RetroSketch: A Retrospective Method for Measuring Emotions and Presence in Virtual Reality

Dominic Potts
dominic.potts@uwe.ac.uk
University of the West of England
Bristol, UK
REVEAL, University of Bath
Bath, UK

Miloni Gada
mg2544@bath.ac.uk
REVEAL, University of Bath
Bath, UK

Aastha Gupta
ag2605@bath.ac.uk
REVEAL, University of Bath
Bath, UK

Kavya Goel
kg674@bath.ac.uk
REVEAL, University of Bath
Bath, UK

Klaus Phillip Krzok
kk2003@bath.ac.uk
REVEAL, University of Bath
Bath, UK

Genieve Pate
genievievepate13@outlook.com
REVEAL, University of Bath
Bath, UK

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

Mark Weston-Arnold
mdwa20@bath.ac.uk
REVEAL, University of Bath
Bath, UK

Jakob Aylott
jma66@bath.ac.uk
REVEAL, University of Bath
Bath, UK

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

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

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

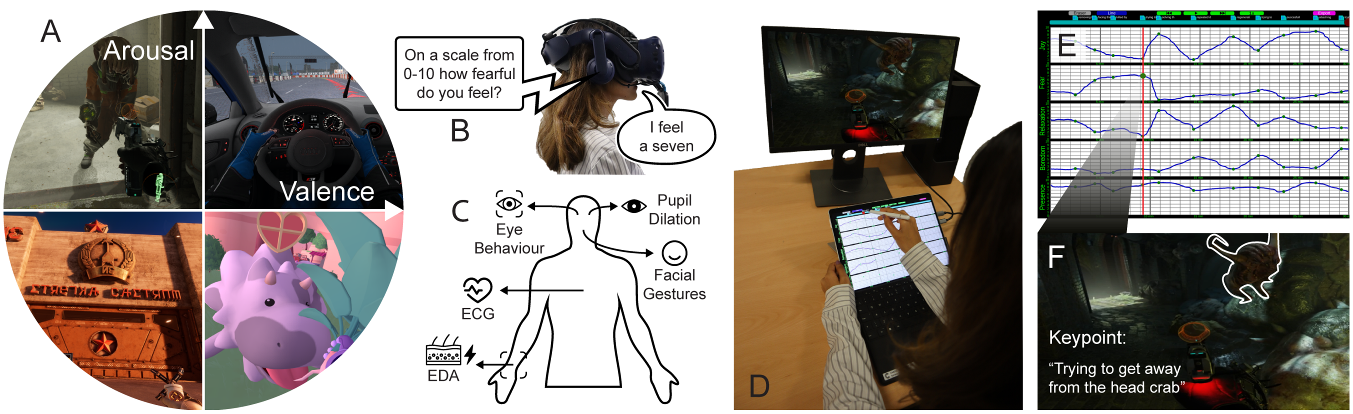


Figure 1: (A) Virtual Reality (VR) experiences are designed to elicit a wide range of emotions, across the dimensions of valence and arousal, as well as a sense of presence. These are commonly measured using (B) Experience Sampling Methodology which can introduce biases and fails to provide continuous measurements, or (C) physiological sensing which is difficult to administer and analyse. We present RetroSketch, a novel measurement method that complements existing methods. (D) Users watch and control a video playback of their VR experience, and provide (E) continuous and temporally aligned measurements for presence and four emotions, as well as (F) keypoints with annotations that offer qualitative insights and provide additional context.

*https://revealcentre.org/

CHI '25, April 26-May 1, 2025, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/2025/04
https://doi.org/10.1145/3706598.3713957

ABSTRACT

Virtual Reality (VR) designers and researchers often need to measure emotions and presence as they evolve over time. The experience sampling method (ESM) is a common way to achieve this, however, ESM disrupts the experience and lacks granularity. We propose RetroSketch, a new method for measuring subjective emotions and presence in VR, where users watch back their VR experience and retrospectively sketch a plot of their feelings. RetroSketch leaves the VR experience undisturbed and yields highly granular data, including information about salient events and qualitative descriptions of their feelings. We compared RetroSketch and ESM in a large study (n=140) using five different VR experiences over one-hour sessions. Our results show that RetroSketch and ESM measures are highly correlated with each other, as well as physiological measures indicative of emotion. The correlations are robust across different VR experiences and user demographics. They also highlight the impact of ESM on users' experience.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Human computer interaction (HCI)**; **HCI theory, concepts and models**.

KEYWORDS

virtual reality, emotion, presence, emotion measurement, presence measurement, emotion appraisal, experience sampling, physiological sensing, physiological correlates, games

ACM Reference Format:

Dominic Potts, Miloni Gada, Aastha Gupta, Kavya Goel, Klaus Phillip Krzok, Genieve Pate, Joseph Hartley, Mark Weston-Arnold, Jakob Aylott, Christopher Clarke, Crescent Jicol, and Christof Lutteroth. 2025. RetroSketch: A Retrospective Method for Measuring Emotions and Presence in Virtual Reality. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3706598.3713957

1 INTRODUCTION

Measuring presence and emotions in Virtual Reality (VR) is important because they fundamentally impact the user experience. Arguably the most important function of VR is to immerse users in virtual environments and invoke a sense of presence – a feeling of *"being there"* [230] that elicits responses *"as if it were real"* [221]. Measuring presence can help researchers understand how it affects VR experiences, e.g. VR skills training is only effective if a sense of presence is invoked [77, 241], and can help developers improve VR experiences, e.g. by identifying breaks in presence due to design flaws [229, 238]. As a result, measuring presence has been a long-standing topic of research [251]. Of similar importance is the ability of VR to elicit emotions. VR experiences often need to be emotionally engaging, e.g. games that are exciting [123] or training simulations that can prepare people for the emotional stress of a real situation [152]. Emotions in VR are related to presence [47, 101] and have also been linked to training effectiveness [49, 257]. By measuring emotions, developers can improve VR experiences, for example by validating and refining experiences to evoke desired emotions or understanding when users are confused or frustrated. This is relevant in research such as when gathering behavioural insights [114, 146, 204], in social and empathy research [100, 161, 208] and research on VR itself [47, 62, 101], as well as for real-world applications such as therapeutic interventions [56, 150], education [4, 49, 257] and entertainment [92, 122]. As both presence and emotions change over time, methods are needed to measure them repeatedly as they evolve.

One of the most established approaches for measuring emotions and presence is the Experience Sampling Methodology (ESM) [119]. ESM provides subjective measures by administering validated questionnaires during, or immediately after, a VR experience. It is easy to administer, ecologically valid, and there are many variations of the methodology [34, 42, 64]. One of the most common ways of administering ESM is during the experience itself by asking the participant how they feel at specific points, e.g. in regular time intervals [46, 264]. However, this type of ESM has several disadvantages. First, it requires participants to report their feelings in the *very moment they are required to experience them*, which can result in an 'observer effect' disrupting the experience and influencing their response [9, 158]. Second, experimenters often ask participants out loud to minimise disruption, which in turn can increase the 'social desirability' bias and the chance participants respond favourably to the experimenter [79, 210]. Finally, ESM can practically only capture a limited number of data points, so important changes and events can easily be missed. Increasing the number of data points exacerbates the previously mentioned issues of disruption, observer effect, and social desirability.

Physiological sensing is another approach that has gained traction in recent years [52, 180, 201]. This is based on the principle that emotions and other internal sensations are associated with automatic bodily responses that can be measured [143]. Physiological sensing is desirable because it can capture changes in emotions continuously, in real-time, and removes the disruptions and biases that ESM suffers from. However, it is difficult to isolate the effects of individual emotions and other sensations in physiological signals, e.g. attempts to measure presence physiologically have not found anything conclusive [76, 251]. In addition, physiological signals are noisy due to factors such as user movement and exertion levels[170, 251] which necessitates non-trivial data cleaning and analysis processes, in addition to careful calibration and setup of the sensors themselves. Finally, due to the complexity of emotions, collecting ground truth data to be able to train physiological sensing models still relies on subjective methods such as ESM.

We propose *RetroSketch*, a new method for measuring emotions and presence in VR experiences. RetroSketch yields continuous subjective data of an experience, complementing existing methods such as physiological sensing and providing an alternative to ESM.

The experience is recorded and users then retrospectively reflect on it by watching back and exploring the video while sketching a graph of their emotions and presence over time (Figure 1D). The VR experience is replayed both visually and audibly from the user's perspective, allowing them to hear themselves and the audio soundscape of the experience. Users sketch and plot emotions and presence over the duration of the experience, resulting in continuous data for each measure (see Figure 1E). Users can also identify and highlight 'keypoints' in their experience such as salient events with emotional consequences. These can be annotated with textual descriptions, providing context and sentiment (Figure 1F).

RetroSketch offers several advantages over existing methods. First, RetroSketch provides high-resolution, continuous data for emotions and presence that are not bound by specific events or time intervals with direct correspondence to a point-of-view video of a VR experience. Second, it does not disrupt the experience and does not suffer from observer bias because it is administered in retrospect. Third, it reduces social desirability bias because data can be entered in private without intervention from the experimenter. However, measuring emotions in retrospect based on recall has been explored previously and can be affected by age and personality traits [149, 253]. It is unclear what effect these factors will have when measuring emotions using RetroSketch. We validate RetroSketch by posing the following research questions:

RQ1 How do RetroSketch measures relate to ESM measures?
RQ2 How reliable is RetroSketch & ESM across different VR experiences and users?
RQ3 How does ESM influence the VR user experience?
RQ4 How do RetroSketch & ESM relate to physiological measures?
RQ5 How do RetroSketch & ESM relate to qualitative measures?

To address these questions, we conducted a user study (n=140), comparing RetroSketch to ESM, as the gold standard for subjective emotion measurement, as well as physiological measures. We measured four emotions (joy, fear, relaxation and boredom) and presence across five popular VR games: *Assetto Corsa Competizione* [218], *Garden of the Sea* [39], *Half-Life Alyx* [254], *I Expect You To Die* [65] and *Red Matter* [189]. Each participant played one of the games over two 30-minute gameplay sessions (one hour total): one session using ESM during the experience and RetroSketch immediately after, and the other session only using RetroSketch. The former allows us to compare the tools directly against each other (RQ1, RQ2), while the latter allows us to understand the influence ESM has on the experience (RQ3). Throughout the experience, we collected ten physiological measures that have been shown to correlate with different emotions (RQ4). Furthermore, we collected qualitative data in the form of RetroSketch annotations of the experience as well as post-experiment questionnaires (RQ5).

Our results show that RetroSketch and ESM measures are highly correlated with each other, however, RetroSketch generally captures a higher variation and range of emotions and presence compared to ESM. Positive emotions (joy and relaxation) tend to be scored lower, and negative emotions (fear and boredom) scored higher in RetroSketch compared with ESM, which may be a result of RetroSketch's ability to reduce social desirability bias (RQ1). These findings are robust across the different VR games and individual characteristics of the participants (RQ2). Furthermore, our study provides evidence that ESM affects the experience in seemingly unpredictable ways across the different games, such as significantly decreasing physical presence in Assetto Corsa Competizione, while increasing it in Half-Life Alyx (RQ3). This suggests that researchers and developers should be mindful when using ESM to compare different experiences. RetroSketch and ESM bear similar relationships to physiological measures, indicating that RetroSketch can be used to collect subjective ground truth data for emotion estimation models (RQ4). While ESM provides comparatively few data points, RetroSketch data is continuous and has a high resolution, making it particularly useful for this purpose. Finally, we found that RetroSketch's temporally anchored and contextualised qualitative annotations are consistent with the quantitative measures reported by participants. The annotations complement the quantitative data and provide extra information for researchers and developers to make sense of the user experience. In summary, we make the following contributions:

(1) RetroSketch, a novel and openly available method for measuring emotions and presence for VR experiences [1].
(2) Empirical evidence that validates RetroSketch against both ESM and physiological measures.
(3) Insights into the impacts of ESM on VR experiences.
(4) A large open dataset of emotions, presence, and ten physiological measures across five VR experiences (n=140), see [172].

2 RELATED WORK

2.1 Models of Emotion

Emotions are internal states associated with feelings, thoughts, behaviours and neurophysiological changes. They can be described using models of varying complexity. Categorical models characterise emotions as fundamental and discrete feelings, such as joy, fear, anger, and sadness, with complex emotions regarded as combinations of these basic feelings [53, 137]. Dimensional models such as Russell's widely-accepted Circumplex Model of Affect (CMA) [157, 193, 194] describe emotions along a few dimensions, e.g. *Valence* (pleasant vs. unpleasant) and *Arousal* (sleepy vs. alert), making it possible to compare different emotions along these dimensions [81, 170].

Barrett's theory of constructed emotion [12], elaborates on dimensional models, describing how bodily feelings are interpreted as a pre-cognitive step based on the context and prior experiences of a person. Similar but distinct to this are appraisal-based models which take into account that emotions are heavily influenced by context, describing emotions as processes derived from our cognitive evaluation or 'appraisal' of events [55, 121, 191, 203, 231]. They explain how "different emotions may emerge from the same event, in different individuals, and on different occasions" [151]. Similar to many ESM studies, RetroSketch uses categorical emotion measures, while allowing users to retrospectively reflect on events and appraise them in the context of the overall experience.

2.2 Emotion Elicitation

Typically, emotion elicitation in VR is highly discretised with VEs designed to target specific emotions [57, 105, 105, 234] and short exposures normally lasting only a few minutes [57, 99, 101, 137]. However, popular VR experiences often take over an hour [1]. More recent work has explored interactive and complex emotional stimuli such as VR games, which often span different levels of valence and arousal that vary as the gameplay unfolds [17, 75, 84, 93, 148, 163, 216, 265]. Taking advantage of this more ecologically valid approach, we evaluated RetroSketch with a cross-section of popular VR games from different genres.

2.3 Subjective Measures of Emotion

Emotions are often measured subjectively by asking participants to rate what they feel using psychometric scales [13, 38, 139], e.g. categorical emotion scales [42], PANAS [261], the Pleasure-Arousal-Dominance scale [145], the Self-Assessment Manikin [25] and the

[1] https://github.com/revealcentre/retrosketch

Affect Slider [18]. All these scales can be used retrospectively or repeated throughout an experience as part of ESM [119].

While retrospective use of psychometric scales minimises 'in-the-moment' disruptions, it relies on accurate recall and may be influenced by recency and primacy effects (i.e. recalling more clearly what was perceived first or last in an experience) [197]. Demographical covariates such as age and personality traits, as well as tiredness, have been shown to influence the recall of emotions [149, 253]. For example, there is evidence for a positivity effect in older adults compared to younger adults [31, 149, 196]. In addition, neuroticism has been shown to result in increased recall of negative emotions, while extraversion increases recall of positive emotions [149, 188]. Moreover, emotions often change and evolve over time [135] so cannot generally be captured by only a few retrospective measurements. RetroSketch aims to reduce the limitations of recall through navigable point-of-view video and audio of the VR experience and we investigate the influence of demographical covariates on RetroSketch measures through our study.

When using psychometric scales during an experience, it can be challenging for participants to gauge and express their emotions 'on the spot' [119, 137]. Furthermore, responses are more likely to be biased by experimenter rapport, participant openness, social desirability and demand characteristics [81, 86, 156]. Even when applied multiple times, psychometric scales are limited in the amount of data they can provide. They cannot collect data continuously and therefore miss key aspects of an experience [119].

Some works tried to address this with continuous emotion measurement tools [67, 138, 192, 206]. These include software interfaces for 2D videos [67], the affect rating dial which involves emotion measurement by continuously rotating a mechanical dial [73, 142, 262], and the emotion slider where users move a physical slider [120]. These interfaces provide highly granular, moment-to-moment emotion measures of a specific emotion measure such as valence [67] or, in the case of Schubert, both valence and arousal captured in a 2D plot [206].

More recently, Xue et al. demonstrated how these techniques can be applied to 360° VR video with valence and arousal recorded continuously and manipulated using a game controller [269, 270]. These methods overcome the data limitations of questionnaires for *non-interactive* media such as videos and music, enabling continuous and granular emotion measures. However, these benefits require continuous input which is challenging during *interactive* experiences such as VR games. A retrospective approach is a promising alternative, and RetroSketch aims to overcome these limitations by allowing users to measure multiple emotions simultaneously, whilst supporting appraisal of the experience through video playback that can be navigated and annotated as the user desires.

Finally, emotions can be captured qualitatively after an experience with methods such as open-ended questionnaires [153], interviews [195] and diaries [41, 165], or during an experience through observation [184] and methods such as 'Think-Aloud' protocols [90]. While they share similar limitations as psychometric scales, e.g. reliance on recall and biases, they can better capture emotional nuances and appraisal due to their open-ended nature. RetroSketch avoids the limitations of 'in the moment' approaches and supports recall with a navigable video walk-through of the experience. It uses scales to collect continuous, high-resolution quantitative data and qualitative annotations to capture nuances, context, and appraisal over time.

