeen Abdrabou, Khaled Kassem, Jailan Salah, Reem El-Gendy, Mahesty Morsy, Yomna Abdelrahman, and Slim Abdennadher. 2018. Exploring the usage of EEG and pupil diameter to detect elicited valence. In *Intelligent Human*

Systems Integration: Proceedings of the 1st International Conference on Intelligent Human Systems Integration (IHSI 2018): Integrating People and Intelligent Systems, January 7-9, 2018, Dubai, United Arab Emirates. Springer, 287–293.

[3] Ashwaq Alhargan, Neil Cooke, and Tareq Binjammaz. 2017. Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 479–486.

[4] Naomi Altman and Martin Krzywinski. 2016. Regression diagnostics: residual plots can be used to validate assumptions about the regression model. *Nature Methods* 13, 5 (2016), 385–387.

[5] Samira Aminihajibashi, Thomas Hagen, Maja Dyhre Foldal, Bruno Laeng, and Thomas Espeseth. 2019. Individual differences in resting-state pupil size: Evidence for association between working memory capacity and pupil size variability. *International Journal of Psychophysiology* 140 (2019), 1–7.

[6] Bradley M Appelhans and Linda J Luecken. 2006. Heart rate variability as an index of regulated emotional responding. *Review of general psychology* 10, 3 (2006), 229–240.

[7] Janice Attard-Johnson, Caoilte Ó Ciardha, and Markus Bindemann. 2019. Comparing methods for the analysis of pupillary response. *Behavior Research Methods* 51 (2019), 83–95.

[8] Değer Ayata, Yusuf Yaslan, and Mustafa Kamaşak. 2016. Emotion recognition via random forest and galvanic skin response: Comparison of time based feature sets, window sizes and wavelet approaches. In *2016 Medical Technologies National Congress (TIPTEKNO)*. IEEE, 1–4.

[9] Ebrahim Babaei, Benjamin Tag, Tilman Dingler, and Eduardo Velloso. 2021. A Critique of Electrodermal Activity Practices at CHI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 177, 14 pages. https://doi.org/10.1145/3411764.3445370

[10] Areej Babiker, Ibrahima Faye, and Aamir Malik. 2013. Pupillary behavior in positive and negative emotions. In *2013 IEEE International Conference on Signal and Image Processing Applications*. 379–383. https://doi.org/10.1109/ICSIPA.2013.6708037

[11] Fabrizio Balducci, Costantino Grana, and Rita Cucchiara. 2017. Affective level design for a role-playing videogame evaluated by a brain–computer interface and machine learning methods. *The Visual Computer* 33 (2017), 413–427.

[12] Serdar Baltaci and Didem Gokcay. 2016. Stress detection in human–computer interaction: Fusion of pupil dilation and facial temperature features. *International Journal of Human–Computer Interaction* 32, 12 (2016), 956–966.

[13] Soumya C Barathi, Daniel J Finnegan, Matthew Farrow, Alexander Whaley, Pippa Heath, Jude Buckley, Peter W Dowrick, Burkhard C Wuensche, James LJ Bilzon, Eamonn O'Neill, et al. 2018. Interactive feedforward for improving performance and maintaining intrinsic motivation in VR exergaming. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[14] Soumya C. Barathi, Daniel J. Finnegan, Matthew Farrow, Alexander Whaley, Pippa Heath, Jude Buckley, Peter W. Dowrick, Burkhard C. Wuensche, James L. J. Bilzon, Eamonn O'Neill, and Christof Lutteroth. 2018. Interactive Feedforward for Improving Performance and Maintaining Intrinsic Motivation in VR Exergaming. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173982

[15] Soumya C Barathi, Michael Proulx, Eamonn O'Neill, and Christof Lutteroth. 2020. Affect recognition using psychophysiological correlates in high intensity vr exergaming. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[16] Gershon Ben-Shakhar. 1985. Standardization within individuals: A simple method to neutralize individual differences in skin conductance. *Psychophysiology* 22, 3 (1985), 292–299.

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

[18] Elaine Biddiss and Jennifer Irwin. 2010. Active video games to promote physical activity in children and youth: a systematic review. *Archives of Pediatrics & Adolescent Medicine* 164, 7 (2010), 664–672.

[19] Raphaël Bize, Jeffrey A Johnson, and Ronald C Plotnikoff. 2007. Physical activity level and health-related quality of life in the general adult population: a systematic review. *Preventive medicine* 45, 6 (2007), 401–415.

[20] Paul Bliese. 2006. Multilevel Modeling in R (2.2)–A Brief Introduction to R, the multilevel package and the nlme package.

[21] Silke Boettger, Christian Puta, Vikram K Yeragani, Lars Donath, Hans-Josef Mueller, Holger H Gabriel, and Karl-Juergen Baer. 2010. Heart rate variability, QT variability, and electrodermal activity during exercise. *Med Sci Sports Exerc* 42, 3 (2010), 443–8.

[22] Gunilla Bohlin. 1976. Delayed habituation of the electrodermal orienting response as a function of increased level of arousal. *Psychophysiology* 13, 4 (1976), 345–351.

[23] John Bolton, Mike Lambert, Denis Lirette, and Ben Unsworth. 2014. PaperDude: A Virtual Reality Cycling Exergame. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI EA '14)*. Association for Computing Machinery, New York, NY, USA, 475–478. https://doi.org/10.1145/2559206.2574827

[24] Felix Born, Linda Graf, and Maic Masuch. 2021. Exergaming: The Impact of Virtual Reality on Cognitive Performance and Player Experience. In *2021 IEEE Conference on Games (CoG)*. 1–8. https://doi.org/10.1109/CoG52621.2021.9619105

[25] Danny Oude Bos et al. 2006. EEG-based emotion recognition. *The influence of visual and auditory stimuli* 56, 3 (2006), 1–17.

[26] Patrícia J. Bota, Chen Wang, Ana L. N. Fred, and Hugo Plácido Da Silva. 2019. A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals. *IEEE Access* 7 (2019), 140990–141020. https://doi.org/10.1109/ACCESS.2019.2944001

[27] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

[28] Margaret M Bradley and Peter J Lang. 2000. Measuring emotion: Behavior, feeling, and physiology. (2000).