2.4 Experience Sampling Method (ESM)

ESM is a methodology designed to measure experience in 'natural' environments and 'in the moment' [119]. It is often used in longitudinal research [22] and relies on participants reporting their thoughts, feelings, and behaviours using quantitative and qualitative measures at designated points in time [264]: In signal-contingent ESM, participants respond when signalled by an experimenter or system (e.g. mobile phone). In event-contingent ESM, participants respond after set events. In interval-contingent ESM, responses happen at set time intervals. All three approaches are affected by observer and social desirability biases (see above) and can miss key moments of an experience. For example, ESM may disrupt a VR experience, create breaks in presence [220, 229] and redirect cognitive resources outside the virtual environment and away from the active elicitation. Nevertheless, ESM is well-established, validated and useful [43] so serves as a 'gold standard' for the comparison to and validation of RetroSketch measures.

2.5 Physiological Measures of Emotion

Emotions can be estimated by analysing unconscious changes in physiological measures associated with the central and autonomic nervous system [29, 137] such as electroencephalography (EEG) [23, 96, 243], eye and facial behaviour (pupilometry, blinks, fixations, and saccades) [190], and cardiovascular dynamics (heart rate, respiration, and electrodermal activity) [11, 82]. Compared to subjective measures, physiological measures are less affected by experimenter bias [81, 86] or recall [197]. Pupil Dilation Level (PDL) and Pupil Dilation Response (PDR) correlate with both arousal [26, 124, 181, 232, 259] and valence [2, 8, 26, 32, 103, 154, 162, 271], as well as categorical emotions such as fear [33, 124, 232]. Heart rate (HR) and HR variability (HRV) correlate with affect [82, 106, 155, 213, 260]. Electrodermal activity (EDA), in particular Skin Conductance Response (SCR) and Skin Conductance Level (SCL) [11, 26, 199], correlate with arousal. Facial gestures such as contractions of the zygomatic major muscles (smiling) correlate with valence [28, 170, 273].

Physiological measures are noisy, especially in VR where users often move naturally [246], requiring non-trivial data cleaning and analysis processes [170] and careful sensor setup and calibration. Furthermore, physiological measures are only indirect markers of emotion that do not clearly map to psychometric scales. Emotion recognition approaches often use machine learning to model the complex relationships between emotions and physiological measures such as EDA [6, 70–72, 97, 106, 107, 113, 186], fMEG [94, 219, 245], HRV [40, 71, 71, 82, 95, 155, 219, 219] and blink information [3, 233]. They are typically 'black boxes' that have been trained on ground truth data collected through subjective methods, with their own biases and uncertainties. As a consequence, we validate RetroSketch by correlating it directly with common physiological measures, using established data cleaning procedures.

2.6 Presence

Presence is arguably the most important quality of VR and a core interest in VR research [251]. Presence is typically defined as the sense of *"being there"* [87, 128, 200, 212, 230, 240, 272], describing the illusion created by VR [220, 221, 223, 225] that leads users to respond to virtual experiences as if "they were real" [200, 220, 222, 237]. Presence and emotions are associated [24, 62, 187, 215], for example, presence has been shown to correlate with emotion intensity [10, 47, 225] as well as with negative valence emotions such as fear [62, 80, 99, 101, 140, 164, 239]. There are different approaches for measuring presence, each with limitations [220].

Presence is most commonly measured using questionnaires [76, 117, 209, 225, 236, 251] such as the WS [266], IPQ [207], SUS [224] and MPS [133], which are typically administered after a VR experience, but can also be administered while still in VR [60, 174, 209]. Presence questionnaires are affected by biases similar to emotion questionnaires [74, 225]. Qualitative approaches include having users write essays about their VR experience [15, 112, 225, 227], which can then be analysed using ML-driven sentiment analysis [126, 160, 225]. We use such ML-driven sentiment analysis to validate RetroSketch's annotations.

In VEs designed to elicit emotions, presence can sometimes be estimated through physiological measures [47] such as ECG and EDA [136, 144, 174, 237], eye movements/pupillometry [116], fEMG [182] or EEG signals [14, 58, 110, 166] because of its association with emotional response [225]. Furthermore, physiological correlates have been explored for breaks in presence (BIPs) [125], such as ECG and EDA [226, 228], blood flow [185] and EEG [111]. We include common physiological correlates of presence in the validation of RetroSketch measures.

3 RETROSKETCH DESIGN

RetroSketch relies on three main design features, which have been developed based on the emotion and presence literature, prior work and iterative pilot testing: (1) unconstrained video and audio playback, (2) continuous quantitative measures, and (3) salient keypoints and qualitative annotations.

Unconstrained Video and Audio Playback: RetroSketch uses video playback of the recorded experience to assist with the recognition of events and to help users recall how they felt. This has been used in related work on emotion measurement [192] and is supported by findings that reported emotions while re-watching a video align with physiological measures from the initial viewing [138]. In contrast to related work, RetroSketch allows users to navigate the video without constraints, in the order and speed they choose. This supports an individual's recall patterns (e.g. linearly or certain salient events first; guided by primacy or recency) with personalised reflection and appraisal of an experience. We also include recorded audio, including the user's voice, to further support recall.

Continuous Quantitative Measures: Users can report continuous levels of presence and four emotions — joy, fear, relaxation, and boredom — each of which is significant for VR experiences and provides comprehensive coverage of the circumplex model [193, 194]. Unlike related work on emotion measurement [192], RetroSketch is designed to capture subjective ratings across multiple measures in a temporally synchronised manner. To achieve this, the graphs of each measure are stacked vertically (Figure 2) ensuring that time points are aligned. Users can freely draw their ratings as line graphs, with the method of input depending on the specific RetroSketch implementation (see below). Each measure uses an 11-point rating scale where 0 represents *"none of that feeling"* and 10 represents *"the most intense version of that feeling possible"*. Supported by theory and simulations [85, 268], this has previously been used for measuring emotions in VR on an interval scale [170, 247].

Salient Keypoints and Qualitative Annotations: To complement continuous quantitative measures of emotions and presence, RetroSketch enables users to specify 'keypoints' that identify particularly memorable or salient events in the experience. Keypoints can be specified before drawing ratings into the graph, e.g. to enable users to draw ratings by "connecting the dots", or they can be specified once ratings have been drawn, e.g. based on peaks and troughs. Keypoints are not provided but specified by the user to reduce 'demand characteristics' [156], i.e. providing keypoints for users may give cues about the aim of a study and may bias the user. In addition to identifying salient events, users can annotate keypoints with rich qualitative insights and additional context as to why they selected them. Depending on the implementation of RetroSketch, annotations can take any form, whether pictorial such as sketches or doodles, or textual such as short descriptions of their feelings or thoughts at the keypoint moments.

To enable flexible and versatile use, we implemented two versions of RetroSketch: a paper-based and a digital version. The paper-based version served as a low-fidelity prototype for the digital version.

3.1 Paper-based Version

The paper-based RetroSketch consists of an A3 sheet of graph paper (Figure 2). Emotion and presence scales are stacked on top of each other, with the Y-axis representing the 0-10 rating scales and the X-axis representing time during the experience. Users can use various drawing tools such as a pencil, ruler and eraser to draw, erase, or correct the lines, keypoints and annotations as they see fit.

Users can free-hand draw, affording them familiarity and flexibility, which makes it more suitable for users who are less accustomed to technology. The only piece of equipment needed in addition to the printed sheet and drawing tools is a device for video playback (e.g. a mobile phone). However, the disadvantage of a paper-based version is that participants' attention is split between the video playback and the tool, and they must manually align their responses with the video playback timestamps. In addition, the measures and annotations may need to be digitised for later analysis.

3.2 Digital Version

We designed and implemented a digital version of RetroSketch (Figure 3) that combines an interactive tablet and stylus, allowing users to sketch emotion measures onto digital graphing paper. Our goal was to retain the flexibility of the paper-based version regarding line drawing and keypoints, while adding several digital features:

Synchronised Timeline Cursor: An interactive timeline cursor allows users to scrub through the video while highlighting the corresponding moment in the graphs below. This integration of the video and emotion graphs helps minimise diverging attention between watching the video and plotting emotions, making it easier

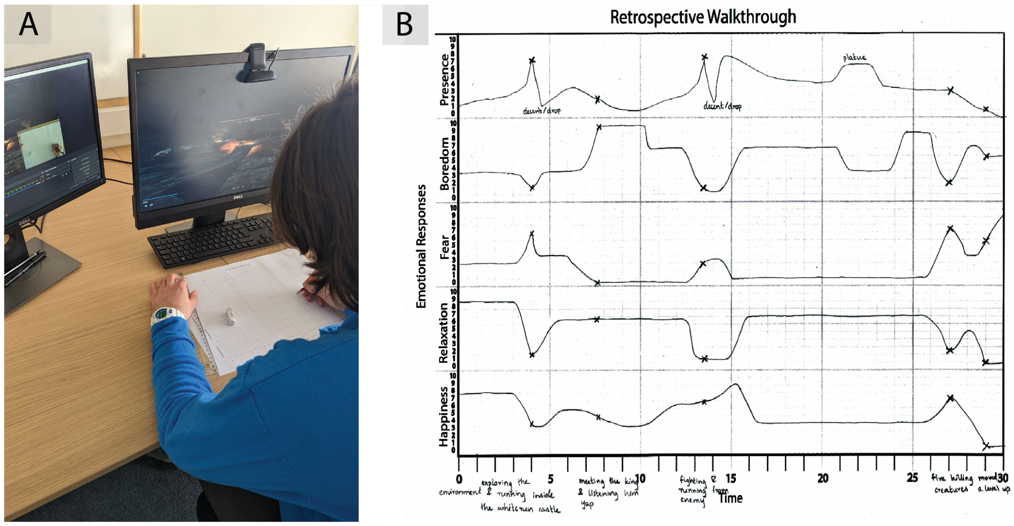


Figure 2: Paper-based RetroSketch. (A) shows a user reviewing a VR gameplay session and sketching their emotions and presence using paper RetroSketch. (B) shows their completed graph with keypoints marked and annotated with brief quotes.

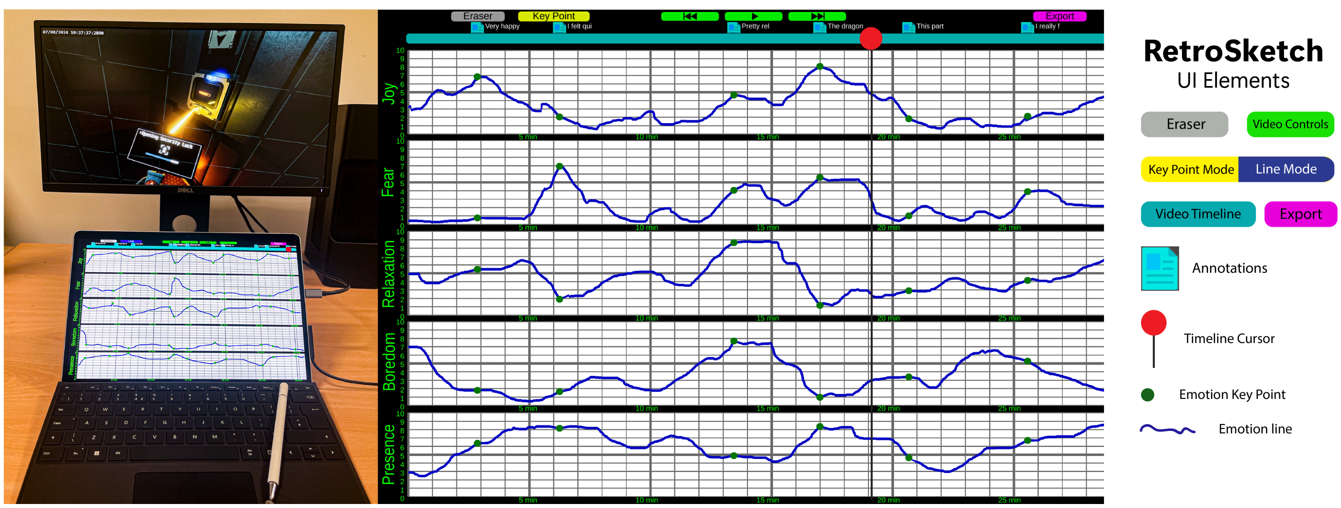


Figure 3: Left: Digital version of RetroSketch showing the tablet and stylus used for input which are connected to a bigger display for video playback. Right: RetroSketch's user interface with buttons for sketching and playback functions, graphs with a timeline cursor, and keypoints with annotations.

to accurately place emotional responses and keypoints for specific moments of the VR experience.

Video Controls: Basic video controls are provided and synchronised with the timeline cursor, including options to play, pause, fast forward 5 seconds, rewind 5 seconds, and adjust playback speed to 1x, 1.5x, 2x, or 2.5x.

Line Drawing: The digital version was specifically designed for a stylus and tablet to mimic the flexibility and expressiveness of sketching with the paper-based version. This includes an eraser tool so that users can refine any erroneous parts of their sketch, including keypoints.

Keypoints & Annotations: The digital version tightly integrates keypoints and annotations. After placing a keypoint, an annotation text box appears, prompting the user to describe the event and their feelings in more detail. The annotation is then displayed above the timeline alongside the keypoints.

Data Export: After completing their sketch, the data can be exported at one sample per second. For example, a 30-minute VR session results in 1,800 samples for each emotion and presence. The export feature also generates a log file of all user actions performed during sketching, an aggregated keypoint file with keypoints and annotations, and a screenshot of the completed sketch.

4 STUDY METHODOLOGY

To address our research questions, we conducted a mixed-design study with VR experience (*VR Game*) as a between-subject factor with five levels (*ACC*, *GotS*, *HLA*, *IEYTD* and *RM*) and whether or not ESM was used alongside RetroSketch as within-subject factor (*ESM* and *NoESM*). Having one session with and one without ESM, counterbalanced using a balanced Latin square design, allows us to investigate how ESM affects the user experience (RQ3). The main study methodology was first informed by a pilot study.

4.1 Pilot Study

We conducted a pilot study with 10 participants using the paper-based RetroSketch to inform the design and evaluation of digital RetroSketch. All participants played Skyrim-VR [242] during two 30-minute gameplay sessions, one with ESM measures and one without. After each session, participants watched back their experience and used the paper-based RetroSketch to sketch their Joy, Fear, Relaxation, Boredom, and Presence. There were significant moderate-to-strong Kendall's τ correlations between RetroSketch and ESM for all emotions and presence. Based on participant feedback and experimenter observations, we found that the paper-based RetroSketch is most suitable for small-scale studies. The Supplementary Material Document provides further details.

4.2 VR Games

We chose five state-of-the-art single-player VR games that cover a broad range of game mechanics, themes, aesthetics, challenges, immersive elements, and emotional components (see Figure 4 and the Video Figure in Supplementary Material). The games use different configurations including controllers, space required, and player movements, which allow us to validate RetroSketch more broadly. We focus on single-player games because online multiplayer gameplay can vary vastly based on other players' actions, making it difficult to control what participants would experience. Based on the pilot study, we excluded Skyrim-VR due to an overly long tutorial and usability issues.

Participants were randomly assigned a VR game (n = 28 per VR game) and completed a 10-minute tutorial and two 30-minute VR sessions, which is close to average VR gameplay times [1] and the recommended time for VR usage [91, 147, 235]. We chose a session duration of 30 minutes to ensure participants had substantial exposure to their respective VR experience and validate RetroSketch's ability to produce accurate measures throughout that exposure.

The procedure for each game was refined through piloting to ensure a natural and sequential flow between the first and second gameplay sessions, minimising disruptions for participants and preserving ecological validity. For more detailed descriptions of each VR game and the gameplay tutorials, please refer to Section 2 of the Supplementary Material Document. Each participant played one of the following games:

Assetto Corsa Competizione (ACC): a racing simulator featuring various cars and circuits across the world. We chose ACC to elicit feelings of high arousal and high valence because it features highly realistic graphics and sound that are designed to closely resemble a real racing experience. ACC uses an immersive haptic driving simulator with a six-degrees-of-freedom motion platform (Figure 5-D) that simulates force feedback from acceleration, cornering, road surfaces, and collisions.

Garden of the Sea (GotS): an open-world crafting, farming and exploration game. GotS offers a more relaxed experience eliciting feelings of low arousal and high valence because of its bright and cartoonish style, emphasising a tranquil and open-ended experience with meditative elements. Quests are followed and unlocked at the player's pace and can be ignored entirely. GotS uses VR controllers and is a standing, room-scale VR experience.

Half-Life: Alyx (HLA): a critically acclaimed first-person action-horror game with a rich, sci-fi story in which Earth is invaded and controlled by an alien race. The game requires players to fight aliens and zombies while solving puzzles and exploring a post-apocalyptic city. We chose HLA to elicit high arousal and low valence because of its horror themes. HLA uses VR controllers and is a standing, room-scale VR experience.

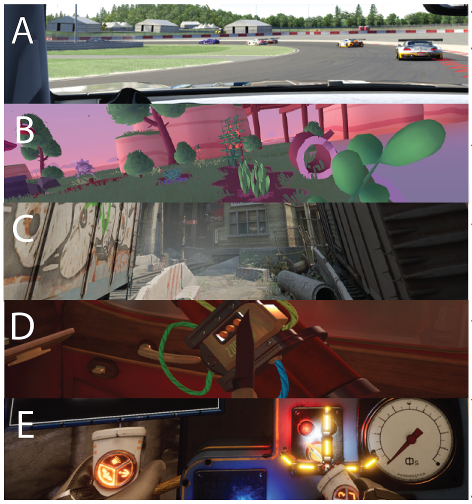
I Expect You To Die (IEYTD): an escape-room style puzzle game where players embody a secret agent. In each mission, the player finds themselves in impossible and deadly scenarios which they need to creatively solve and escape by interacting with the environment. We chose IEYTD to elicit high arousal and medium levels of valence because of its high-risk scenarios and cartoonish style with a 'tongue-in-cheek' comedic tone. IEYTD uses VR controllers and is a seated VR experience.

Red Matter (RM): a story-driven puzzle and adventure game set during a dystopian sci-fi Cold War in which the player embodies an astronaut sent on a mission to an abandoned space station. We chose RM to elicit low arousal and low valence because the story unfolds slowly, being uncovered through scanning objects and documents in a mysterious and unsettling atmosphere. The game uses VR controllers and is a standing, room-scale VR experience.

4.3 Apparatus

We used a Vive Pro Eye VR headset for all experiences. All study sessions were completed in a private University lab space which afforded an open tracking space of more than 3×3 meters for the room-scale VR experiences. The driving simulator used for ACC was in the same lab space and is composed of a Next Level Racing V3 motion platform offering six degrees of freedom [179], a Fanatec haptic steering wheel [59], and a vibrotactile seat [178] (Figure 5D).

Physiological measures were collected through eye (pupillometry) and lip tracking (facial gestures) from the VR headset (Figure 5A), a Shimmer3 GSR+ [214] with electrodes on the participant's middle and ring finger [44] (Figure 5B), and a Polar H10 chest strap HR monitor [202, 252] (Figure 5C). All physiological measures were collected using the EmoSense Unity SDK [170, 171, 173] which we modified to run in the background of each VR experience, allowing access to both the inbuilt VR eye and lip tracker, as well as the Polar and Shimmer devices. Data was streamed to the same PC running the VR experiences (Intel Core i9 Extreme, 64GB DDR5 RAM, Nvidia RTX 3090) over Bluetooth (BLE protocol), and recorded at a sample rate of 60Hz. Gameplay footage from the participant's



Game	VR Configuration	Genre	Mechanics	Themes
Assetto Corsa Competizione (*ACC*)		Sport, Simulation	Racing, Competitive, Realism	Action, Adrenaline, Exciting
Garden of the Sea (*GotS*)		Indie, Simulation	Crafting, Farming, Exploration	Wholesome, Cute, Relaxing
Half-Life: Alyx (*HLA*)		Horror, Survival	First-person Shooter, Loot Management, Story Driven	Cinematic, Alien Invasion, Exciting
I Expect You To Die (*IEYTD*)		Thriller, Mystery	Puzzle, Escape Room, Missions	Comedic, Noir, Nostalgic
Red Matter (*RM*)		Scifi, Adventure	Puzzle, Exploration, Story Driven	Cinematic, Secretive, Ominous

Figure 4: A-E shows the five VR experiences used in the study: (A) Assetto Corsa Competizione - ACC, (B) Garden of the Sea - GotS, (C) Half Life Alyx - HLA, (D) I Expect You To Die - IEYTD, and (E) Red Matter - RM. The right table summarises the VR configuration and controls, game genres, mechanics and themes.

perspective was captured using OBS 29.1.3 [118], recording the SteamVR view at a resolution of 1920×1080.

4.4 Measures

We collected a range of measures, including subjective emotion and presence ratings via RetroSketch and ESM, ten physiological metrics, post-session VR user experience questionnaires, and qualitative data on the use of RetroSketch and ESM.

4.4.1 RetroSketch & ESM. After each 30-minute gameplay session of every VR game, participants used RetroSketch to measure their Joy, Fear, Relaxation, Boredom, and Presence each on a 0-10 scale as described in subsection 3.2.

In one of the two sessions, interval-contingent ESM was used based on best practices from the literature [119]. In every VR experience and every five minutes, participants were asked via automated voice recordings through the headset speakers to rate their Joy, Fear, Relaxation, Boredom and Presence, in random order. Interval-contingent sampling was chosen over event- or signal-contingent sampling to 1) avoid demand characteristics [156], 2) maintain consistent sampling across sessions and participants, and 3) mitigate surprise and prevent participants from spending time reflecting on when to complete a sample, as this could disrupt presence.

Participants answered 11-point Likert-scale questions (0-10) such as "*On a scale of 0 to 10, how Joyful do you feel?*" with 0 being "*none of that feeling*" and 10 being "*the most intense version of that feeling possible*". Five-minute intervals were chosen to strike a balance between disruptions caused by ESM and the amount of data collected. To mitigate disruption, we did not pause the gameplay during ESM samples as pilot testing showed that participants could answer while continuing to play.

4.4.2 Physiological Measures. Physiological measures were recorded for the whole VR experience and aggregated over consecutive 60-second windows, resulting in 30 data points per session. Pupillometry was recorded using the Vive eye tracker [258] including pupil dilation level (PDL) as mean pupil diameter in millimetres, and pupil dilation response (PDR) as standard deviation of the pupil diameter. The standard deviation has been previously used to quantify phasic responses during prolonged exposures [5, 205], including in VR settings [137, 170]. We also recorded Blink Rate (BR) as the mean inter-blink interval in seconds, and Blink Duration (BD) as the mean blink length in milliseconds. Electrodermal Activity (EDA) was recorded using the Shimmer device as mean Skin Conductance Level (SCL) in micro Siemens (μS), and Skin Conductance Response (SCR) as standard deviation of SCL. Cardiac activity was recorded using the Polar H10 as beats per minute (HR), and heart rate variability (HRV) as root mean square of successive differences (RMSSD) of interbeat (RR) intervals [211]. Facial gestures were tracked by observing the movements of the zygomaticus major muscle (Smile) and the orbicularis oris muscle (O-Shape). These were quantified using the Vive facial tracker's blend shape weightings for 'Mouth Smile' and 'Mouth O-Shape', respectively [258]. We removed erroneous sensor measures based on absolute thresholds: we filtered out skin conductance values above 100μS and below 0.1 μS [7, 27], RR interval values above 2000 ms and below 200 ms (30-200 bpm) [109], and any pupil dilation measures recorded while the eyes were closed.

4.4.3 Pre-Study Questionnaires. Participants completed a demographics questionnaire (age, gender, VR/video game experience) as well as measures assessing their personality (Big5 [68, 69]), video game player type (Tondello [248]), and immersive tendencies (ITQ [266]). These measures were used to test the reliability and consistency of RetroSketch and ESM measures across different covariates. Additionally, we collected baselines for each emotion sampled in RetroSketch and ESM, allowing us to understand a participant's baseline disposition, as well as baselines for simulator/motion sickness (SSQ [20]) and ratings of perceived exertion (BORG-RPE [168]) – two factors that are important to control for when measuring physiological markers [83, 170].

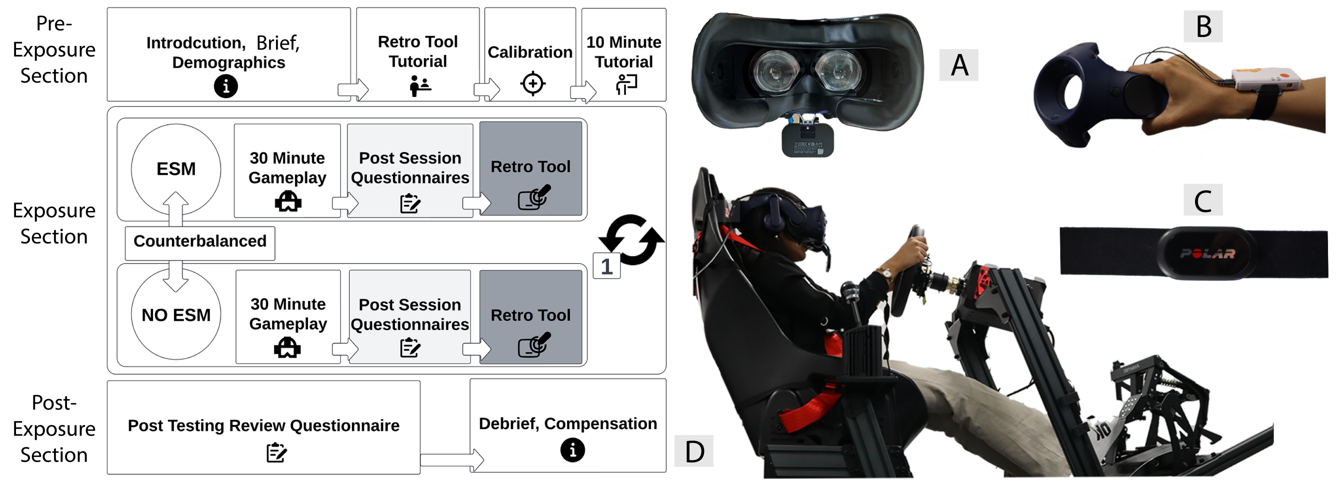


Figure 5: Left: A diagram of the study procedure. The pre-exposure section includes introduction, debrief, demographics, RetroSketch tutorial, calibration, and a gameplay tutorial. The exposure section includes two counterbalanced 30-minute gameplay sessions (*ESM* and *NoESM*) followed by questionnaires and RetroSketch use. The post-exposure section includes questionnaires and debrief. Right: (A) HTC Vive Pro Eye VR headset with eye and lip tracker, (B) VR controller and Shimmer skin conductance sensor, (C) Polar H10 heart rate sensor and chest strap, and (D) Next-level racing simulator with 6-degrees-of-freedom haptic and motion platform.

4.4.4 Post-Session Questionnaires. Immediately after each 30-minute VR session, we took measures of participants' experienced VR presence using MPS [133], intrinsic motivation using IMI [141], and flow-state using PPL-FSQ [132], all of which have been used extensively in prior VR research. Additionally, we again measured participants' simulator sickness levels (SSQ [20]) and ratings of perceived exertion (BORG-RPE [168]). These measures were collected to test whether ESM influences the VR user experience.

4.4.5 Post-Study Questionnaires. At the end of the study, participants answered open-ended questions asking them to evaluate and compare RetroSketch and ESM, such as how accurate they perceived the respective method to be, how easy it was to recall feelings with RetroSketch, whether ESM influenced their VR experience, and which method they preferred.

4.5 Procedure

Participants were first screened for health risks when using VR technology, such as epilepsy, mobility impairments and severe visual impairments (see Supplementary Material Document). Participants were then assigned to a VR experience using a balanced randomisation process designed to ensure gender balance across the five VR games. While most VR games were tested in parallel, the ACC sample was completed in a single batch due to logistical constraints with the driving simulator. Participants were briefed about the VR game they would play and verbally screened by the experimenter for content sensitivities. If any concerns were raised, the participant was reassigned to a different VR experience at random. After providing informed consent and completing pre-study questionnaires, participants were introduced to both RetroSketch and ESM.

For RetroSketch, participants were given a 10-minute introduction and tutorial demonstrating the keypoint, annotation, line and eraser features using a stock video as example. To set an expectation of the level of detail for the keypoints when using RetroSketch, participants were recommended that they should create at least one keypoint every 5 minutes. However, participants were informed this was not a strict rule, and it was emphasised that they should decide where to place keypoints by themselves.

Participants were introduced to the VR headset and the physiological sensors, and the sensors were calibrated. Participants were then introduced to the allocated VR gaming experience through a 10-minute tutorial. This was followed immediately by the first 30-minute gameplay session. After completing the first gameplay session, participants were given verbal cues by the experimenter instructing them to pause the game and exit VR. This was followed by a post-session questionnaire and a short break which in total took approximately 10-15 minutes, before participants used RetroSketch to measure their VR experience. Participants were told that they would have approximately 15 minutes to complete their sketch, but that it was perfectly fine if they needed more time. In practice, participants took on average 25 minutes to complete their sketch. The entire experimental procedure took approximately 3 hours and participants were compensated with £30 for their time.

4.6 Participants

We recruited 140 participants (58 female, 78 male, 3 non-binary, 1 undisclosed) aged between 18-61 (M = 25.379, SD = 7.720), who were predominantly staff and students of the University of Bath. In total 167 participants were recruited. However, 18 participants withdrew due to VR sickness, three were excluded because of technical issues, and two chose to withdraw. Additionally, the data of four participants was excluded due to sensor data errors. Most participants had used VR occasionally (42 never, 91 occasionally,

Table 1: Demographics and experience of participants across the different VR experiences: Assetto Corsa Competizione (ACC), Garden of the Sea (GotS), Half-Life Alyx (HLA), I Expect You To Die (IEYTD), and Red Matter (RM).

Game	Gender	Age	VR Exp.	Game Exp.
ACC	M= 16, F= 10 NB= 1 Other= 1	23.750 ±4.486	Occa.= 16 ≥ Weekly=7 Never= 5	Never= 24 Once= 3 > Once= 1
GotS	M= 16, F= 12	25.000 ±8.590	Occa.= 22 Never= 6	Never
HLA	M= 15, F= 13	24.714 ±6.588	Occa.= 19 Never= 9	Never= 27 Once= 1
IEYTD	M= 15, F= 13	26.786 ±9.183	Occa.= 16 Never= 12	Never
RM	M= 16, F= 10 NB= 2	26.643 ±8.841	Occa.= 18 Never= 10	Never

4 weekly, 3 daily). Most participants had no prior experience with the VR game they were allocated (135 never, 4 once, 1 more than once). Table 1 shows a breakdown of the demographics for each VR experience. A power analysis using G*Power 3.1.9.7 showed that we can detect medium-sized differences between RetroSketch and ESM measures at $\alpha = .05$ with a power of 0.999, even when simple non-parametric Wilcoxon-signed rank tests are used which do not take advantage of the multiple RetroSketch and ESM samples for each participant.

5 RESULTS

Data was analysed with `R` v4.4.1 using various packages. For all tests, we used a significance threshold of $\alpha = .05$ (*), as well as $\alpha = .01$ to denote 'highly significant' results (**) and $\alpha = .001$ for 'very highly significant' results (***). For clarity, we report effect sizes mainly using the popular Cohen's d, converting other forms of effect sizes such as η^2 and log odds ratios to d using the effect size conversion functions of the `effectsize` package [16]. Tables indicate the magnitudes of significant effects by highlighting table cells in green: the stronger the colour, the larger the effect. The analysis script, aggregated dataset, and additional results are available in Supplementary Material.

5.1 Emotion Manipulation

Figure 6 left and right summarise the emotional footprints of each experience as captured by RetroSketch and ESM. We first performed a manipulation check to ascertain that the experiences elicited different emotions, and to provide an understanding of the range and intensity of emotions elicited. Anderson-Darling tests from the `nortest` package [177] confirmed non-normality of the data (they are more reliable for large sample sizes than the more common Shapiro-Wilk [66]), and Levene's tests confirmed violations of homogeneity of variance. Therefore, we used non-parametric Kruskal-Wallis ANOVAs to test the main effect of VR Game on emotion and presence ratings as measured by RetroSketch in all sessions, followed by pairwise Dunn's tests with Holm-Bonferroni posthoc correction.

The main effect of *VRGame* was very highly significant for all emotions and presence ($\chi^2(4) \geq 81.255, p \leq .001^{***}, \eta^2 \geq .009$). Pairwise comparisons showed that all experiences elicited significantly different Presence. 42 of the 50 pairwise comparisons of emotions were also significant, with more closely related games such as HLA, IEYTD and RM not always showing significant differences (details in Supplementary Material Document).

5.2 Internal Consistency of RetroSketch

To assess the internal consistency of quantitative and qualitative measures of RetroSketch, two researchers independently reviewed 40 RetroSketch sketches (8 for each game). The researchers examined the keypoints and annotations created by participants and cross-referenced them with the respective VR gameplay footage. In addition, the annotations of all RetroSketch sketches were analysed using a Twitter-roBERTa-base model fine-tuned for sentiment analysis [30, 130, 131]. The two researchers assessed the sketches using the following three criteria:

(1) **Annotations correspond to the associated gameplay footage:** The coder assessed whether an annotation related to the associated gameplay footage (true or false), e.g. ensuring the participant reflected on the associated moment and not a different moment.
(2) **Keypoint ratings are consistent with the associated annotations:** The coder assessed whether the RetroSketch ratings assigned to a keypoint are consistent with its annotation (true or false). For example, if a participant annotated a keypoint as *"The most enjoyable part of the experience"* but the *Joy* rating was not the highest for the experience then this would be marked as false. Importantly, this was not a value judgement of a participant's ratings (i.e. whether they are too low or too high) but a solely a judgement of consistency with the rest of the sketch.
(3) **Sentiment analysis scores are accurate:** The coder assessed whether the sentiment score produced by the sentiment analysis model accurately reflected the conveyed sentiment of the annotation (true or false). For example, if an annotation for a moment in *ACC* stated *"I was extremely happy to overtake two other drivers on this corner"* then the positive sentiment should be high (≥ 0.7) and the negative sentiment should be low (≤ 0.3). This was done to validate the sentiment analysis method in the context of RetroSketch, with a view to applying it to address RQ5.