[29] Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology* 45, 4 (2008), 602–607.

[30] Jason J Braithwaite, Derrick G Watson, Robert Jones, and Mickey Rowe. 2013. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology* 49, 1 (2013), 1017–1034.

[31] Fenja T Bruns and Frank Wallhoff. 2022. Design of a Personalized Affective Exergame to Increase Motivation in the Elderly.. In *HEALTHINF*. 657–663.

[32] Christian Burgers, Allison Eden, Mélisande D van Engelenburg, and Sander Buningh. 2015. How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior* 48 (2015), 94–103.

[33] John T Cacioppo, Richard E Petty, Mary E Losch, and Hai Sook Kim. 1986. Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of personality and social psychology* 50, 2 (1986), 260.

[34] Rafael A Calvo and Sunghwan Mac Kim. 2013. Emotions in text: dimensional and categorical models. *Computational Intelligence* 29, 3 (2013), 527–543.

[35] Marc Cavazza, David Pizzi, Fred Charles, Thurid Vogt, and Elisabeth André. 2009. Emotional Input for Character-Based Interactive Storytelling. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1* (Budapest, Hungary) *(AAMAS '09)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 313–320.

[36] C Richard Chapman, Shunichi Oka, David H Bradshaw, Robert C Jacobson, and Gary W Donaldson. 1999. Phasic pupil dilation response to noxious stimulation in normal volunteers: relationship to brain evoked potentials and pain report. *Psychophysiology* 36, 1 (1999), 44–52.

[37] Hao Chen, Arindam Dey, Mark Billinghurst, and Robert W Lindeman. 2017. Exploring pupil dilation in emotional virtual reality environments. (2017).

[38] Jong-Bae Choi, Suzi Hong, Richard Nelesen, Wayne A Bardwell, Loki Natarajan, Christian Schubert, and Joel E Dimsdale. 2006. Age and ethnicity differences in short-term heart-rate variability. *Psychosomatic medicine* 68, 3 (2006), 421–426.

[39] Avital Cnaan, Nan M Laird, and Peter Slasor. 1997. Using the General Linear Mixed Model to Analyse Unbalanced Repeated Measures and Longitudinal Data. *Statistics in Medicine* 16, 20 (1997), 2349–2380.

[40] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. 2006. Color harmonization. In *ACM SIGGRAPH 2006 Papers*. 624–630.

[41] Gloria Cosoli, Angelica Poli, Lorenzo Scalise, and Susanna Spinsante. 2021. Heart rate variability analysis with wearable devices: Influence of artifact correction method on classification accuracy for emotion recognition. In *2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 1–6.

[42] Francois Cottin, Claire Médigue, Pierre-Marie Leprêtre, Yves Papelier, Jean-Pierre Koralsztein, and Véronique Billat. 2004. Heart rate variability during exercise performed below and above ventilatory threshold. *Medicine & Science in Sports & Exercise* 36, 4 (2004), 594–600.

[43] Cora L Craig, Alison L Marshall, Michael Sjöström, Adrian E Bauman, Michael L Booth, Barbara E Ainsworth, Michael Pratt, ULF Ekelund, Agneta Yngve, James F Sallis, et al. 2003. International physical activity questionnaire: 12-country reliability and validity. *Medicine & science in sports & exercise* 35, 8 (2003), 1381–1395.

[44] Mihaly Csikszentmihalyi. 1975. Play and intrinsic rewards. *Journal of Humanistic Psychology* (1975).

[45] Mihaly Csikszentmihalyi. 2014. *Flow and the foundations of positive psychology*. Springer.

[46] Mihaly Csikszentmihalyi and Mihaly Csikszentmihalyi. 2014. Toward a psychology of optimal experience. *Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi* (2014), 209–226.

[47] Mihaly Csikszentmihalyi, Mihaly Csikszentmihalyi, and Reed Larson. 2014. Validity and reliability of the experience-sampling method. *Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi* (2014), 35–54.

[48] Muhammad Najam Dar, Amna Rahim, Muhammad Usman Akram, Sajid Gul Khawaja, and Aqsa Rahim. 2022. YAAD: Young Adult's Affective Data Using Wearable ECG and GSR sensors. In *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)*. 1–7. https://doi.org/10.1109/ICoDT255437.2022.9787465

[49] Michael E Dawson, Anne M Schell, and Diane L Filion. 2007. The electrodermal system. *Handbook of psychophysiology* 2 (2007), 200–223.

[50] Januka Dharmapriya, Lahiru Dayarathne, Tikiri Diasena, Shiromi Arunathilake, Nihal Kodikara, and Primal Wijesekera. 2021. Music Emotion Visualization through Colour. In *2021 International Conference on Electronics, Information, and Communication (ICEIC)*. 1–6. https://doi.org/10.1109/ICEIC51217.2021.9369788

[51] Adamantios Diamantopoulos, Marko Sarstedt, Christoph Fuchs, Petra Wilczynski, and Sebastian Kaiser. 2012. Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective. *Journal of the Academy of Marketing Science* 40 (2012), 434–449.

[52] Esther Eijlers, Ale Smidts, and Maarten AS Boksem. 2019. Implicit measurement of emotional experience and its dynamics. *PLoS One* 14, 2 (2019), e0211496.

[53] Panteleimon Ekkekakis, Eric E Hall, and Steven J Petruzzello. 2008. The relationship between exercise intensity and affective responses demystified: to crack the 40-year-old nut, replace the 40-year-old nutcracker! *Annals of Behavioral Medicine* 35, 2 (2008), 136–149.

[54] Paul Ekman, Wallace V Freisen, and Sonia Ancoli. 1980. Facial signs of emotional experience. *Journal of personality and social psychology* 39, 6 (1980), 1125.

[55] Hendrik Enders, Filomeno Cortese, Christian Maurer, Jennifer Baltich, Andrea B Protzner, and Benno M Nigg. 2016. Changes in cortical activity measured with EEG during a high-intensity cycling exercise. *Journal of neurophysiology* 115, 1 (2016), 379–388.

[56] Stefan Engeser. 2012. *Advances in flow research*. Springer.

[57] R Frank Falk and Nancy B Miller. 1992. *A primer for soft modeling*. University of Akron Press.

[58] Irving Fatt and Barry A Weissman. 2013. *Physiology of the eye: an introduction to the vegetative functions*. Butterworth-Heinemann.

[59] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

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

[61] Samantha Finkelstein, Andrea Nickel, Tiffany Barnes, and Evan A. Suma. 2010. Astrojumper: Motivating Children with Autism to Exercise Using a VR Game. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI EA '10)*. Association for Computing Machinery, New York, NY, USA, 4189–4194. https://doi.org/10.1145/1753846.1754124

[62] Ronald Fischer. 2004. Standardization to account for cross-cultural response bias: A classification of score adjustment procedures and review of research in JCCP. *Journal of Cross-Cultural Psychology* 35, 3 (2004), 263–282.

[63] Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures. 2012. Publication recommendations for electrodermal measurements. *Psychophysiology* 49, 8 (2012), 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.2012.01384.x

[64] Thomas W Frazier, Milton E Strauss, and Stuart R Steinhauer. 2004. Respiratory sinus arrhythmia as an index of emotional response in young adults. *Psychophysiology* 41, 1 (2004), 75–83.

[65] James H Geer. 1966. Fear and autonomic arousal. *Journal of Abnormal Psychology* 71, 4 (1966), 253.

[66] Erik Geslin, Laurent Jégou, and Danny Beaudoin. 2016. How color properties can be used to elicit emotions in video games. *International Journal of Computer Games Technology* 2016 (2016), 1–1.

[67] M Gleeson. 1998. Temperature regulation during exercise. *International Journal of Sports Medicine* 19, S 2 (1998), S96–S99.

[68] Stefan Göbel, Sandro Hardy, Viktor Wendel, Florian Mehm, and Ralf Steinmetz. 2010. Serious Games for Health: Personalized Exergames. In *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. Association for Computing Machinery, New York, NY, USA, 1663–1666. https://doi.org/10.1145/1873951.1874316

[69] JF Golding. 1992. Phasic skin conductance activity and motion sickness. *Aviation, space, and environmental medicine* 63, 3 (March 1992), 165—171. http://europepmc.org/abstract/MED/1567315

[70] Yulia Golland, Adam Hakim, Tali Aloni, Stacey Schaefer, and Nava Levit-Binnun. 2018. Affect dynamics of facial EMG during continuous emotional experiences. *Biological Psychology* 139 (2018), 47–58. https://doi.org/10.1016/j.biopsycho.2018.10.003

[71] Joshua J Gooley, Ivan Ho Mien, Melissa A St Hilaire, Sing-Chen Yeo, Eric Chern-Pin Chua, Eliza Van Reen, Catherine J Hanley, Joseph T Hull, Charles A Czeisler, and Steven W Lockley. 2012. Melanopsin and rod–cone photoreceptors play different roles in mediating pupillary light responses during exposure to continuous light in humans. *Journal of Neuroscience* 32, 41 (2012), 14242–14253.

[72] Atefeh Goshvarpour, Ataollah Abbasi, and Ateke Goshvarpour. 2017. An accurate emotion recognition system using ECG and GSR signals and matching pursuit method. *Biomedical journal* 40, 6 (2017), 355–368.

[73] Atefeh Goshvarpour, Ataollah Abbasi, and Ateke Goshvarpour. 2017. Fusion of heart rate variability and pulse rate variability for emotion recognition using lagged poincare plots. *Australasian physical & engineering sciences in medicine* 40 (2017), 617–629.

[74] Atefeh Goshvarpour, Ataollah Abbasi, Ateke Goshvarpour, and Sabalan Daneshvar. 2017. Discrimination between different emotional states based on the chaotic behavior of galvanic skin responses. *Signal, Image and Video Processing* 11 (2017), 1347–1355.

[75] Hatice Gunes and Björn Schuller. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31, 2 (2013), 120–136. https://doi.org/10.1016/j.imavis.2012.06.016 Affect Analysis In Continuous Input.

[76] Han-Wen Guo, Yu-Shun Huang, Chien-Hung Lin, Jen-Chien Chien, Koichi Haraikawa, and Jiann-Shing Shieh. 2016. Heart rate variability signal features for emotion recognition by using principal component analysis and support vectors machine. In *2016 IEEE 16th international conference on bioinformatics and bioengineering (BIBE)*. IEEE, 274–277.

[77] Andreas Haag, Silke Goronzy, Peter Schaich, and Jason Williams. 2004. Emotion Recognition Using Bio-sensors: First Steps towards an Automatic System. In *Affective Dialogue Systems*, Elisabeth André, Laila Dybkjær, Wolfgang Minker, and Paul Heisterkamp (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 36–48.

[78] Charles J Hardy and W Jack Rejeski. 1989. Not what, but how one feels: the measurement of affect during exercise. *Journal of sport and exercise psychology* 11, 3 (1989), 304–317.

[79] Michael R Harwell and Guido G Gatti. 2001. Rescaling ordinal data to interval data in educational research. *Review of Educational Research* 71, 1 (2001), 105–131.

[80] Jennifer Healey. 2011. Recording affect in the field: Towards methods and metrics for improving ground truth labels. In *Affective Computing and Intelligent Interaction: 4th International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part I 4*. Springer, 107–116.

[81] Sungpyo Hong, Joanna Narkiewicz, and Randy H Kardon. 2001. Comparison of pupil perimetry and visual perimetry in normal eyes: decibel sensitivity and variability. *Investigative ophthalmology & visual science* 42, 5 (2001), 957–965.

[82] Md-Billal Hossain, Youngsun Kong, Hugo F Posada-Quintero, and Ki H Chon. 2022. Comparison of electrodermal activity from multiple body locations based on standard EDA indices' quality and robustness against motion artifact. *Sensors* 22, 9 (2022), 3177.

[83] Torsten Hothorn, Kurt Hornik, Mark A Van De Wiel, and Achim Zeileis. 2008. Implementing a class of permutation tests: the coin package. *Journal of statistical software* 28, 8 (2008), 1–23.

[84] Christos Ioannou, Patrick Archard, Eamonn O'Neill, and Christof Lutteroth. 2019. Virtual Performance Augmentation in an Immersive Jump & Run Exergame. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300388

[85] Susan A Jackson and Herbert W Marsh. 1996. Development and validation of a scale to measure optimal experience: The Flow State Scale. *Journal of sport and exercise psychology* 18, 1 (1996), 17–35.