The two researchers (R1 and R2) found that over 99% of annotations correctly corresponded to the associated video footage (R1: 99.2%, R2: 99.5%), over 98% of the keypoint ratings were consistent with their annotations (R1: 98.4%, R2: 98.9%), and the majority of sentiment analysis scores accurately reflected the annotations (R1: 72.0%, R2: 67.5%). Agreement scores between the two reviewers were computed using Prevalence-Adjusted Biased-Adjusted Kappa (PABAK) [159] as opposed to Cohen's Kappa due to the distribution being highly skewed towards 'valid' scores compared to 'invalid' scores. The agreement between coders was almost perfect (0.98 and 0.97) for the first two criteria which internally validate RetroSketch, and the agreement was substantial (0.65) for the sentiment analysis.

DV	Method	ACC	GotS	HLA	IEYTD	RM
Joy	Retro	6.269	6.516	6.056	5.752	5.953
	ESM	6.771	6.774	6.238	6.714	5.524
Fear	Retro	2.833	1.037	4.019	2.065	2.045
	ESM	2.518	0.786	4.458	2.113	2.232
Relaxation	Retro	3.590	5.972	4.337	4.847	4.663
	ESM	4.211	7.077	5.054	5.482	4.827
Boredom	Retro	2.110	3.108	2.326	2.313	2.421
	ESM	2.205	3.321	1.958	1.780	2.411
Presence	Retro	7.051	7.241	7.600	7.326	6.842
	ESM	7.139	7.917	8.214	7.923	6.887

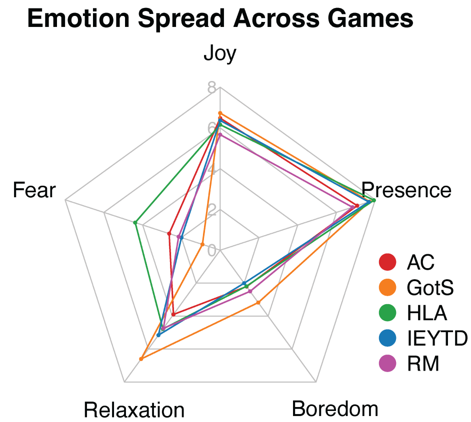


Figure 6: Left: Table showing averages for joy, fear, relaxation, boredom, and presence for RetroSketch and ESM in all five VR games. Right: Spider chart of average emotion and presence RetroSketch ratings across the five VR games.

Table 2: Correlations between RetroSketch and ESM across all VR games. Pearson's r and Kendall's τ values describe overall correlations. $\tau_5, \ldots, \tau_{30}$ are separate Kendall's Tau correlations for the six ESM samples at $5, 10, \ldots, 30$ minutes, and $r_{min} = min(r_5, \ldots, r_{30})$ describes the worst case Pearson's correlation across the six samples. All τ values are tested for significance.

DV	r	r_{min}	τ	τ_5	τ_{10}	τ_{15}	τ_{20}	τ_{25}	τ_{30}
Joy	.661	.418	**.537******	**.350******	**.421******	**.367******	**.448******	**.445******	**.460******
Fear	.761	.581	**.586******	**.441******	**.466******	**.457******	**.507******	**.496******	**.556******
Relaxation	.747	.500	**.561******	**.361******	**.428******	**.407******	**.487******	**.446******	**.483******
Boredom	.725	.454	**.575******	**.442******	**.443******	**.476******	**.495******	**.487******	**.529******
Presence	.732	.521	**.549******	**.453******	**.519******	**.438******	**.490******	**.552******	**.438******

5.3 RQ1: How do RetroSketch measures relate to ESM measures?

5.3.1 Correlations between RetroSketch and ESM.. We first analysed the correlations between RetroSketch and ESM emotion and presence ratings.Scatterplots suggest an approximately linear relationship between RetroSketch and ESM ratings, therefore we use Pearson correlation coefficients r to describe the strength of the relationships. However, Shapiro-Wilk tests and QQ plots showed that normality was violated, therefore we used non-parametric Kendall's Tau (τ) correlation tests with Holm-Bonferroni posthoc correction to confirm the relationships statistically. Table 2 summarises the correlations overall (r and τ) and per sample interval ($\tau_5, \ldots, \tau_{30}$), showing that they were highly significant with compellingly 'strong' effects ($\tau \geq 0.4$).

We assessed the stability of the correlations across the different VR experiences by testing interactions with *VR Game* through regression analysis. For linear regressions, residual plots showed that the assumptions of normality and heteroscedasticity were violated. Therefore we performed repeated-measures ordinal logistic regressions using the `ordLORgee` function of the `multgee` package [249, 250]. No significant interaction effects were found, indicating that the correlations between RetroSketch and ESM are robust and stable across all five VR experiences.

5.3.2 Distribution Characteristics of RetroSketch and ESM. Next, we compared the distribution characteristics of RetroSketch and ESM ratings. To address non-normality, we performed two-way Align Rank Transform (ART) ANOVAs [267] using the `ARTool` package [104]. We tested the effects of the measurement *Method* (RetroSketch or ESM) and the *VR Game* on the median, mean absolute deviation (MAD), minimum and maximum values of each participant's emotion and presence ratings, respectively. Pairwise comparisons were performed using ART-C tests [54] with Holm-Bonferroni posthoc correction.

Table 3 shows the overall differences (Δ) in distribution characteristics of RetroSketch compared to ESM, their significance, and the size of their effect (η^2 converted to Cohen's d). The table also shows significant interactions broken down by *VR Game*, i.e. when differences are particularly strong for a particular game. RetroSketch generally captures a higher variation and range for emotions and presence. For example, when using RetroSketch participants' ratings have significantly higher MAD and maximum values for Joy ('large' and 'medium' effect), and lower minimum values for Joy ('large' effect). Additionally, RetroSketch generally yields lower ratings for 'positive' emotions compared to ESM (see in particular Relaxation), lower Presence ratings, and higher Boredom ratings.

Table 3: Distribution characteristics of RetroSketch and ESM: Median, Mean Absolute Deviation (MAD), Minimum, and Maximum values. Δ is the difference between RetroSketch and ESM (*RetroSketch – ESM*), which is tested for significance. Effect sizes Cohen's *d* is visualised using shades of green. Significant interactions with specific games are shown in separate rows.

DV	Game	*Median*		*MAD*		*Min*		*Max*	
		Δ	Cohen's *d*	Δ	Cohen's *d*	Δ	Cohen's *d*	Δ	Cohen's *d*
Joy	ALL	−0.226	0.074	**0.535******	**1.237**	**−0.958******	**1.066**	**0.437******	**.0.613**
Fear	ALL	0.006	0.005	0.113	0.049	−0.083	0.083	0.000	0.065
	GotS	**−0.259******	**0.210**						
Relaxation	ALL	**−0.6041******	**0.387**	0.17	0.257	**−0.594******	**0.679**	**−0.486******	**0.606**
Boredom	ALL	**0.162***	**0.135**	**0.226***	**0.613**	0.032	0.373	0.312	0.242
	IEYTD	**−0.446***	**0.235**						
Presence	ALL	**−0.305******	**0.198**	**0.280******	**0.829**	**−0.578***	**0.496**	**−0.210***	**0.401**
	HLA	**0.569***	**0.327**						
	IEYTD	**0.512***	**0.257**						

5.4 RQ2: How reliable is RetroSketch & ESM across different VR experiences and users?

To address RQ2, we analyzed interactions between 15 user covariates across all 140 participants. These covariates include demographics, personality traits, video game player types, immersive tendencies, and participants' methodological preference (RetroSketch or ESM). For each measure, we aggregated ratings across the six ESM samples taken per session, as well as the six corresponding RetroSketch ratings, using the mean. Then we performed regression analyses to determine whether the covariates influence the correlations between RetroSketch and ESM by testing interactions with those covariates. Finally, we tested the influences of the covariates on the RetroSketch and ESM ratings themselves, e.g. how gender influences measured emotions.

5.4.1 The Influence of Covariates on the Correlations between RetroSketch and ESM. Due to violations of normality, ordinal logistic regressions with the `polr` package [175, 256] were used. Demographic variables such as Age, Gender, and VR Experience did not significantly influence the correlations between RetroSketch and ESM, indicating that RetroSketch is robust across different demographic groups. Similarly, participants' preference for RetroSketch or ESM had no significant influence. Two significant interactions were observed for Big Five personality traits: for Presence, there were interactions with Agreeableness ($d = 0.032$) and Conscientiousness ($d = 0.024$).

Lastly, for immersive tendencies, two significant interactions were detected: with Presence ($d = -0.002$) and Joy ($d = -0.003$). This indicates that as immersive tendencies increase, the correlation between RetroSketch and ESM for Presence and Joy decreases slightly. The effect sizes of these interactions were 'tiny' ($d < 0.1$), suggesting these effects are negligible and the correlations between RetroSketch and ESM are robust across different types of people.

5.4.2 The Influence of Covariates on RetroSketch and ESM Measures. Table 4 and Table 5 provide an overview of the influences of covariates on RetroSketch and ESM measures. For continuous covariates (e.g. Age), correlation analyses were used. Scatterplots suggest approximately linear relationships, therefore we used Pearson correlation coefficients r to describe the strength of the relationships. However, normality was violated, therefore we used non-parametric Kendall's Tau (τ) correlation tests with Holm-Bonferroni posthoc correction to confirm the relationships statistically. For categorical covariates (e.g. Gender), we used three-way ART-ANOVAs with the covariate, measurement method (RetroSketch and ESM) and *VR Game* as factors, followed by pairwise ART-C tests with Holm-Bonferroni posthoc correction. For gender analysis, only male and female identities were considered due to the low sample size of non-binary (3) and undisclosed (1) identities.

A key finding from Table 4 is that ESM Joy ratings were significantly higher for males compared to females with a medium effect size – an effect not observed in RetroSketch measures. Another notable finding is the significant positive correlation between Presence and Age for RetroSketch where none was found for ESM. Various Big-5 personality traits also showed significant effects, some influencing only RetroSketch (e.g. Extroversion and Fear), others only ESM (e.g. Fear and Openness), and some both (e.g. Relaxation and Conscientiousness). However, the observed effect sizes were small to negligible.

Regarding player types, Table 5 highlights that Tondello Challenge was significantly correlated with both RetroSketch and ESM Joy. The correlation was moderate for ESM and weaker for RetroSketch, suggesting that players who seek challenges report higher Joy with ESM than with RetroSketch. Additionally, ESM Joy correlated with the Social trait, ESM Relaxation with the Challenge trait, and ESM Boredom with the Narrative trait, whereas the corresponding RetroSketch ratings did not. While these correlations are significant, they are weak ($\tau < 0.2$). Finally, both RetroSketch and ESM Presence measures significantly correlated with immersive tendencies (ITQ), indicating higher Presence ratings for those with stronger immersive tendencies.

While Table 4 and Table 5 present the results of covariates across all VR games, we also tested each covariate for interactions with *VR Game*, using ordinal logistic regressions for continuous covariates

Table 4: Relationships of RetroSketch and ESM ratings with different user covariates. Gender (Male = Male - Female) and methodological preference (Pref. Retro = RetroSketch - ESM) were tested with ART-ANOVAs and described using the mean group difference $\Delta Mean$ and effect size Cohen's d. The relationships with continuous covariates, VR experience (VR Exp.) and Big5 personality traits, are described with Pearson's r and Kendall's τ correlations, with significance tests performed on the τ.

DV	Method	Male		Pref. Retro		Age		VR Exp.		Extroversion		Agreeableness		Conscientious		Neuroticism		Openness	
		$\Delta Mean$	d	$\Delta Mean$	d	r	τ	r	τ	r	τ	r	τ	r	τ	r	τ	r	τ
Joy	*Retro*	0.541	0.293	0.263	0.138	−.088	.027	.077	.025	.035	.005	.083	.052	.064	.034	.110	.074	.049	.036
	ESM	**0.908****	0.516	−0.054	−0.030	−.122	.065	−.002	−.048	.079	.060	.087	.059	.117	.078	.224	.166	−.040	−.070
Fear	*Retro*	−0.751	−0.417	0.835	0.460	−.044	.046	.014	−.005	−.200	**−.130***	−.064	−.057	−.163	−.103	−.106	−.048	−.090	−.070
	ESM	−0.889	−0.456	0.783	0.387	−.024	.0240	.017	.004	−.15	−.099	−.104	−.085	−.225	**−.150***	−.070	−.030	−.179	**−.150***
Relaxation	*Retro*	0.706	0.352	−0.353	−0.172	−.055	.057	.011	.012	.155	.089	.255	**.177****	.203	**.130***	.140	.068	.085	.029
	ESM	0.762	0.388	−0.764	−0.388	−.143	−.052	−.059	−.049	.167	.088	.24	**.169****	.234	**.132***	.170	.080	.043	.016
Boredom	*Retro*	−0.149	−0.091	−0.358	−0.221	.005	−.014	−.030	.013	.044	.026	.028	.031	.02	.039	−.109	−.095	.073	.057
	ESM	−0.116	−0.064	−0.396	−0.230	−.027	−.088	.001	.04	.099	.051	.044	.016	.004	−.012	−.150	−.135	.116	.056
Presence	*Retro*	−0.183	−0.110	0.069	0.038	.077	**.118***	.090	.042	.079	.052	.050	.045	.110	.059	.142	.088	.084	.048
	ESM	−0.054	−0.035	−0.356	−0.230	.024	.108	−.038	−.030	.017	.003	.050	.029	.08	.070	.146	.090	.036	.037

Table 5: Correlations of RetroSketch and ESM ratings with different user covariates. Tondello player traits (T) and immersive tendencies (ITQ) are described with Pearson's r and Kendall's τ, with significance tests performed on the τ.

DV	Method	T. Challenge		T. Aesthetic		T. Narrative		T. Goal		T. Social		ITQ	
		r	τ	r	τ	r	τ	r	τ	r	τ	r	τ
Joy	*Retro*	0.217	**0.140***	0.115	0.088	0.171	0.114	0.075	0.052	0.129	0.097	0.053	0.037
	ESM	0.308	**0.205*****	0.050	0.068	0.161	0.092	0.053	0.053	0.212	**0.142***	0.077	0.068
Fear	*Retro*	−0.074	−0.037	0.025	0.015	0.013	−0.006	0.022	−0.000	0.005	0.007	0.137	0.085
	ESM	−.119	−0.053	0.024	0.008	−0.047	−0.020	0.039	0.034	−0.32	−0.028	0.143	0.105
Relaxation	*Retro*	0.136	0.092	−0.004	0.012	−0.092	−0.075	0.055	0.037	0.145	0.078	0.013	0.014
	ESM	0.187	**0.119***	0.015	0.023	−0.030	−0.028	0.038	0.0153	0.211	0.109	−0.0136	−0.024
Boredom	*Retro*	0.155	−0.092	−0.154	−0.087	−0.307	−0.185	0.100	−0.048	−0.016	−0.028	0.019	0.020
	ESM	−0.043	−0.060	−0.135	−0.093	−0.227	**−0.131***	−0.049	−0.034	0.094	0.003	0.013	−0.033
Presence	*Retro*	0.069	0.068	0.006	0.025	0.052	0.056	0.000	0.000	0.083	0.068	0.172	**0.121***
	ESM	0.039	0.068	0.072	0.039	−0.128	−0.051	−0.012	0.017	0.066	0.083	0.249	**0.157****

and three-way ART ANOVAs for categorical covariates. Numerous significant interactions were found, suggesting that covariates influence RetroSketch and ESM scores differently depending on the VR experience (see Supplementary Material Document). However, all significant interactions have tiny to very small effect sizes, suggesting that they have little relevance in practice.

5.5 RQ3: How does ESM influence the VR user experience?

To answer RQ3, we tested the differences in user experience measures between the *ESM* and *NoESM* conditions using two-way ART-ANOVAs, with factors *ESM* vs. *NoESM* and *VR Game*, followed by ART-C tests with Holm-Bonferroni posthoc correction. Table 6 summarises the effects of *ESM* compared to *NoESM*, both overall and per *VR Game* to highlight interactions.

Notable findings from Table 6 include a significant increase in IMI Pressure/Tension during *ESM* sessions overall. However, this effect is not consistent across all games. Specifically, in *HLA* and *IEYTD*, using ESM significantly reduced the experienced pressure. Additionally, there are significant effects on various presence measures. Overall, ESM significantly reduced how physically present participants felt, while significantly increasing their feelings of self and social presence. Similar to pressure, the effects on presence vary across different VR games.

Overall, ESM significantly decreased Flow Absorption, which was particularly pronounced for *ACC*, *HLA*, and *RM*, albeit with small effect sizes. However, once again the effects are not the same for all VR experiences and ESM significantly increased Flow Absorption in *IEYTD*, although with a much smaller effect size. We also note the comparatively larger effects seen in *RM* across several measures. Although the effect sizes for the differences between *ESM* and *NoESM* range from small to tiny, it is clear that ESM influences the user experience in a measurable yet seemingly unpredictable way, heavily dependent on the specific VR experience.