[86] Akriti Jaiswal, A Krishnama Raju, and Suman Deb. 2020. Facial emotion detection using deep learning. In *2020 international conference for emerging technology (INCET)*. IEEE, 1–5.

[87] Eun-Hye Jang, Byoung-Jun Park, Mi-Sook Park, Sang-Hyeob Kim, and Jin-Hun Sohn. 2015. Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. *Journal of physiological anthropology* 34, 1 (2015), 1–12.

[88] Robert Jenke, Angelika Peer, and Martin Buss. 2014. Feature extraction and selection for emotion recognition from EEG. *IEEE Transactions on Affective computing* 5, 3 (2014), 327–339.

[89] S Jerritta, M Murugappan, R Nagarajan, and Khairunizam Wan. 2011. Physiological signals based human emotion recognition: a review. In *2011 IEEE 7th international colloquium on signal processing and its applications*. IEEE, 410–415.

[90] Keita Kamijo, Yoshiaki Nishihira, Arihiro Hatta, Takeshi Kaneda, Tetsuo Kida, Takuro Higashiura, and Kazuo Kuroiwa. 2004. Changes in arousal level by differential exercise intensity. *Clinical Neurophysiology* 115, 12 (2004), 2693–2698.

[91] Marcus Karlsson, Rolf Hörnsten, Annika Rydberg, and Urban Wiklund. 2012. Automatic filtering of outliers in RR intervals before analysis of heart rate variability in Holter recordings: a comparison with carefully edited data. *Biomedical engineering online* 11 (2012), 1–12.

[92] Khaled Kassem, Jailan Salah, Yasmeen Abdrabou, Mahesty Morsy, Reem El-Gendy, Yomna Abdelrahman, and Slim Abdennadher. 2017. DiVA: exploring the usage of pupil diameter to elicit valence and arousal. In *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia*. 273–278.

[93] Kathi J Kemper, Craig Hamilton, and Mike Atkinson. 2007. Heart rate variability: impact of differences in outlier identification and management strategies on common measures in three clinical populations. *Pediatric research* 62, 3 (2007), 337–342.

[94] Joni Kettunen, Niklas Ravaja, Petri Näätänen, and Liisa Keltikangas-Järvinen. 2000. The relationship of respiratory sinus arrhythmia to the co-activation of autonomic and facial responses during the Rorschach test. *Psychophysiology* 37, 2 (2000), 242–250. https://doi.org/10.1111/1469-8986.3720242 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/1469-8986.3720242

[95] Stéphanie Khalfa, Peretz Isabelle, Blondin Jean-Pierre, and Robert Manon. 2002. Event-related skin conductance responses to musical emotions in humans. *Neuroscience letters* 328, 2 (2002), 145–149.

[96] Stéphanie Khalfa, Peretz Isabelle, Blondin Jean-Pierre, and Robert Manon. 2002. Event-related skin conductance responses to musical emotions in humans. *Neuroscience Letters* 328, 2 (2002), 145–149. https://doi.org/10.1016/S0304-3940(02)00462-7

[97] Jonghwa Kim and Elisabeth André. 2008. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 12 (2008), 2067–2083. https://doi.org/10.1109/TPAMI.2008.26

[98] Kyung Hwan Kim, Seok Won Bang, and Sang Ryong Kim. 2004. Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing* 42 (2004), 419–427.

[99] Ana Carolina Tomé Klock, Isabela Gasparini, Marcelo Soares Pimenta, and Juho Hamari. 2020. Tailored gamification: A review of literature. *International Journal of Human-Computer Studies* 144 (2020), 102495.

[100] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing* 3, 1 (2012), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

[101] Julian Koenig and Julian F. Thayer. 2016. Sex differences in healthy human heart rate variability: A meta-analysis. *Neuroscience & Biobehavioral Reviews* 64 (2016), 288–310. https://doi.org/10.1016/j.neubiorev.2016.03.007

[102] Marcin Kozak and H-P Piepho. 2018. What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. *Journal of agronomy and crop science* 204, 1 (2018), 86–98.

[103] Sylvia D Kreibig, Frank H Wilhelm, Walton T Roth, and James J Gross. 2007. Cardiovascular, electrodermal, and respiratory response patterns to fear-and sadness-inducing films. *Psychophysiology* 44, 5 (2007), 787–806.

[104] Lars Kuchinke, Melissa L-H Võ, Markus Hofmann, and Arthur M Jacobs. 2007. Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology* 65, 2 (2007), 132–140.

[105] Dana Kulic and Elizabeth A. Croft. 2007. Affective State Estimation for Human–Robot Interaction. *IEEE Transactions on Robotics* 23, 5 (2007), 991–1000. https://doi.org/10.1109/TRO.2007.904899

[106] Kyung-Ah Kwon, Rebecca J Shipley, Mohan Edirisinghe, Daniel G Ezra, Geoff Rose, Serena M Best, and Ruth E Cameron. 2013. High-speed camera characterization of voluntary eye blinking kinematics. *Journal of the Royal Society Interface* 10, 85 (2013), 20130227.

[107] Kate Lambourne and Phillip Tomporowski. 2010. The effect of exercise-induced arousal on cognitive task performance: a meta-regression analysis. *Brain research* 1341 (2010), 12–24.

[108] PETER J. LANG, MARK K. GREENWALD, MARGARET M. BRADLEY, and ALFONS O. HAMM. 1993. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology* 30, 3 (1993), 261–273. https://doi.org/10.1111/j.1469-8986.1993.tb03352.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.1993.tb03352.x

[109] Jeff T Larsen, Catherine J Norris, and John T Cacioppo. 2003. Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology* 40, 5 (2003), 776–785.

[110] Howard B Lee. 2008. Using the Chow test to analyze regression discontinuities. *Tutorials in Quantitative Methods for Psychology* 4, 2 (2008), 46–50.

[111] Paul H Lee, Duncan J Macfarlane, Tai Hing Lam, and Sunita M Stewart. 2011. Validity of the international physical activity questionnaire short form (IPAQ-SF): A systematic review. *International journal of behavioral nutrition and physical activity* 8, 1 (2011), 1–11.

[112] Dominik Leiner, Andreas Fahr, and Hannah Früh. 2012. EDA Positive Change: A Simple Algorithm for Electrodermal Activity to Measure General Audience Arousal During Media Exposure. *Communication Methods and Measures* 6 (12 2012), 237–250. https://doi.org/10.1080/19312458.2012.732627

[113] Laura Leuchs, Max Schneider, Michael Czisch, and Victor I Spoormaker. 2017. Neural correlates of pupil dilation during human fear learning. *Neuroimage* 147 (2017), 186–197.

[114] Ying Liu, Guangyuan Liu, Dongtao Wei, Qiang Li, Guangjie Yuan, Shifu Wu, Gaoyuan Wang, and Xingcong Zhao. 2018. Effects of musical tempo on musicians' and non-musicians' emotional experience when listening to music. *Frontiers in Psychology* 9 (2018), 2118.

[115] Elizabeth J Lyons. 2015. Cultivating engagement and enjoyment in exergames using feedback, challenge, and rewards. *Games for health journal* 4, 1 (2015), 12–18.

[116] Antonio Maffei and Alessandro Angrilli. 2019. Spontaneous blink rate as an index of attention and emotion during film clips viewing. *Physiology & Behavior* 204 (2019), 256–263.

[117] Anna Lisa Martin-Niedecken and Ulrich Götz. 2017. Go with the dual flow: evaluating the psychophysiological adaptive fitness game environment "Plunder Planet". In *Serious Games: Third Joint International Conference, JCSG 2017, Valencia, Spain, November 23-24, 2017, Proceedings 3*. Springer, 32–43.

[118] Anna Lisa Martin-Niedecken, Katja Rogers, Laia Turmo Vidal, Elisa D Mekler, and Elena Márquez Segura. 2019. 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*. 1–15.

[119] Edward McAuley, Terry Duncan, and Vance V Tammen. 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport* 60, 1 (1989), 48–58.

[120] Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology* 14 (1996), 261–292.

[121] Julia Moeller. 2015. A word on standardization in longitudinal studies: don't. , 1389 pages.

[122] Javier Monedero, Elizabeth J Lyons, and Donal J O'Gorman. 2015. Interactive video game cycling leads to higher energy expenditure and is more enjoyable than conventional exercise in adults. *PloS one* 10, 3 (2015), e0118470.

[123] Larissa Müller, Sebastian Zagaria, Arne Bernin, Abbes Amira, Naeem Ramzan, Christos Grecos, and Florian Vogt. 2015. Emotionbike: a study of provoking emotions in cycling exergames. In *Entertainment Computing-ICEC 2015: 14th International Conference, ICEC 2015, Trondheim, Norway, September 29-Ocotober 2, 2015, Proceedings 14*. Springer, 155–168.

[124] Larissa Müller, Sebastian Zagaria, Arne Bernin, Abbes Amira, Naeem Ramzan, Christos Grecos, and Florian Vogt. 2015. EmotionBike: A Study of Provoking Emotions in Cycling Exergames. In *Entertainment Computing - ICEC 2015*, Konstantinos Chorianopoulos, Monica Divitini, Jannicke Baalsrud Hauge, Letizia Jaccheri, and Rainer Malaka (Eds.). Springer International Publishing, Cham, 155–168.

[125] Nijika Murokawa and Minoru Nakayama. 2021. Pupil responses by level of valence sensitivity to emotion-evoking pictures. In *2021 25th International Conference Information Visualisation (IV)*. IEEE, 143–147.

[126] Larissa Müller, Arne Bernin, Sobin Ghose, Wojtek Gozdzielewski, Qi Wang, Christos Grecos, Kai von Luck, and Florian Vogt. 2016. Physiological data analysis for an emotional provoking exergame. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. 1–8. https://doi.org/10.1109/SSCI.2016.7850042

[127] Larissa Müller, Arne Bernin, Andreas Kamenz, Sobin Ghose, Kai von Luck, Christos Grecos, Qi Wang, and Florian Vogt. 2017. Emotional journey for an emotion provoking cycling exergame. In *2017 IEEE 4th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. 104–108. https://doi.org/10.1109/ISCMI.2017.8279607

[128] Lennart E Nacke and Craig A Lindley. 2010. Affective ludology, flow and immersion in a first-person shooter: Measurement of player experience. *arXiv preprint arXiv:1004.0248* (2010).

[129] Lennart E Nacke, Sophie Stellmach, and Craig A Lindley. 2011. Electroencephalographic assessment of player experience: A pilot study in affective ludology. *Simulation & Gaming* 42, 5 (2011), 632–655.

[130] Mimma Nardelli, Gaetano Valenza, Alberto Greco, Antonio Lanata, and Enzo Pasquale Scilingo. 2015. Recognizing emotions induced by affective sounds through heart rate variability. *IEEE Transactions on Affective Computing* 6, 4 (2015), 385–394.

[131] Patrick Ng and Keith Nesbitt. 2013. Informative Sound Design in Video Games. In *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death* (Melbourne, Australia) *(IE '13)*. Association for Computing Machinery, New York, NY, USA, Article 9, 9 pages. https://doi.org/10.1145/2513002.2513015

[132] Sachiyo Ozawa, Hiromasa Yoshimoto, Kazuo Okanoya, and Kazuo Hiraki. 2020. Pupil constrictions and their associations with increased negative affect during responses to recalled memories of interpersonal stress. *Journal of Psychophysiology* (2020).

[133] Timo Partala, Maria Jokiniemi, and Veikko Surakka. 2000. Pupillary responses to emotionally provocative stimuli. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. 123–129.

[134] Timo Partala and Veikko Surakka. 2003. Pupil size variation as an indication of affective processing. *International journal of human-computer studies* 59, 1-2 (2003), 185–198.

[135] Cornelia A Pauls and Gerhard Stemmler. 2003. Repressive and defensive coping during fear and anger. *Emotion* 3, 3 (2003), 284.

[136] Marco Pedrotti, Mohammad Ali Mirzaei, Adrien Tedesco, Jean-Rémy Chardonnet, Frédéric Mérienne, Simone Benedetto, and Thierry Baccino. 2014. Automatic stress classification with pupil diameter analysis. *International Journal of Human-Computer Interaction* 30, 3 (2014), 220–236.