5.6 RQ4: How do RetroSketch & ESM relate to physiological measures?

To address RQ4, we examined the relationships between both measurement methods (ESM and RetroSketch) and ten physiological measures commonly used in the VR emotion recognition literature [88, 137, 244]. Scatterplots suggest approximately linear relationships between physiological measures and emotion and presence ratings, therefore we use standardised linear regression coefficients β to describe the strength of the relationships. The regression coefficients were estimated with multi-level linear regression

Table 6: The effects of ESM on the VR user experience shown across all games (ALL) and for each of the five VR games (ACC, GotS, HLA, IEYTD, and RM), tested with ART-ANOVAs. The effects of ESM are given as mean differences $\Delta Mean$ (ESM - $NoESM$) and Cohen's d effect sizes.

DV	ALL		ACC		GotS		HLA		IEYTD		RM	
ESM − NoESM	Δ *Mean*	Cohen's *d*	Δ *Mean*	Cohen's *d*	Δ *Mean*	Cohen's *d*	Δ *Mean*	Cohen's *d*	Δ *Mean*	Cohen's *d*	Δ *Mean*	Cohen's *d*
IMI Competence	0.019	0.007	**0.143****	0.090	**0.220*****	0.151	**0.226*****	0.129	0.101	0.070	**−0.595*****	−0.375
IMI Interest	−0.027	−0.020	−0.087	−0.072	0.158	0.113	0.020	0.018	**0.173*****	0.162	**−0.398*****	−0.310
IMI Pressure	**0.093*****	0.074	**0.257*****	0.197	**0.121*****	0.116	**−0.236*****	−0.161	**−0.107*****	−0.076	**0.429*****	0.302
IMI Effort	**−0.050****	−0.061	−0.050	−0.054	**−0.086*****	−0.072	**−0.171*****	−0.144	**0.136*****	0.121	−0.079	−0.065
MPS Physical	**−0.009*****	−0.104	**−0.061*****	−0.333	**0.024*****	0.152	**0.036*****	0.365	−0.000	−0.000	**−0.043*****	−0.255
MPS Self	**0.006*****	0.204	**−0.003****	−0.013	0.011	0.064	**−0.017*****	−0.105	**0.033*****	0.158	**0.007****	0.035
MPS Social	**0.017*****	0.212	**−0.020*****	−0.101	**0.031*****	0.148	**0.036*****	0.222	**0.033*****	0.151	0.003	0.017
Flow Absorp	**−0.048*****	−0.004	**−0.095*****	−0.122	0.083	0.112	**−0.111*****	−0.166	**0.060***	0.086	**−0.179*****	−0.250
Flow Challenge	0.001	0.014	**0.075*****	0.080	**−0.019****	−0.025	0.049	0.054	**0.162*****	0.188	**−0.260*****	−0.301
SSQ	−0.164	−0.020	0.087	0.006	1.328	0.102	**−1.114*****	−0.088	**0.694*****	0.079	**−1.817****	−0.200
BORG RPE	0.179	0.022	**1.000*****	0.043	**1.786*****	0.062	**−4.429*****	−0.140	0.321	0.010	2.214	0.094
Joy	−0.031	−0.001	−0.209	−0.085	0.265	0.123	−0.407	−0.168	**0.466*****	0.179	−0.271	−0.115
Fear	**0.028*****	0.104	−0.098	−0.044	−0.005	−0.004	0.489	0.187	−0.251	−0.179	−0.004	−0.002
Relaxation	**0.109***	0.067	−0.012	−0.005	0.305	0.142	0.275	0.125	−0.123	−0.049	−0.102	−0.044
Boredom	−0.029	−0.009	0.091	0.051	−0.406	−0.168	0.222	0.110	−0.272	−0.137	−0.220	−0.107
Presence	−0.036	−0.005	−0.280	−0.133	0.377	0.193	−0.133	−0.070	0.111	0.054	−0.254	−0.118

Table 7: Linear relationships between emotion and presence ratings and physiological measures for RetroSketch and ESM, expressed as standardised linear regression coefficients β. If either RetroSketch or ESM has a significantly stronger relationship with a physiological measure, the respective cells are highlighted. The difference in the strength of the relationships is quantified as Δd, with a positive Δd indicating a stronger relationship with RetroSketch. The physiological measures are pupil dilation level (PDL) and response (PDR), skin conductance level (SCL) and response (SCR), heart rate (HR), heart rate variability (HRV), blink rate (BR), blink duration (BD), zygomaticus major activity (Smile), and orbicularis oris activity (O-Shape).

DV	Method	PDL		PDR		SCL		SCR		HR		HRV		BR		BD		Smile		O-Shape	
		β	Δ d	β	Δ d	β	Δ d	β	Δ d	β	Δ d	β	Δ d	β	Δ d	β	Δ d	β	Δ d	β	Δ d
Joy	Retro	−.025	0.005	−.006	0.005	−.003	−0.011	−.007	−0.008	.054	0.053	.015	**−0.100**	.016	−0.002	.014	−0.020	−.027	−.008	.009	−0.009
	ESM	.012		.010		.008		−.014		.080		**−.053***		−.009		−.041		−.035		.014	
Fear	Retro	.098	0.009	.074	**−0.055**	−.002	**−0.159**	**.062***	**0.123**	.019	**−0.099**	**.038***	**0.106**	−.028	.005	.024	−0.004	.048	**−0.071**	.033	0.004
	ESM	.117		**.119***		**.026****		−.009		**.056****		.069		−.036		.030		**.151*****		.032	
Relax	Retro	−.035	**−0.086**	−.023	0.002	−.022	−0.001	−.027	−0.010	−.024	−0.006	−.010	−0.036	.052	0.024	.052	0.028	−.011	−0.007	.031	0.041
	ESM	**−.075***		−.020		−.022		−.008		−.024		−.040		.023		−.024		−.071		−.003	
Bored	Retro	**−.084****	**0.025**	−.058	0.003	.015	0.042	.000	−0.005	−.033	**−0.073**	−.040	**−0.062**	.018	−0.001	.021	0.001	−.020	0.004	.019	−0.011
	ESM	−.058		−.060		.000		.011		**−.062****		**.007***		−.016		−.015		−.002		−.023	
Pres.	Retro	.038	0.029	−.051	−0.040	−.006	−0.006	−.073	−0.022	.045	**−0.001**	.020	0.007	−.001	−0.027	−.084	**−0.064**	.020	0.002	−.008	0.002
	ESM	−.005		−.078		.009		−.081		**.077****		.005		.050		**.001***		−.015		.001	

models using the `nlme` package [21], which can take advantage of our repeated measures [36, 167]. Although the regression models cannot be used to test the relationships directly due to violations of normality, the models are still valid descriptions of the linear relationships and we can compare their goodness of fit using encompassing tests [45] from the `lmtest` package [176]. An encompassing test can detect if either RetroSketch or ESM explains significantly less variance in a physiological measure, potentially allowing us to identify a 'winner'. Based on the explained variances of each model, we calculate Cohen's f^2 and corresponding d values (see Supplementary Material Document for details) and use the difference Δd between the two d values to quantify how much one model is better than the other (RetroSketch vs. ESM).

The results in Table 7 show, for example, that RetroSketch Fear explained SCR and HRV significantly better than ESM Fear. In contrast, ESM Fear better explained PDR, SCL and Smile. Overall, the regression results indicate that RetroSketch and ESM measures bear similar relationships with physiological measures. The encompassing tests revealed only a few cases where either RetroSketch or ESM better explained physiological measures, and in all cases, the effect sizes of these differences were 'very small' or 'tiny'. When considering only significant differences with at least 'very small' effect sizes ($d > 0.1$), ESM is slightly superior in three cases whereas RetroSketch is slightly superior in two. RetroSketch explained more variance in phasic physiological measures (i.e. those related to quick changes) such as SCR and HRV, while ESM explained more variance in tonic measures (i.e. those related to slow changes in baseline) such as SCL.

We performed ordinal logistic regressions to determine whether the relationships between RetroSketch and ESM measures and physiological measures were influenced by the *VR Game*, using interaction tests with Holm-Bonferroni posthoc correction. There were very few significant interactions for both RetroSketch and ESM. However, we observed consistent interactions for several ESM measures – such as Fear, Relaxation, Boredom, and Presence – and HR, which varied significantly depending on the VR experience. A detailed breakdown of these interactions can be found in the Supplementary Material Document.

Table 8: Correlations between RetroSketch and ESM ratings and sentiment scores using Pearson's r and Kendall's τ with significance tests performed on τ. If either RetroSketch or ESM has a significantly stronger relationship with sentiment scores, the respective regression coefficients β are marked with *. The difference in the strength of the relationships is quantified as Δd, with a positive Δd indicating a stronger relationship with RetroSketch. Some rows describe significant interactions of such relationships with specific VR games (e.g. HLA and RM for Joy); in these rows Δd describes the size of the interaction effect.

DV	Method	Game	r	τ	β	Δd
Joy	*Retro*	ALL	.499	**.335*****	**.478*****	**.540**
	ESM	ALL	.355	**.242***	.266	
	ESM	*HLA – RM*				**0.162***
	ESM	*HLA – GotS*				**0.356***
Fear	*Retro*	ALL	−.069	−.063	−.149	.119
	ESM	ALL	−.092	−.049	−.100	
Relaxation	*Retro*	ALL	.342	.227***	**.322*****	.153
	ESM	ALL	.195	**.155***	.236	
	Retro	*ACC – GotS*				**0.189*****
	ESM	*HLA – RM*				**0.153****
Boredom	*Retro*	ALL	−.196	−.131	**−.279*****	0.252
	ESM	ALL	−.096	−.060	−.093	
Presence	*Retro*	ALL	.199	**.153****	.172	0.200
	ESM	ALL	.070	.053	.112	
	ESM	*GotS – IEYTD*				**0.380*****

5.7 RQ5: How do RetroSketch & ESM relate to qualitative measures?

To answer RQ5, we analysed the qualitative annotations participants made about salient moments during their VR experiences using RetroSketch. Each participant created on average 20 annotations, resulting in 2,799 annotations on 280 sketches. The annotations were analysed using a Twitter-roBERTa-base model fine-tuned for sentiment analysis [30, 130, 131], which we evaluated in the context of RetroSketch in subsection 5.2. We correlated the sentiment scores with RetroSketch and ESM ratings using Pearson's r and Kendall's τ, with significance tests performed on Kendall's τ due to non-normality. Analogously to our approach for RQ4, we described the linear relationships between RetroSketch and the sentiment scores, and ESM and the sentiment scores, respectively, with regression coefficients β. We compared the two relationships with encompassing tests, and quantified the difference between the strengths of the two relationships as Δd.

Table 8 shows both RetroSketch Joy and ESM Joy correlated moderately with the sentiment scores. Furthermore, RetroSketch Relaxation correlated moderately with sentiment, whereas ESM correlations, though significant, were weaker. Additionally, RetroSketch Presence was significantly positively correlated with sentiment, whereas ESM Presence was not. Interestingly, negative valence measures such as Boredom and Fear did not significantly correlate with sentiment for both RetroSketch and ESM. The encompassing tests and Δd values indicate that RetroSketch consistently explained more variance in sentiment than ESM, with small to medium effects.

Lastly, we performed ordinal logistic regressions to determine whether the relationships between RetroSketch and ESM measures and sentiment scores were influenced by the *VR Game*, using interaction tests with Holm-Bonferroni posthoc correction, as presented in Table 8. For these analyses, Δd describes the size of an interaction effect. There was only one significant interaction for RetroSketch and three for ESM, all with 'small' effects. This suggests that the correlations between RetroSketch and ESM measures and sentiment scores are fairly robust across different VR experiences.

5.8 User Feedback on RetroSketch and ESM

We examined participant responses to open-ended questions regarding their experiences with RetroSketch and ESM, which included their methodological preferences. Overall, users' preferences were split between ESM and RetroSketch (RetroSketch: 44.29%, ESM: 42.86%, No preference: 14.29%), highlighting both the strengths and limitations in the perceived accuracy of each method. Their responses were analysed and deductively grouped into themes [35] to better capture the reasoning behind preferences for either method.

ESM:. Views on ESM were divided, with over a third expressing mixed opinions. On the positive side, 54 participants felt that being asked questions during gameplay allowed them to accurately recognise their emotions in the moment and respond to them in *"a very natural manner"* (P88). However, 25 participants felt that the questioning caused them to disengage, especially in highly engaging games (*"jarring and a little distracting"*, P29). 86 participants felt the questioning during gameplay was disruptive and frustrating, which could result in missed key moments and incomplete capture of fluctuating emotions. 13 participants mentioned that multitasking affected their focus or the accuracy of their responses, while 10 participants talked about ESM as a study reminder ("*it was a reminder that I was in a study throughout, so I felt a bit less immersed*", P5). One participant noted "*I couldn't give extreme values which I would've otherwise given without someone else hearing me cuz I tend to conform quite a lot*" (P89), which points to a social desirability bias. In contrast, 32 participants found ESM not disruptive ("*I don't believe it influenced or interrupted my feeling as it was focussing on the experience itself*", P3).

RetroSketch: 75 participants appreciated the ability to rewatch their experience, which helped them look at the broader picture and recall key moments ("*I was able to look through and remember what I felt in the moment without being interrupted*", P8). 6 participants expressed difficulty in recalling their entire feelings for the whole gameplay session ("*the key parts definitely felt more memorable, but the parts in between less so*", P37), but participants also expressed that ratings were "*still in the right area*" (P5). 12 participants mentioned difficulty recalling presence "*when you are not in the environment*" (P72), whereas 19 participants suggested that the short duration between the experience and RetroSketch helped with recall ("*It was easy because it was just after the gameplay*", P23). 26 participants mentioned difficulties quantifying multiple feelings within a limited amount of time ("*you're limited in the amount of accurate information you can give when just watching a video on a normal computer screen*", P13).

97 participants found RetroSketch user friendly and effective in recalling emotions ("*it is quite good a way to express feeling based on time and the video*", P70). However, 13 participants talked about video control challenges ("*I also really wanted to use the 2.5x option but that just made the footage rough*", P3). 7 participants mentioned challenges with the interface layout. 27 participants described challenges sketching lines ("*the line tool was a bit hard to use when drawing really steep lines*", P57), with other participants suggesting "*the addition of being able to input an exact number*" (P1). RetroSketch was generally perceived as usable (97 participants) and accurate (74 participants).

6 DISCUSSION

RetroSketch demonstrated both internal and external validity in collecting quantitative and qualitative measures of emotions and presence across a wide range of VR experiences. In this section, we discuss the answers to our research questions in more depth and discuss future work on continuous emotion measurement.

6.1 RQ1: How do RetroSketch measures relate to ESM measures?

RetroSketch and ESM Strongly Correlate: We found a significant correlation between RetroSketch and ESM scores across all dependent variables that is consistent across different VR experiences. This, along with the majority of users reporting ease in recalling their emotions via RetroSketch, supports the validity of RetroSketch as an appraisal-based emotion measurement method and alternative to ESM without needing to disrupt the experience. These correlations exist despite RetroSketch being administered approximately 10-15 minutes after the VR experience, not immediately afterwards due to the natural break in the study for questionnaires. A key feature which enabled this was the flexibility provided by the video-aided recall that allowed participants to reflect on the 'bigger picture' of the VR experience. These findings also bolster the growing literature validating retrospective emotional appraisal as a reliable method for measuring emotions [134, 142, 198].

RetroSketch and ESM Have Key Differences: Despite the strong correlation, we observed significant differences in the distribution characteristics between RetroSketch and ESM measures. RetroSketch typically captured a broader range and variation of emotions and presence compared to ESM, possibly because its open graphing format encourages users to depict emotions more dynamically. However, in VR experiences with weak narrative elements, like *GotS* and *ACC*, RetroSketch often showed emotional flatlines. RetroSketch encourages participants to reflect on specific, salient events in the overall context of an experience. In contrast, ESM may lead to middle-scale responses due to central-tendency bias, a common effect observed in the literature [48, 115, 263]. RetroSketch's continuous, appraisal-based approach may encourage users to capture the ebb and flow of 'simmering' emotions [135], including "different and conflicting emotions from the same event" [151].

ESM Reminds Users They Are Being Observed: RetroSketch consistently recorded lower 'positive' emotions and presence, and higher 'negative' emotions compared to ESM. This may result from social desirability bias [255] in ESM, where participants are reminded of the study. While RetroSketch is not immune to such bias, it allows users to self-report in a more private, less pressured setting. Additionally, interaction results from Table 3 reveal significant differences in distribution characteristics between RetroSketch and ESM, particularly for less dominant emotions in each VR experience. For instance, *GotS* was least associated with fear overall, yet showed a significant distribution difference for Fear, similar to Boredom in *IEYTD*. This suggests that ESM may be more prone to impulsive fluctuation, whereas RetroSketch may be better attuned to the overall emotional characteristics and context of a game.

6.2 RQ2: How reliable is RetroSketch & ESM across different VR experiences and users?

The relationships between RetroSketch and ESM remained robust across various user demographics, personalities, player types and immersive tendencies. Few significant interactions with covariates were observed, and even users' preferences for using ESM over RetroSketch did not significantly influence the correlations. As shown in Table 4 and Table 5, covariates generally influence RetroSketch and ESM similarly, e.g. both Presence measures correlate with immersive tendencies (ITQ) [266], both Joy measures correlate with Tondello Challenge traits [248], and more agreeable individuals rate the experience as more relaxing with both measures [78].

For example, we observed a moderate effect of gender on ESM Joy, with males reporting higher ratings, whereas no significant gender effect was found for RetroSketch Joy. This could be due to social desirability bias in ESM [255], particularly self-enhancement bias [217], where males may want to appear as being more successful, and therefore enjoying the experience more. Alternatively, previous studies have shown gender differences in self-reported emotions [63], including in the context of video games [108], which ESM might capture.

Additionally, an expected negative correlation was observed between RetroSketch Fear and Extroversion [19, 98], but not for ESM, suggesting that less extroverted individuals do not always report higher Fear, possibly due to social desirability bias [255]. Interestingly, while we would expect Openness to correlate positively with Fear, no significant correlation was found with RetroSketch, and a significant negative correlation was observed with ESM. Overall, the influence of personality traits on either measure is minimal, with both measures generally aligning well.

Previous work has shown evidence for a positivity effect in older adults' emotional recall [149, 253]. While we do not collect data from older adults (65+), age did not affect RetroSketch or ESM measures in the context of VR games, with no significant correlations across the four emotions. Age showed a significant positive correlation with RetroSketch presence that was not observed for ESM presence, and indeed, some prior literature suggests age correlates with presence [127]. RetroSketch may be capturing real differences in presence ratings across age demographics, however, the observed correlation was weak and other previous work contests whether age influences presence in VR [61, 129].

6.3 RQ3: How does ESM influence the VR user experience?

While the effects of ESM on the VR experience are *measurable*, they are also seemingly *unpredictable*, depending heavily on the VR experience. For instance, participants generally experienced higher pressure with ESM (Table 6), likely due to increased cognitive load. However, in more cognitively demanding experiences like *HLA* and *IEYTD*, ESM actually reduced pressure, likely because it provided artificial breaks during intense moments. *ACC* is a notable exception, as the experience is intense but involves driving, where holding a conversation is common.

Another interesting finding is the inconsistent effects of ESM on Presence [133]. The 'immediate reflection' [225] required by ESM had varying impacts on Presence depending on the VR context and user activity. Overall increased social presence could result from ESM being administered through a human voice, however, in *ACC* ESM affected all three Presence measures negatively. The most consistent outcome was the negative impact of ESM on Flow Absorption [132], which is supported by qualitative feedback that ESM was disruptive and disengaging. The largest effects of ESM were observed in *RM*, particularly on intrinsic motivation, flow and physical presence. *RM* has a focus on open exploration and environmental storytelling [102], and ESM may disrupt immersion in such experiences. However, the effects of ESM were overall small.

6.4 RQ4: How do RetroSketch & ESM relate to physiological measures?

Both RetroSketch and ESM showed similar relationships with common physiological measures of emotion (Table 7), with only 'very small' to 'tiny' differences. Affective computing and emotion recognition approaches [29, 88, 137, 169] usually model the relationship between emotions and physiological signals based on subjective 'ground truth' data. RetroSketch is well placed to provide such data continuously in high resolution for multiple emotion variables, and could be particularly useful when subjective ground truths cannot be captured in the moment or immediately after, making retrospective appraisal necessary.

According to subsection 5.6, tonic measures, which reflect gradual shifts in physiological baselines, may drive emotional measures more in ESM because these physiological responses are more readily perceptible by the user. According to Barrett's Constructed Emotion Theory [12], "*users make meaning of physical responses, based on context and prior experience, before they know what emotion is attached to the situation*".

While both ESM and RetroSketch are likely influenced by users' interoceptive awareness [50, 142], this influence may be more immediate in ESM. This is supported by the many interactions observed across different games when relating HR measures to ESM emotions. These interactions highlight that the interpretation and association of one's HR to different emotions is highly context dependent [12], e.g. elevated HR in *HLA* may be more associated with fear while in *ACC* it may be associated with joy or excitement.

RetroSketch explained more variance in phasic responses such as SCR and HRV. Phasic responses are typically tied to specific events and are only perceptible in certain moments, which RetroSketch can capture due to its continuous and granular data collection. This further supports the validity of using appraisal and recall-based approaches for emotion measurement [134, 192, 198].

6.5 RQ5: How do RetroSketch & ESM relate to qualitative measures?

Both RetroSketch and ESM Joy scores significantly correlated with the sentiment of annotations made using RetroSketch. Unsurprisingly, RetroSketch showed a stronger correlation than ESM, validating the use of sentiment analysis [126, 160, 225] on user-generated annotations. Compared to other sentiment-based approaches, RetroSketch provides a rich combination of video footage, temporally anchored qualitative annotations and quantitative ratings.

When reflecting on users' preferences for RetroSketch or ESM, we observed an almost 50/50 split. This divide is surprising, considering that RetroSketch typically takes longer to use and requires more effort from the user. A deeper look into the reasons revealed that some users felt they could not accurately reflect on their emotions or presence in the moment of a VR experience. Conversely, others mentioned that they struggled to recall their emotions or presence when using RetroSketch, even with video-aided recall. This points to interpersonal differences in both immediate reflection and emotional recall. User preference did not significantly affect the correlation between RetroSketch and ESM, indicating robustness of RetroSketch irrespective of preference.

6.6 Limitations & Future Work

Generalizability & Scalability: RetroSketch was successfully applied in both paper form during the pilot study and digital form in the main study, specifically within the context of VR games. While this has demonstrated promising results, there remain opportunities to explore the application of RetroSketch in other VR experiences such as training scenarios, education, therapeutic environments, and non-interactive storytelling like immersive films. Although we did extend our exploration of RetroSketch to high-end immersive simulators commonly used in training (motion platform combined with *ACC*), a formal evaluation in this context is warranted.

Further investigation should determine the best practices for RetroSketch, including how it can be scaled to measure experiences of different durations, how to optimise its interface to reduce complexity and user workload, and what other emotions can be captured by RetroSketch such as direct measures of valence and arousal. We used RetroSketch to appraise 30-minute VR sessions, which took between 15 and 40 minutes depending on the participant. This raises questions about the upper duration limit where RetroSketch can still provide reliable appraisals, and whether RetroSketch can capture highly granular data during brief emotional events.

RetroSketch could be used to explore in-person player dyads or online multiplayer games, where interpersonal relationships influence emotional responses in players [183]. Beyond VR, RetroSketch holds potential for application in Augmented Reality (AR) and even non-immersive experiences as a general emotional appraisal tool. However, questions remain about how to reliably facilitate video-aided recall across different contexts of use.

Emotional Recall and Appraisal: In our study, participants completed questionnaires and were given an opportunity for a short break after a VR session, which took approximately 10-15 minutes, before using RetroSketch.. While this was effective, it raises questions about how emotional recall with RetroSketch might change over time and how this could influence the appraisal process. Moreover, there are a multitude of emotional models beyond appraisal, and RetroSketch should be evaluated in light of these other theories, e.g. considering the Facial Feedback Hypothesis [37].

Do RetroSketch measures retain their robustness when applied days, weeks or even months after an experience? Can RetroSketch be reliably utilised for multiple appraisal sessions of the same experience? Exploring these possibilities could significantly extend the utility of RetroSketch. Additionally, previous research has considered incorporating physiological measurements during the appraisal process [89], which could also be promising for RetroSketch.

Beyond Video-Aided Recall: While RetroSketch was effectively used to appraise VR experiences through video-aided recall, there is potential for incorporating more immersive elements to enhance the recall process. Building on the work of continuous emotion measurement for 360° video in VR [269, 270], VR itself could be leveraged as the platform for retrospection. For instance, enabling users to relive VR experiences and events within the VR environment itself could offer a richer appraisal experience. With the findings of this paper and the existing body of work on video-based emotional measurement and annotation tools [67, 269], the advantages of using immersive over non-immersive retrospection could be explored and formalised.

While this merits further investigation, it also presents several challenges. Replaying events or experiences in the 3D world may not be technically feasible (e.g. if replay features are not built into the experience), and replaying 3D experiences from a first-person perspective could induce simulator sickness, especially in scenarios involving locomotion [51].

6.7 Recommendations for RetroSketch

Based on our findings, we make the following recommendations for using RetroSketch to measure emotions and presence in VR experiences:

(1) RetroSketch is more suitable than ESM for capturing continuous data at high resolution. However, if this is not required then consider the ease and simplicity of ESM.
(2) RetroSketch is particularly relevant for experiences that elicit a variety of fluctuating emotions.
(3) RetroSketch is available as both a digital and paper tool both of which showed strong correlations with ESM. However, the paper-based RetroSketch has not been as thoroughly evaluated so should be used with caution.
(4) Be careful when using RetroSketch for experiences longer than 30 minutes because the time required for retrospection increases with the duration of the experience, and we have not yet explored the limits of RetroSketch.
(5) Allow users the freedom to use RetroSketch as they see fit.
(6) Setting a minimum number of key points can set user expectations for the level of detail required and can be used to ensure good coverage across the experience. However, be careful to avoid demand characteristics.
(7) Allow for at least half the time of the session duration and allow additional time if users need it to complete their sketch.
(8) If time is limited, or RetroSketch can not be administered soon after the experience, consider using other methods such as ESM or standard questionnaires.
(9) Provide instructions about how RetroSketch works and highlight the importance of annotated keypoints to help contextualise and understand responses to events.
(10) Sentiment analysis models can be used to quantify large amounts of annotations for easier analysis.

6.8 Impact

RetroSketch advances the measurement of emotion and presence for VR experiences by offering an appraisal-based approach that provides highly granular and continuous data for categorical emotions, core affect and presence. It also highlights salient events and provides temporally anchored qualitative annotations. RetroSketch can be applied in research, VR design, and user experience testing to better understand the impact of design choices, e.g. conducting more granular comparisons of presence in immersive experiences or measuring specific emotional effects of individual events and characters in VR. The strong correspondence between RetroSketch and physiological measures suggests that the fine-grained emotion data provided by RetroSketch can be used as ground truth in the development of affective systems (e.g. for emotion recognition) providing an alternative to ESM.

7 CONCLUSION

We presented RetroSketch, a retrospective method for measuring emotions and presence in VR experiences. We evaluated RetroSketch in a large VR user study ($n = 140$), comparing it to state-of-the-art methods including ESM and physiological sensing (dataset found here [172]). We validated RetroSketch across five different

VR experiences, which participants played for one hour in two 30-minute sessions. This led us to the following conclusions:

(1) RetroSketch can be used to measure emotions and presence continuously in VR experiences.
(2) RetroSketch correlates strongly and robustly with ESM across various VR experiences.
(3) RetroSketch shows correspondences with physiological measures of emotion similar to ESM and can be used for the development of emotion recognition systems.
(4) ESM influences the VR user experience in small but seemingly unpredictable ways.
(5) RetroSketch annotations relating to salient events provide time-anchored qualitative data that can be analysed automatically with sentiment analysis models.

Our findings support the use of RetroSketch for measuring emotions and presence in VR providing that there is sufficient time for the user to navigate, recall, and reflect on their experience.

ACKNOWLEDGMENTS

We would like to acknowledge Jovana Cuxac, Jasmine Chung, Katherine Baldrian, and Tarini Sehgal for their support and assistance in conducting the study. 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/), as well as the Royal Society Industry Fellowship under grant agreement *No INF\R1\221058*.

REFERENCES

[1] 2021. https://www.anses.fr/en/content/what-are-risks-virtual-reality-and-augmented-reality-and-what-good-practices-does-anses
[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] Devon Allcoat and Adrian von Mühlenen. 2018. Learning in virtual reality: Effects on performance, emotion and engagement. *Research in Learning Technology* 26 (2018).
[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] 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.
[7] 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
[8] 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
[9] Kenneth Baclawski. 2018. The observer effect. In *2018 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, 83–89.
[10] Rosa M Baños, Cristina Botella, Isabel Rubió, Soledad Quero, Azucena García-Palacios, and Mariano Alcañiz. 2008. Presence and emotions in virtual environments: The influence of stereoscopy. *CyberPsychology & Behavior* 11, 1 (2008), 1–8.
[11] 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.
[12] Lisa Feldman Barrett. 2016. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience* 12, 1 (10 2016), 1–23. https://doi.org/10.1093/scan/nsw154 arXiv:https://academic.oup.com/scan/article-pdf/12/1/1/27103603/nsw154.pdf
[13] Lisa Feldman Barrett and Christiana Westlin. 2021. Navigating the science of emotion. In *Emotion measurement*. Elsevier, 39–84.
[14] Thomas Baumgartner, Lilian Valko, Michaela Esslen, and Lutz Jäncke. 2006. Neural correlate of spatial presence in an arousing and noninteractive virtual reality: an EEG and psychophysiology study. *CyberPsychology & Behavior* 9, 1 (2006), 30–45.
[15] Alejandro Beacco, Ramon Oliva, Carlos Cabreira, Jaime Gallego, and Mel Slater. 2021. Disturbance and plausibility in a virtual rock concert: A pilot study. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 538–545.
[16] Mattan S Ben-Shachar, Daniel Lüdecke, and Dominique Makowski. 2020. effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software* 5, 56 (2020), 2815.
[17] Stuart M Bender and Billy Sung. 2021. Fright, attention, and joy while killing zombies in virtual reality: A psychophysiological analysis of VR user experience. *Psychology & Marketing* 38, 6 (2021), 937–947.
[18] 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.
[19] Adam Bibbey, Douglas Carroll, Tessa J. Roseboom, Anna C. Phillips, and Susanne R. de Rooij. 2013. Personality and physiological reactions to acute psychological stress. *International Journal of Psychophysiology* 90, 1 (2013), 28–36. https://doi.org/10.1016/j.ijpsycho.2012.10.018 Blunted Cardiovascular Reactivity - What Does It Mean?.
[20] Pauline Bimberg, Tim Weissker, and Alexander Kulik. 2020. On the usage of the simulator sickness questionnaire for virtual reality research. In *2020 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*. IEEE, 464–467.
[21] Paul Bliese. 2006. Multilevel Modeling in R (2.2)–A Brief Introduction to R, the multilevel package and the nlme package.
[22] Francesca Borghesi, Vittorio Murtas, Valentina Mancuso, and Alice Chirico. 2023. Continuous Time Elicitation Through Virtual Reality to Model Affect Dynamics. In *Computer-Human Interaction Research and Applications*, Hugo Plácido da Silva and Pietro Cipresso (Eds.). Springer Nature Switzerland, Cham, 258–276.
[23] Danny Oude Bos et al. 2006. EEG-based emotion recognition. *The influence of visual and auditory stimuli* 56, 3 (2006), 1–17.
[24] Stéphane Bouchard, Julie St-Jacques, Geneviève Robillard, and Patrice Renaud. 2008. Anxiety increases the feeling of presence in virtual reality. *Presence: Teleoperators and Virtual Environments* 17, 4 (2008), 376–391.
[25] 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
[26] 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.
[27] 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.
[28] 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.
[29] Rafael A Calvo and Sidney D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing* 1, 1 (2010), 18–37.
[30] Jose Camacho-collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, Eugenio Martínez Cámara, et al. 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Abu Dhabi, UAE, 38–49. https://aclanthology.org/2022.emnlp-demos.5
[31] Laura L Carstensen, Bulent Turan, Susanne Scheibe, Nilam Ram, Hal Ersner-Hershfield, Gregory R Samanez-Larkin, Kathryn P Brooks, and John R Nesselroade. 2011. Emotional experience improves with age: evidence based on over 10 years of experience sampling. *Psychology and aging* 26, 1 (2011), 21.

[32] 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.

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

[34] Tamlin Conner Christensen, Lisa Feldman Barrett, Eliza Bliss-Moreau, Kirsten Lebo, and Cynthia Kaschub. 2003. A practical guide to experience-sampling procedures. *Journal of Happiness Studies* 4, 1 (2003), 53–78.

[35] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology* 12, 3 (2017), 297–298.

[36] 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.

[37] Nicholas A Coles, David S March, Fernando Marmolejo-Ramos, Jeff T Larsen, Nwadiogo C Arinze, Izuchukwu LG Ndukaihe, Megan L Willis, Francesco Foroni, Niv Reggev, Aviv Mokady, et al. 2022. A multi-lab test of the facial feedback hypothesis by the Many Smiles Collaboration. *Nature human behaviour* 6, 12 (2022), 1731–1742.

[38] Géraldine Coppin and David Sander. 2021. Theoretical approaches to emotion and its measurement. In *Emotion measurement*. Elsevier, 3–37.

[39] Neat Corporation. 2022. Garden of the Sea (VR) on Steam. https://store.steampowered.com/app/1086850/Garden_of_the_Sea_VR/

[40] 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.

[41] Marci D Cottingham and Rebecca J Erickson. 2020. Capturing emotion with audio diaries. *Qualitative Research* 20, 5 (2020), 549–564.