[137] Wei Peng, Jih-Hsuan Lin, and Julia Crouse. 2011. Is playing exergames really exercising? A meta-analysis of energy expenditure in active video games. *Cyberpsychology, Behavior, and Social Networking* 14, 11 (2011), 681–688.

[138] James L Peugh. 2010. A Practical Guide to Multilevel Modeling. *Journal of School Psychology* 48, 1 (2010), 85–112.

[139] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[140] Aurélien P Pichon, Claire de Bisschop, Manuel Roulaud, André Denjean, and Yves Papelier. 2004. Spectral analysis of heart rate variability during exercise in trained subjects. *Medicine & Science in sports & exercise* 36, 10 (2004), 1702–1708.

[141] Jacek Polechoński, Małgorzata Dębska, and Paweł G Dębski. 2019. Exergaming can be a health-related aerobic physical activity. *BioMed Research International* 2019 (2019).

[142] Hugo F Posada-Quintero and Ki H Chon. 2020. Innovations in electrodermal activity data collection and signal processing: A systematic review. *Sensors* 20, 2 (2020), 479.

[143] Hugo F Posada-Quintero, Natasa Reljin, Craig Mills, Ian Mills, John P Florian, Jaci L VanHeest, and Ki H Chon. 2018. Time-varying analysis of electrodermal activity during exercise. *PloS one* 13, 6 (2018), e0198328.

[144] Dominic Potts, Zoe Broad, Tarini Sehgal, Joseph Hartley, Eamonn O'Neill, Crescent Jicol, Christopher Clarke, and Christof Lutteroth. 2024. *EmoSense SDK*. REVEAL, University of Bath. https://github.com/RevealBath/EmoSense

[145] Dominic Potts, Joseph Hartley, Crescent Jicol, Christopher Clarke, and Christof Lutteroth. 2024. *Dataset for "Sweating The Details: Emotion Recognition and the Influence of Physical Exertion in Virtual Reality Exergaming" and EmoSense SDK*. https://doi.org/10.15125/BATH-01372

[146] Pallavi Raiturkar, Andrea Kleinsmith, Andreas Keil, Arunava Banerjee, and Eakta Jain. 2016. Decoupling light reflex from pupillary dilation to measure emotional arousal in videos. In *Proceedings of the ACM Symposium on Applied Perception*. 89–96.

[147] Ryan E. Rhodes, Darren E.R. Warburton, and Shannon S.D. Bredin. 2009. Predicting the effect of interactive video bikes on exercise adherence: An efficacy trial. *Psychology, Health & Medicine* 14, 6 (2009), 631–640. https://doi.org/10.1080/13548500903281088 arXiv:https://doi.org/10.1080/13548500903281088 PMID: 20183536.

[148] Rafaela Larsen Ribeiro, Flávia Teixeira-Silva, Sabine Pompéia, and Orlando Francisco Amodeo Bueno. 2007. IAPS includes photographs that elicit low-arousal physiological responses in healthy volunteers. *Physiology & behavior* 91, 5 (2007), 671–675.

[149] G. Rigas, C. D. Katsis, G. Ganiatsas, and D. I. Fotiadis. 2007. A User Independent, Biosignal Based, Emotion Recognition Method. In *User Modeling 2007*, Cristina Conati, Kathleen McCoy, and Georgios Paliouras (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 314–318.

[150] Raquel Robinson, Katelyn Wiley, Amir Rezaeivahdati, Madison Klarkowski, and Regan L. Mandryk. 2020. "Let's Get Physiological, Physiological!": A Systematic Review of Affective Gaming. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Virtual Event, Canada) *(CHI PLAY '20)*. Association for Computing Machinery, New York, NY, USA, 132–147. https://doi.org/10.1145/3410404.3414227

[151] Nadinne Roman, Cozmin Baseanu, Vlad Ionut Tuchel, Cristina Nicolau, Angela Repanovici, Adina Manaila, Diana Minzatanu, and Roxana Steliana Miclaus. 2023. The Benefits of Combining Mixed Virtual Reality Exergaming with Occupational Therapy for Upper Extremity Dexterity. *Electronics* 12, 6 (2023). https://doi.org/10.3390/electronics12061431

[152] Jennifer Romano Bergstrom, Sabrina Duda, David Hawkins, and Mike McGill. 2014. 4 - Physiological Response Measurements. In *Eye Tracking in User Experience Design*, Jennifer Romano Bergstrom and Andrew Jonathan Schall (Eds.). Morgan Kaufmann, Boston, 81–108. https://doi.org/10.1016/B978-0-12-408138-3.00004-2

[153] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[154] James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 1 (2003), 145.

[155] Valorie N Salimpoor, Mitchel Benovoy, Gregory Longo, Jeremy R Cooperstock, and Robert J Zatorre. 2009. The rewarding aspects of music listening are related to degree of emotional arousal. *PloS one* 4, 10 (2009), e7487.

[156] Marko Sarstedt and Petra Wilczynski. 2009. More for less? A comparison of single-item and multi-item measures. *Die Betriebswirtschaft* 69, 2 (2009), 211.

[157] Kenzo Sato and Richard L Dobson. 1970. Regional and individual variations in the function of the human eccrine sweat gland. *Journal of Investigative Dermatology* 54, 6 (1970), 443–449.

[158] K Sato and F Sato. 1983. Individual variations in structure and function of human eccrine sweat gland. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 245, 2 (1983), R203–R208.

[159] Wataru Sato, Takanori Kochiyama, and Sakiko Yoshikawa. 2020. Physiological correlates of subjective emotional valence and arousal dynamics while viewing films. *Biological Psychology* 157 (2020), 107974.

[160] Shekhar Saxena, M Van Ommeren, KC Tang, and TP Armstrong. 2005. Mental health benefits of physical activity. *Journal of Mental Health* 14, 5 (2005), 445–451.

[161] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and emotion* 24, 7 (2010), 1153–1172.

[162] Marcelle Schaffarczyk, Bruce Rogers, Rüdiger Reer, and Thomas Gronwald. 2022. Validity of the polar H10 sensor for heart rate variability analysis during resting state and incremental exercise in recreational men and women. *Sensors* 22, 17 (2022), 6536.

[163] Karen L Schmidt, Zara Ambadar, Jeffrey F Cohn, and L Ian Reed. 2006. Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling. *Journal of nonverbal behavior* 30 (2006), 37–52.

[164] Simon Schröder, Ekaterina Chashchina, Edgar Janunts, Alan Cayless, and Achim Langenbucher. 2018. Reproducibility and normal values of static pupil diameters. *European Journal of Ophthalmology* 28, 2 (2018), 150–156. https://doi.org/10.5301/ejo.5001027 arXiv:https://doi.org/10.5301/ejo.5001027 PMID: 28885673.

[165] Katie Seaborn and Deborah I Fels. 2015. Gamification in theory and action: A survey. *International Journal of human-computer studies* 74 (2015), 14–31.

[166] Fred Shaffer and Jay P Ginsberg. 2017. An overview of heart rate variability metrics and norms. *Frontiers in public health* (2017), 258.

[167] Hongyu Shi, Licai Yang, Lulu Zhao, Zhonghua Su, Xueqin Mao, Li Zhang, and Chengyu Liu. 2017. Differences of heart rate variability between happiness and sadness emotion states: a pilot study. *Journal of Medical and Biological Engineering* 37 (2017), 527–539.

[168] Daniel Shookster, Bryndan Lindsey, Nelson Cortes, and Joel R Martin. 2020. Accuracy of commonly used age-predicted maximal heart rate equations. *International journal of exercise science* 13, 7 (2020), 1242.

[169] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A review of emotion recognition using physiological signals. *Sensors* 18, 7 (2018), 2074.

[170] Jeff Sinclair, Philip Hingston, and Martin 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*. 289–295.

[171] Lowell Nathaniel B Singson, Maria Trinidad Ursula R Sanchez, and Jocelyn Flores Villaverde. 2021. Emotion recognition using short-term analysis of heart rate variability and ResNet architecture. In *2021 13th International Conference on Computer and Automation Engineering (ICCAE)*. IEEE, 15–18.

[172] Daniel Smilek, Jonathan SA Carriere, and J Allan Cheyne. 2010. Out of mind, out of sight: Eye blinking as indicator and embodiment of mind wandering. *Psychological science* 21, 6 (2010), 786–789.

[173] Robert J Snowden, Katherine R O'Farrell, Daniel Burley, Jonathan T Erichsen, Naomi V Newton, and Nicola S Gray. 2016. The pupil's response to affective pictures: Role of image duration, habituation, and viewing mode. *Psychophysiology* 53, 8 (2016), 1217–1223.

[174] MOHAMMAD SOLEYMANI, GUILLAUME CHANEL, JOEP J. M. KIERKELS, and THIERRY PUN. 2009. AFFECTIVE CHARACTERIZATION OF MOVIE SCENES BASED ON CONTENT ANALYSIS AND PHYSIOLOGICAL CHANGES. *International Journal of Semantic Computing* 03, 02 (2009), 235–254. https://doi.org/10.1142/S1793351X09000744 arXiv:https://doi.org/10.1142/S1793351X09000744

[175] Mohammad Soleymani, Joep J.M. Kierkels, Guillaume Chanel, and Thierry Pun. 2009. A Bayesian framework for video affective representation. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. 1–7. https://doi.org/10.1109/ACII.2009.5349563

[176] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing* 3, 1 (2011), 42–55.

[177] Rukshani Somarathna, Tomasz Bednarz, and Gelareh Mohammadi. 2023. Virtual Reality for Emotion Elicitation – A Review. *IEEE Transactions on Affective Computing* 14, 4 (2023), 2626–2645. https://doi.org/10.1109/TAFFC.2022.3181053

[178] Robert Stojan and Claudia Voelcker-Rehage. 2019. A Systematic Review on the Cognitive Benefits and Neurophysiological Correlates of Exergaming in Healthy Older Adults. *Journal of Clinical Medicine* 8, 5 (2019). https://doi.org/10.3390/jcm8050734

[179] H Storm, K Myre, M Rostrup, O Stokland, MD Lien, and JC Raeder. 2002. Skin conductance correlates with perioperative stress. *Acta anaesthesiologica scandinavica* 46, 7 (2002), 887–895.

[180] Nazmi Sofian Suhaimi, James Mountstephens, Jason Teo, et al. 2020. EEG-based emotion recognition: A state-of-the-art review of current trends and opportunities. *Computational intelligence and neuroscience* 2020 (2020).

[181] Jun-Wen Tan, Adriano O Andrade, Hang Li, Steffen Walter, David Hrabal, Stefanie Rukavina, Kerstin Limbrecht-Ecklundt, Holger Hoffman, and Harald C Traue. 2016. Recognition of intensive valence and arousal affective states via facial electromyographic activity in young and senior adults. *PloS one* 11, 1 (2016), e0146691.

[182] Jun-Wen Tan, Steffen Walter, Andreas Scheck, David Hrabal, Holger Hoffmann, Henrik Kessler, and Harald C Traue. 2012. Repeatability of facial electromyography (EMG) activity over corrugator supercilii and zygomaticus major on differentiating various emotions. *Journal of Ambient Intelligence and Humanized Computing* 3 (2012), 3–10.

[183] Andrew K Tate and Steven J Petruzzello. 1995. Varying the intensity of acute exercise: implications for changes in affect. *The Journal of sports medicine and physical fitness* 35, 4 (1995), 295–302.

[184] Jan-Philipp Tauscher, Fabian Wolf Schottky, Steve Grogorick, Paul Maximilian Bittner, Maryam Mustafa, and Marcus Magnor. 2019. Immersive EEG: Evaluating Electroencephalography in Virtual Reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 1794–1800. https://doi.org/10.1109/VR.2019.8797858

[185] Kuldar Taveter and Eliise Marie Taveter. 2021. Case Study on Using Colours in Constructing Emotions by Interactive Digital Narratives. *arXiv e-prints*, Article arXiv:2104.12154 (April 2021), arXiv:2104.12154 pages. https://doi.org/10.48550/arXiv.2104.12154 arXiv:2104.12154 [cs.HC]

[186] Scott Thomas, Jeff Reading, and Roy J Shephard. 1992. Revision of the physical activity readiness questionnaire (PAR-Q). *Canadian journal of sport sciences= Journal canadien des sciences du sport* 17, 4 (1992), 338–345.