[42] Mihaly Csikszentmihalyi and Reed Larson. 1987. Validity and reliability of the experience-sampling method. *The Journal of nervous and mental disease* 175, 9 (1987), 526–536.

[43] Mihaly Csikszentmihalyi and Reed Larson. 2014. *Validity and Reliability of the Experience-Sampling Method*. Springer Netherlands, Dordrecht, 35–54. https://doi.org/10.1007/978-94-017-9088-8_3

[44] 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

[45] R Davidson. 1993. Estimation and Inference in Econometrics.

[46] Silvia de Haan-Rietdijk, Manuel C. Voelkle, L. Keijsers, and Ellen L. Hamaker. 2017. Discrete- vs. continuous-time modeling of unequally spaced experience sampling method data. *Frontiers in Psychology* 8 (2017). https://doi.org/10.3389/fpsyg.2017.01849

[47] Julia Diemer, Georg W Alpers, Henrik M Peperkorn, Youssef Shiban, and Andreas Mühlberger. 2015. The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Frontiers in psychology* 6 (2015), 26.

[48] Igor Douven. 2018. A Bayesian perspective on Likert scales and central tendency. *Psychonomic bulletin & review* 25 (2018), 1203–1211.

[49] Ilana Dubovi. 2022. Cognitive and emotional engagement while learning with VR: The perspective of multimodal methodology. *Computers & Education* 183 (2022), 104495.

[50] Barnaby D Dunn, Hannah C Galton, Ruth Morgan, Davy Evans, Clare Oliver, Marcel Meyer, Rhodri Cusack, Andrew D Lawrence, and Tim Dalgleish. 2010. Listening to your heart: How interoception shapes emotion experience and intuitive decision making. *Psychological science* 21, 12 (2010), 1835–1844.

[51] Natalia Dużmańska, Paweł Strojny, and Agnieszka Strojny. 2018. Can simulator sickness be avoided? A review on temporal aspects of simulator sickness. *Frontiers in psychology* 9 (2018), 2132.

[52] Maria Egger, Matthias Ley, and Sten Hanke. 2019. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science* 343 (2019), 35–55.

[53] Paul Ekman. 1992. Are there basic emotions?. *Psychological Review* 99 (1992), 550–553. https://search.ebscohost.com/login.aspx?direct=true&db=eue&AN=508429547&site=ehost-live

[54] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 754–768. https://doi.org/10.1145/3472749.3474784

[55] Phoebe C Ellsworth and Klaus R Scherer. 2002. Appraisal Processes In Emotion. *Handbook of Affective Sciences* (12 2002), 572–595. https://doi.org/10.1093/oso/9780195126013.003.0029

[56] Paul MG Emmelkamp and Katharina Meyerbröker. 2021. Virtual reality therapy in mental health. *Annual review of clinical psychology* 17, 1 (2021), 495–519.

[57] Sergio Estupiñán, Francisco Rebelo, Paulo Noriega, Carlos Ferreira, and Emília Duarte. 2014. Can virtual reality increase emotional responses (Arousal and Valence)? A pilot study. In *Design, User Experience, and Usability. User Experience Design for Diverse Interaction Platforms and Environments: Third International Conference, DUXU 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part II 3*. Springer, 541–549.

[58] Kirill A Fadeev, Alexey S Smirnov, Olga P Zhigalova, Polina S Bazhina, Alexey V Tumialis, and Kirill S Golokhvast. 2020. Too real to be virtual: Autonomic and EEG responses to extreme stress scenarios in virtual reality. *Behavioural neurology* 2020, 1 (2020), 5758038.

[59] Fanatec. 2024. Fanatec Direct Drive Wheel. url=https://fanatec.com/eu-en/racing-wheels-direct-drive-bases/direct-drive-bases/podium-wheel-base-dd2.

[60] Martin Feick, Niko Kleer, Anthony Tang, and Antonio Krüger. 2020. The Virtual Reality Questionnaire Toolkit. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 68–69. https://doi.org/10.1145/3379350.3416188

[61] Anna Felnhofer, Oswald D. Kothgassner, Nathalie Hauk, Leon Beutl, Helmut Hlavacs, and Ilse Kryspin-Exner. 2014. Physical and social presence in collaborative virtual environments: Exploring age and gender differences with respect to empathy. *Computers in Human Behavior* 31 (2014), 272–279. https://doi.org/10.1016/j.chb.2013.10.045

[62] Anna Felnhofer, Oswald D Kothgassner, Mareike Schmidt, Anna-Katharina Heinzle, Leon Beutl, Helmut Hlavacs, and Ilse Kryspin-Exner. 2015. Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios. *International journal of human-computer studies* 82 (2015), 48–56.

[63] Agneta H. Fischer, Mariska E. Kret, and Joost Broekens. 2018. Gender differences in emotion perception and self-reported emotional intelligence: A test of the emotion sensitivity hypothesis. *PLOS ONE* 13, 1 (01 2018), 1–19. https://doi.org/10.1371/journal.pone.0190712

[64] Allison S Gabriel, Nathan P Podsakoff, Daniel J Beal, Brent A Scott, Sabine Sonnentag, John P Trougakos, and Marcus M Butts. 2019. Experience sampling methods: A discussion of critical trends and considerations for scholarly advancement. *Organizational Research Methods* 22, 4 (2019), 969–1006.

[65] Schell Games. 2017. I Expect You To Die on Steam. https://store.steampowered.com/app/587430/I_Expect_You_To_Die/

[66] Asghar Ghasemi and Saleh Zahediasl. 2012. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism* 10, 2 (2012), 486.

[67] Jeffrey M Girard. 2014. CARMA: Software for continuous affect rating and media annotation. *Journal of open research software* 2, 1 (2014).

[68] Lewis R Goldberg. 1992. The development of markers for the Big-Five factor structure. *Psychological assessment* 4, 1 (1992), 26.

[69] Lewis R Goldberg. 1992. Possible questionnaire format for administering the 50-item set of IPIP Big-Five factor markers. *Psychol. Assess* 4 (1992), 26–42.

[70] 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.

[71] 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.

[72] 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.

[73] John M Gottman and Robert W Levenson. 1985. A valid procedure for obtaining self-report of affect in marital interaction. *Journal of consulting and clinical psychology* 53, 2 (1985), 151.

[74] Sarah Graf and Valentin Schwind. 2020. Inconsistencies of Presence Questionnaires in Virtual Reality. In *Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology* (Virtual Event, Canada) *(VRST '20)*. Association for Computing Machinery, New York, NY, USA, Article 60, 3 pages. https://doi.org/10.1145/3385956.3422105

[75] Marco Granato, Davide Gadia, Dario Maggiorini, and Laura A Ripamonti. 2020. An empirical study of players' emotions in VR racing games based on a dataset of physiological data. *Multimedia tools and applications* 79, 45 (2020), 33657–33686.

[76] Simone Grassini and Karin Laumann. 2020. Questionnaire measures and physiological correlates of presence: A systematic review. *Frontiers in psychology* 11 (2020), 349.

[77] Simone Grassini, Karin Laumann, and Martin Rasmussen Skogstad. 2020. The use of virtual reality alone does not promote training performance (but sense of presence does). *Frontiers in psychology* 11 (2020), 1743.

[78] William G Graziano, Lauri A Jensen-Campbell, and Elizabeth C Hair. 1996. Perceiving interpersonal conflict and reacting to it: the case for agreeableness. *Journal of personality and social psychology* 70, 4 (1996), 820.

[79] Pamela Grimm. 2010. Social desirability bias. *Wiley international encyclopedia of marketing* (2010).

[80] Daniel Gromer, Max Reinke, Isabel Christner, and Paul Pauli. 2019. Causal Interactive Links Between Presence and Fear in Virtual Reality Height Exposure. *Frontiers in Psychology* 10 (2019). https://doi.org/10.3389/fpsyg.2019.00141

[81] 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.

[82] 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.

[83] Deborah L Harm. 2002. Motion sickness neurophysiology, physiological correlates, and treatment. In *Handbook of virtual environments*. CRC Press, 677–702.

[84] Jonathon D. Hart, Thammathip Piumsomboon, Louise Lawrence, Gun A. Lee, Ross T. Smith, and Mark Billinghurst. 2018. Emotion Sharing and Augmentation in Cooperative Virtual Reality Games. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts* (Melbourne, VIC, Australia) *(CHI PLAY '18 Extended Abstracts)*. Association for Computing Machinery, New York, NY, USA, 453–460. https://doi.org/10.1145/3270316.3271543

[85] 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.

[86] 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.

[87] Richard Held. 1992. Telepresence. *The Journal of the Acoustical Society of America* 92, 4_Supplement (1992), 2458–2458.

[88] Lee B. Hinkle, Kamrad Khoshhal Roudposhti, and Vangelis Metsis. 2019. Physiological Measurement for Emotion Recognition in Virtual Reality. In *2019 2nd International Conference on Data Intelligence and Security (ICDIS)*. 136–143. https://doi.org/10.1109/ICDIS.2019.00028

[89] Simon M Hofmann, Felix Klotzsche, Alberto Mariola, Vadim Nikulin, Arno Villringer, and Michael Gaebler. 2021. Decoding subjective emotional arousal from EEG during an immersive virtual reality experience. *Elife* 10 (2021), e64812.

[90] Talke Klara Hoppmann. 2009. Examining the 'point of frustration'. The think-aloud method applied to online search tasks. *Quality & Quantity* 43 (2009), 211–224.

[91] HTC. 2018. HTC Vive Pro Support. https://www.vive.com/au/support/vive-pro/

[92] Katherine Isbister. 2016. *How games move us: Emotion by design*. Mit Press.

[93] Syem Ishaque, Alice Rueda, Binh Nguyen, Naimul Khan, and Sridhar Krishnan. 2020. Physiological signal analysis and classification of stress from virtual reality video game. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 867–870.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] Tahira Jibeen, Mirza Muhammad Zubair Baig, and Mudassar Mahmood Ahmad. 2019. Fear of negative evaluation and communication apprehension: The moderating role of communicative competence and extraversion personality trait in Pakistani academia. *Journal of Rational-Emotive & Cognitive-Behavior Therapy* 37 (2019), 185–201.

[99] Crescent Jicol, Christopher Clarke, Emilia Tor, Rebecca M Dakin, Tom Charlie Lancaster, Sze Tung Chang, Karin Petrini, Eamonn O'Neill, Michael J Proulx, and Christof Lutteroth. 2023. Realism and Field of View Affect Presence in VR but Not the Way You Think. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[100] Crescent Jicol, Julia Feltham, Jinha Yoon, Michael J Proulx, Eamonn O'Neill, and Christof Lutteroth. 2022. Designing and Assessing a Virtual Reality Simulation to Build Resilience to Street Harassment. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[101] Crescent Jicol, Chun Hin Wan, Benjamin Doling, Caitlin H Illingworth, Jinha Yoon, Charlotte Headey, Christof Lutteroth, Michael J Proulx, Karin Petrini, and Eamonn O'Neill. 2021. Effects of emotion and agency on presence in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[102] Marta Clavero Jimenez and Thomas P. Buijtenweg. 2013. ATUM: applying multi-layer game design and environmental storytelling. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 2623–2626. https://doi.org/10.1145/2468356.2479479

[103] 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.

[104] Matthew Kay, Lisa A. Elkin, James J. Higgins, and Jacob O. Wobbrock. 2021. *mjskay/ARTool: ARTool 0.11.0*. https://doi.org/10.5281/zenodo.4721941

[105] Aelee Kim, Minha Chang, Yeseul Choi, Sohyeon Jeon, and Kyoungmin Lee. 2018. The effect of immersion on emotional responses to film viewing in a virtual environment. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 601–602.

[106] 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

[107] 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.

[108] J. Matias Kivikangas, Jari Kätsyri, Simo Järvelä, and Niklas Ravaja. 2014. Gender Differences in Emotional Responses to Cooperative and Competitive Game Play. *PLOS ONE* 9, 7 (07 2014), 1–16. https://doi.org/10.1371/journal.pone.0100318

[109] Robert E Kleiger, Phyllis K Stein, and J Thomas Bigger Jr. 2005. Heart rate variability: measurement and clinical utility. *Annals of Noninvasive Electrocardiology* 10, 1 (2005), 88–101.

[110] Silvia Erika Kober, Jürgen Kurzmann, and Christa Neuper. 2012. Cortical correlate of spatial presence in 2D and 3D interactive virtual reality: an EEG study. *International Journal of Psychophysiology* 83, 3 (2012), 365–374.

[111] Silvia Erika Kober and Christa Neuper. 2012. Using auditory event-related EEG potentials to assess presence in virtual reality. *International Journal of Human-Computer Studies* 70, 9 (2012), 577–587.

[112] Makrina Viola Kosti, Nefeli Georgakopoulou, Sotiris Diplaris, Theodora Pistola, Konstantinos Chatzistavros, Vasileios-Rafail Xefteris, Athina Tsanousa, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2023. Assessing Virtual Reality Spaces for Elders Using Image-Based Sentiment Analysis and Stress Level Detection. *Sensors* 23, 8 (2023), 4130.

[113] 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

[114] Saskia F Kuliga, Tyler Thrash, Ruth Conroy Dalton, and Christoph Hölscher. 2015. Virtual reality as an empirical research tool—Exploring user experience in a real building and a corresponding virtual model. *Computers, environment and urban systems* 54 (2015), 363–375.

[115] Imam Kusmaryono, Dyana Wijayanti, and Hevy Risqi Maharani. 2022. Number of Response Options, Reliability, Validity, and Potential Bias in the Use of the Likert Scale Education and Social Science Research: A Literature Review. *International Journal of Educational Methodology* 8, 4 (2022), 625–637.

[116] J Laarni, N Ravaja, and T Saari. 2003. Using eye tracking and psychophysiological methods to study spatial presence. In *Annual International Workshop on Presence, USA, 2003*.

[117] Jari Laarni, Niklas Ravaja, Timo Saari, Saskia Böcking, Tilo Hartmann, and Holger Schramm. 2015. *Ways to Measure Spatial Presence: Review and Future Directions*. Springer International Publishing, Cham, 139–185. https://doi.org/10.1007/978-3-319-10190-3_8

[118] Lain. 2024. Open Broadcaster Software. https://obsproject.com/.

[119] Reed Larson and Mihaly Csikszentmihalyi. 2014. *The Experience Sampling Method*. Springer Netherlands, Dordrecht, 21–34. https://doi.org/10.1007/978-94-017-9088-8_2

[120] Gaël Laurans, Pieter MA Desmet, and Paul Hekkert. 2009. The emotion slider: A self-report device for the continuous measurement of emotion. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. Ieee, 1–6.

[121] Richard S Lazarus. 1991. Progress on a cognitive-motivational-relational theory of emotion. *American psychologist* 46, 8 (1991), 819.

[122] Nicole Lazzaro. 2009. Why we play: affect and the fun of games. *Human-computer interaction: Designing for diverse users and domains* 155 (2009), 679–700.

[123] Jeroen S Lemmens, Monika Simon, and Sindy R Sumter. 2022. Fear and loathing in VR: the emotional and physiological effects of immersive games. *Virtual Reality* 26, 1 (2022), 223–234.

[124] 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.