[187] Sinika Timme and Ralf Brand. 2020. Affect and exertion during incremental physical exercise: Examining changes using automated facial action analysis and experiential self-report. *PloS one* 15, 2 (2020), e0228739.

[188] Christian Tronstad, Maryam Amini, Dominik R Bach, and Ørjan G Martinsen. 2022. Current trends and opportunities in the methodology of electrodermal activity measurement. *Physiological measurement* 43, 2 (2022), 02TR01.

[189] Kemal S Türker. 1993. Electromyography: Some Methodological Problems and Issues. *Physical Therapy* 73, 10 (10 1993), 698–710. https://doi.org/10.1093/ptj/73.10.698 arXiv:https://academic.oup.com/ptj/article-pdf/73/10/698/10760206/ptj0698.pdf

[190] Marco C Uchida, Renato Carvalho, Vitor Daniel Tessutti, Reury Frank Pereira Bacurau, Hélio José Coelho-Júnior, Luciane Portas Capelo, Heloiza Prando Ramos, Marcia Calixto dos Santos, Luís Felipe Milano Teixeira, and Paulo Henrique Marchetti. 2018. Identification of muscle fatigue by tracking facial expressions. *PLoS One* 13, 12 (2018), e0208834.

[191] Mariaconsuelo Valentini and Gianfranco Parati. 2009. Variables Influencing Heart Rate. *Progress in Cardiovascular Diseases* 52, 1 (2009), 11–19. https://doi.org/10.1016/j.pcad.2009.05.004 Heart Rate and Cardiovascular Disease.

[192] EH Van Olst, JF Orlebeke, and SD Fokkema. 1967. Skin conductance as a measure of tonic and phasic arousal. *Acta psychologica* 27 (1967), 262.

[193] HTC Vive. 2023. *Eye and Facial Tracking SDK - Developer Resources*. https://developer.vive.com/resources/vive-sense/eye-and-facial-tracking-sdk/

[194] Chin-An Wang, Talia Baird, Jeff Huang, Jonathan D Coutinho, Donald C Brien, and Douglas P Munoz. 2018. Arousal effects on pupil size, heart rate, and skin conductance in an emotional face task. *Frontiers in neurology* 9 (2018), 1029.

[195] Hua Wang, Mark Chignell, and Mitsuru Ishizuka. 2006. Empathic tutoring software agents using real-time eye tracking. In *Proceedings of the 2006 symposium on Eye tracking research & applications*. 73–78.

[196] Youfa Wang and Hsin-Jen Chen. 2012. Use of percentiles and z-scores in anthropometry. In *Handbook of anthropometry: Physical measures of human form in health and disease*. Springer, 29–48.

[197] Darren ER Warburton, Shannon SD Bredin, Leslie TL Horita, Dominik Zbogar, Jessica M Scott, Ben TA Esch, and Ryan E Rhodes. 2007. The health benefits of interactive video game exercise. *Applied Physiology, Nutrition, and Metabolism* 32, 4 (2007), 655–663.

[198] Darren ER Warburton, Crystal Whitney Nicol, and Shannon SD Bredin. 2006. Health benefits of physical activity: the evidence. *Cmaj* 174, 6 (2006), 801–809.

[199] Claudia AF Wascher. 2021. Heart rate as a measure of emotional arousal in evolutionary biology. *Philosophical Transactions of the Royal Society B* 376, 1831 (2021), 20200479.

[200] SV Wass, K De Barbaro, and K Clackson. 2015. Tonic and phasic co-variation of peripheral arousal indices in infants. *Biological Psychology* 111 (2015), 26–39.

[201] Dan Wichterle, Jan Simek, Maria Teresa La Rovere, Peter J Schwartz, A John Camm, and Marek Malik. 2004. Prevalent low-frequency oscillation of heart rate: novel predictor of mortality after myocardial infarction. *Circulation* 110, 10 (2004), 1183–1190.

[202] Urban Wiklund, Rolf Hörnsten, Marcus Karlsson, Ole B Suhr, and Steen M Jensen. 2008. Abnormal heart rate variability and subtle atrial arrhythmia in patients with familial amyloidotic polyneuropathy. *Annals of Noninvasive Electrocardiology* 13, 3 (2008), 249–256.

[203] Glenn D Wilson. 1967. GSR responses to fear-related stimuli. *Perceptual and Motor Skills* 24, 2 (1967), 401–402.

[204] Huiping Wu and Shing-On Leung. 2017. Can Likert scales be treated as interval scales?—A Simulation study. *Journal of social service research* 43, 4 (2017), 527–532.

[205] Saori Yamashita, K Iwai, T Akimoto, J Sugawara, and I Kono. 2006. Effects of music during exercise on RPE, heart rate and the autonomic nervous system. *Journal of sports medicine and physical fitness* 46, 3 (2006), 425.

[206] Jennifer Yih, Harry Sha, Danielle E Beam, Josef Parvizi, and James J Gross. 2019. Reappraising faces: effects on accountability appraisals, self-reported valence, and pupil diameter. *Cognition and Emotion* 33, 5 (2019), 1041–1050.

[207] Peter Zachar and Ralph D Ellis. 2012. *Categorical versus dimensional models of affect: a seminar on the theories of Panksepp and Russell*. Vol. 7. John Benjamins Publishing.

[208] Marcel Zentner, Didier Grandjean, and Klaus R Scherer. 2008. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion* 8, 4 (2008), 494.

[209] Jing Zhang, Ottmar V Lipp, Tian PS Oei, and Renlai Zhou. 2011. The effects of arousal and valence on facial electromyographic asymmetry during blocked picture viewing. *International journal of psychophysiology* 79, 3 (2011), 378–384.

[210] Janis H Zickfeld, Patrícia Arriaga, Sara Vilar Santos, Thomas W Schubert, and Beate Seibt. 2020. Tears of joy, aesthetic chills and heartwarming feelings: Physiological correlates of Kama Muta. *Psychophysiology* 57, 12 (2020), e13662.