[125] Benny Liebold, Michael Brill, Daniel Pietschmann, Frank Schwab, and Peter Ohler. 2017. Continuous measurement of breaks in presence: psychophysiology

and orienting responses. *Media Psychology* 20, 3 (2017), 477–501.
[126] Bing Liu. 2022. *Sentiment analysis and opinion mining.* Springer Nature.
[127] Qian Liu, Yanyun Wang, Qingyang Tang, and Ziwei Liu. 2020. Do You Feel the Same as I Do? Differences in Virtual Reality Technology Experience and Acceptance Between Elderly Adults and College Students. *Frontiers in Psychology* 11 (2020). https://doi.org/10.3389/fpsyg.2020.573673
[128] Matthew Lombard and Theresa Ditton. 1997. At the heart of it all: The concept of presence. *Journal of computer-mediated communication* 3, 2 (1997), JCMC321.
[129] Mario Lorenz, Jennifer Brade, Philipp Klimant, Christoph-E. Heyde, and Niels Hammer. 2023. Age and gender effects on presence, user experience and usability in virtual environments–first insights. *PLOS ONE* 18, 3 (03 2023), 1–16. https://doi.org/10.1371/journal.pone.0283565
[130] Daniel Loureiro. 2022. twitter-roberta-base-sentiment-latest. https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest?text=ugh%2Bi%2Bdefinitenly%2Bhate%2Bthe%2Bsecond%2Btrack.%2Bnot%2Bwide%2Benough%2Band%2Bfeels%2Blike%2Bsooo%2Bmany%2Bstuff%2Baround%2Bme [Accessed 09-09-2024].
[131] Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic Language Models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Association for Computational Linguistics, Dublin, Ireland, 251–260. https://doi.org/10.18653/v1/2022.acl-demo.25
[132] Tímea Magyaródi, Henriett Nagy, Péter Soltész, Tamás Mózes, and Attila Oláh. 2013. Psychometric properties of a newly established flow state questionnaire. *The Journal of Happiness & Well-Being* 1, 2 (2013), 85–96.
[133] Guido Makransky, Lau Lilleholt, and Anders Aaby. 2017. Development and validation of the Multimodal Presence Scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior* 72 (2017), 276–285.
[134] Valentina Mancuso, Francesca Bruni, Chiara Stramba-Badiale, Giuseppe Riva, Pietro Cipresso, and Elisa Pedroli. 2023. How do emotions elicited in virtual reality affect our memory? A systematic review. *Computers in Human Behavior* 146 (2023), 107812. https://doi.org/10.1016/j.chb.2023.107812
[135] Wetherell Margaret. 2012. *Affect and Emotion: A New Social Science Understanding.* SAGE Publications Ltd. https://doi.org/10.4135/9781446250945
[136] Marieke AG Martens, Angus Antley, Daniel Freeman, Mel Slater, Paul J Harrison, and Elizabeth M Tunbridge. 2019. It feels real: physiological responses to a stressful virtual reality environment and its impact on working memory. *Journal of Psychopharmacology* 33, 10 (2019), 1264–1273.
[137] Javier Marín-Morales, Carmen Llinares, Jaime Guixeres, and Mariano Alcañiz. 2020. Emotion Recognition in Immersive Virtual Reality: From Statistics to Affective Computing. *Sensors* 20, 18 (2020). https://doi.org/10.3390/s20185163
[138] Iris B Mauss, Robert W Levenson, Loren McCarter, Frank H Wilhelm, and James J Gross. 2005. The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion* 5, 2 (2005), 175.
[139] Iris B. Mauss and Michael D. Robinson. 2009. Measures of emotion: A review. *Cognition and Emotion* 23, 2 (2009), 209–237. https://doi.org/10.1080/02699930802204677 arXiv:https://doi.org/10.1080/02699930802204677 PMID: 19809584.
[140] Christopher N Maymon, Matthew T Crawford, Katie Blackburne, André Botes, Kieran Carnegie, Samuel A Mehr, Jeremy Meier, Justin Murphy, Nicola L Miles, Kealagh Robinson, et al. 2024. The presence of fear: How subjective fear, not physiological changes, shapes the experience of presence. *Journal of experimental psychology: general* (2024).
[141] 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.
[142] Cade McCall, Lea K. Hildebrandt, Boris Bornemann, and Tania Singer. 2015. Physiophenomenology in retrospect: Memory reliably reflects physiological arousal during a prior threatening experience. *Consciousness and Cognition* 38 (2015), 60–70. https://doi.org/10.1016/j.concog.2015.09.011
[143] James L McGaugh. 2013. *Emotions and bodily responses: A psychophysiological approach.* Academic Press.
[144] Michael Meehan, Brent Insko, Mary Whitton, and Frederick P Brooks Jr. 2002. Physiological measures of presence in stressful virtual environments. *Acm transactions on graphics (tog)* 21, 3 (2002), 645–652.
[145] Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology* 14 (1996), 261–292.
[146] Miguel Melo, Hugo Coelho, Guilherme Gonçalves, Nieves Losada, Filipa Jorge, Mário Sérgio Teixeira, and Maximino Bessa. 2022. Immersive multisensory virtual reality technologies for virtual tourism: A study of the user's sense of presence, satisfaction, emotions, and attitudes. *Multimedia Systems* 28, 3 (2022), 1027–1037.
[147] Meta. 2024. Quest Health and Safety Warnings. https://www.meta.com/gb/legal/quest/health-and-safety-warnings/?utm_source=statics.teams.cdn.office.net&utm_medium=oculusredirect
[148] Ben Meuleman and David Rudrauf. 2021. Induction and Profiling of Strong Multi-Componential Emotions in Virtual Reality. *IEEE Transactions on Affective Computing* 12, 1 (2021), 189–202. https://doi.org/10.1109/TAFFC.2018.2864730
[149] Aire Mill, Anu Realo, and Jüri Allik. 2016. Retrospective Ratings of Emotions: the Effects of Age, Daily Tiredness, and Personality. *Frontiers in Psychology* 6 (2016). https://doi.org/10.3389/fpsyg.2015.02020
[150] Jessica Isbely Montana, Marta Matamala-Gomez, Marta Maisto, Petar Aleksandrov Mavrodiev, Cesare Massimo Cavalera, Barbara Diana, Fabrizia Mantovani, and Olivia Realdon. 2020. The benefits of emotion regulation interventions in virtual reality for the improvement of wellbeing in adults and older adults: a systematic review. *Journal of clinical medicine* 9, 2 (2020), 500.
[151] Agnes Moors, Phoebe C. Ellsworth, Klaus R. Scherer, and Nico H. Frijda. 2013. Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review* 5, 2 (2013), 119–124. https://doi.org/10.1177/1754073912468165 arXiv:https://doi.org/10.1177/1754073912468165
[152] Sarah Morélot, Alain Garrigou, Julie Dedieu, and Bernard N'Kaoua. 2021. Virtual reality for fire safety training: Influence of immersion and sense of presence on conceptual and procedural acquisition. *Computers & Education* 166 (2021), 104145.
[153] Kevin W Mossholder, Randall P Settoon, Stanley G Harris, and Achilles A Armenakis. 1995. Measuring emotion in open-ended survey responses: An application of textual data analysis. *Journal of management* 21, 2 (1995), 335–355.
[154] 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.
[155] 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.
[156] Martin T Orne. 2009. Demand characteristics and the concept of quasi-controls. *Artifacts in behavioral research: Robert Rosenthal and Ralph L. Rosnow's classic books* 110 (2009), 110–137.
[157] Charles Egerton Osgood, William H May, and Murray S Miron. 1975. *Cross-cultural universals of affective meaning.* Vol. 1. University of Illinois Press.
[158] David Oswald, Fred Sherratt, and Simon Smith. 2014. Handling the Hawthorne effect: The challenges surrounding a participant observer. *Review of social studies* 1, 1 (2014), 53–73.
[159] Shaun O'Leary, Marte Lund, Tore Johan Ytre-Hauge, Sigrid Reiersen Holm, Kaja Naess, Lars Nagelstad Dalland, and Steven M McPhail. 2014. Pitfalls in the use of kappa when interpreting agreement between multiple raters in reliability studies. *Physiotherapy* 100, 1 (2014), 27–35.
[160] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval* 2, 1–2 (2008), 1–135.
[161] Thomas D Parsons, Andrea Gaggioli, and Giuseppe Riva. 2017. Virtual reality for research in social neuroscience. *Brain sciences* 7, 4 (2017), 42.
[162] 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.
[163] Xiaolan Peng, Jin Huang, Alena Denisova, Hui Chen, Feng Tian, and Hongan Wang. 2020. A palette of deepened emotions: exploring emotional challenge in virtual reality games. In *Proceedings of the 2020 CHI conference on human factors in computing systems.* 1–13.
[164] Henrik M. Peperkorn, Julia Diemer, and Andreas Mühlberger. 2015. Temporal dynamics in the relation between presence and fear in virtual reality. *Computers in Human Behavior* 48 (2015), 542–547. https://doi.org/10.1016/j.chb.2015.02.028
[165] Elizabeth R Peterson, Gavin TL Brown, and Miriam C Jun. 2015. Achievement emotions in higher education: A diary study exploring emotions across an assessment event. *Contemporary Educational Psychology* 42 (2015), 82–96.
[166] Igor V Petukhov, Andrey E Glazyrin, Andrey V Gorokhov, Luydmila A Steshina, and Ilya O Tanryverdiev. 2020. Being present in a real or virtual world: a EEG study. *International journal of medical informatics* 136 (2020), 103977.
[167] James L Peugh. 2010. A Practical Guide to Multilevel Modeling. *Journal of School Psychology* 48, 1 (2010), 85–112.
[168] Karin A Pfeiffer, James M Pivarnik, Christopher J Womack, Mathew J Reeves, and Robert M Malina. 2002. Reliability and validity of the Borg and OMNI rating of perceived exertion scales in adolescent girls. *Medicine & Science in Sports & Exercise* 34, 12 (2002), 2057–2061.
[169] Rosalind W Picard. 2000. *Affective computing.* MIT press.
[170] 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* (Honolulu, HI, USA) *(CHI '24).* Association for Computing Machinery, New York, NY, USA, Article 757, 21 pages. https://doi.org/10.1145/3613904.3642611
[171] 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
[172] Dominic Potts, Miloni Gada, Aastha Gupta, Kavya Goel, Klaus Phillip Krzok, Genieve Pate, Joseph Hartley, Mark Weston-Arnold, Jakob Aylott, Christopher Clarke, Crescent Jicol, and Christof Lutteroth. 2025. Dataset for "RetroSketch: A Retrospective Method for Measuring Emotions and Presence in Virtual Reality". Bath: University of Bath Research Data Archive. https://doi.org/10.15125/BATH-01489
[173] Dominic Potts, Joe 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://researchdata.bath.ac.uk/1372/
[174] Susanne Putze, Dmitry Alexandrovsky, Felix Putze, Sebastian Höffner, Jan David Smeddinck, and Rainer Malaka. 2020. Breaking the experience: Effects of questionnaires in vr user studies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[175] R. 2024. polr function - RDocumentation — rdocumentation.org. https://www.rdocumentation.org/packages/MASS/versions/7.3-61/topics/polr.
[176] R. 2024. R Documentation - encomptest. https://www.rdocumentation.org/packages/lmtest/versions/0.9-40/topics/encomptest.
[177] R. 2024. R Documentation - nortest. https://www.rdocumentation.org/packages/nortest/versions/1.0-4.
[178] Next Level Racing. 2024. HF8 Haptic Gaming Pad. url=https://nextlevelracing.com/products/hf8-haptic-gaming-pad/.
[179] Next Level Racing. 2024. Motion Platform V3. url=https://nextlevelracing.com/products/next-level-racing-motion-platform-v3/.
[180] Martin Ragot, Nicolas Martin, Sonia Em, Nico Pallamin, and Jean-Marc Diverrez. 2018. Emotion recognition using physiological signals: laboratory vs. wearable sensors. In *Advances in Human Factors in Wearable Technologies and Game Design: Proceedings of the AHFE 2017 International Conference on Advances in Human Factors and Wearable Technologies, July 17-21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*. Springer, 15–22.
[181] 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.
[182] Niklas Ravaja, Jari Laarni, T Saari, K Kallinen, Mikko Salminen, Jussi Holopainen, and Aki Järvinen. 2004. Spatial presence and emotional responses to success in a video game: A psychophysiological study. *Proceedings of the PRESENCE* (2004), 112–116.
[183] Niklas Ravaja, Timo Saari, Marko Turpeinen, Jari Laarni, Mikko Salminen, and Matias Kivikangas. 2006. Spatial presence and emotions during video game playing: Does it matter with whom you play? *Presence: Teleoperators and virtual environments* 15, 4 (2006), 381–392.
[184] Rainer Reisenzein, Martin Junge, Markus Studtmann, and Oswald Huber. 2014. Observational approaches to the measurement of emotions. In *International handbook of emotions in education*. Routledge, 580–606.
[185] Beatriz Rey, Vera Parkhutik, and Mariano Alcañiz. 2011. Breaks in Presence in Virtual Environments: An Analysis of Blood Flow Velocity Responses. *Presence* 20, 3 (2011), 273–286. https://doi.org/10.1162/PRES_a_00049
[186] 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.
[187] Giuseppe Riva, Fabrizia Mantovani, Claret Samantha Capideville, Alessandra Preziosa, Francesca Morganti, Daniela Villani, Andrea Gaggioli, Cristina Botella, and Mariano Alcañiz. 2007. Affective interactions using virtual reality: the link between presence and emotions. *CyberPsychology & Behavior* 10, 1 (2007), 45–56.
[188] Michael D Robinson and Gerald L Clore. 2002. Belief and feeling: evidence for an accessibility model of emotional self-report. *Psychological bulletin* 128, 6 (2002), 934.
[189] Vertical Robot. 2018. Red Matter on Steam. https://store.steampowered.com/app/966680/Red_Matter/
[190] 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
[191] Ira J Roseman and Craig A Smith. 2001. Appraisal theory. *Appraisal processes in emotion: Theory, methods, research* (2001), 3–19.
[192] Anna Marie Ruef and Robert W Levenson. 2007. Continuous measurement of emotion. *Handbook of emotion elicitation and assessment* (2007), 286–297.
[193] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.
[194] James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 1 (2003), 145.
[195] M Rutter and A Cox. 1981. Psychiatric interviewing techniques: I. Methods and measures. *The British Journal of Psychiatry* 138, 4 (1981), 273–282.
[196] Christina Röcke, Christiane A. Hoppmann, and Petra L. Klumb. 2011. Correspondence Between Retrospective and Momentary Ratings of Positive and Negative Affect in Old Age: Findings From a One-Year Measurement Burst Design. *The Journals of Gerontology: Series B* 66B, 4 (03 2011), 411–415. https://doi.org/10.1093/geronb/gbr024 arXiv:https://academic.oup.com/psychsocgerontology/article-pdf/66B/4/411/16746257/gbr024.pdf
[197] Saeed Shafiee Sabet, Carsten Griwodz, and Sebastian Möller. 2019. Influence of primacy, recency and peak effects on the game experience questionnaire. In *Proceedings of the 11th ACM Workshop on Immersive Mixed and Virtual Environment Systems*. 22–27.
[198] Sinué Salgado and Osman Skjold Kingo. 2019. How is physiological arousal related to self-reported measures of emotional intensity and valence of events and their autobiographical memories? *Consciousness and Cognition* 75 (2019), 102811. https://doi.org/10.1016/j.concog.2019.102811
[199] 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.
[200] Maria V Sanchez-Vives and Mel Slater. 2005. From presence to consciousness through virtual reality. *Nature Reviews Neuroscience* 6, 4 (2005), 332–339.
[201] Wataru Sato. 2024. Advancements in sensors and analyses for emotion sensing. , 4166 pages.
[202] 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.
[203] Klaus R Scherer. 2001. Appraisal Considered as a Process of Multilevel Sequential Checking. *Appraisal Processes in Emotion* (05 2001), 92–120. https://doi.org/10.1093/oso/9780195130072.003.0005
[204] Alexander Schnack, Malcolm J Wright, and Jonathan Elms. 2021. Investigating the impact of shopper personality on behaviour in immersive Virtual Reality store environments. *Journal of Retailing and Consumer Services* 61 (2021), 102581.
[205] 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.
[206] Emery Schubert. 1999. Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology* 51, 3 (1999), 154–165. https://doi.org/10.1080/00049539908255353 arXiv:https://doi.org/10.1080/00049539908255353
[207] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. 2001. The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments* 10, 3 (2001), 266–281.
[208] Nicola S Schutte and Emma J Stilinović. 2017. Facilitating empathy through virtual reality. *Motivation and emotion* 41 (2017), 708–712.
[209] Valentin Schwind, Pascal Knierim, Nico Haas, and Niels Henze. 2019. Using Presence Questionnaires in Virtual Reality. 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, 1–12. https://doi.org/10.1145/3290605.3300590
[210] Christie N Scollon, Chu Kim-Prieto, and Ed Diener. 2003. Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness studies* 4, 1 (2003), 5–34.
[211] Fred Shaffer and Jay P Ginsberg. 2017. An overview of heart rate variability metrics and norms. *Frontiers in public health* (2017), 258.
[212] Thomas B Sheridan et al. 1992. Musings on telepresence and virtual presence. *Presence Teleoperators Virtual Environ.* 1, 1 (1992), 120–125.
[213] 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.
[214] Shimmer. 2024. Shimmer Wearable Sensor Technology. url=https://shimmersensing.com/product/shimmer3-gsr-unit/.
[215] Dong-Hee Shin. 2017. The role of affordance in the experience of virtual reality learning: Technological and affective affordances in virtual reality. *Telematics and Informatics* 34, 8 (2017), 1826–1836. https://doi.org/10.1016/j.tele.2017.05.013
[216] Ilia Shumailov and Hatice Gunes. 2017. Computational analysis of valence and arousal in virtual reality gaming using lower arm electromyograms. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 164–169.
[217] Chris G. Sibley and Niki Harré. 2009. A gender role socialization model of explicit and implicit biases in driving self-enhancement. *Transportation Research Part F: Traffic Psychology and Behaviour* 12, 6 (2009), 452–461. https://doi.org/10.1016/j.trf.2009.08.006

[218] Kunos Simulazioni. 2019. Assetto Corsa Competizione on Steam. https://store.steampowered.com/app/244210/Assetto_Corsa/
[219] 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.
[220] Richard Skarbez, Frederick P. Brooks, and Mary C. Whitton. 2021. Immersion and Coherence: Research Agenda and Early Results. *IEEE Transactions on Visualization and Computer Graphics* 27, 10 (2021), 3839–3850. https://doi.org/10.1109/TVCG.2020.2983701
[221] Mel Slater. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3549–3557.
[222] Mel Slater. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3549–3557. https://doi.org/10.1098/rstb.2009.0138 arXiv:https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2009.0138
[223] Mel Slater. 2018. Immersion and the illusion of presence in virtual reality. *British Journal of Psychology* 109, 3 (2018), 431–433.
[224] Mel Slater et al. 1999. Measuring presence: A response to the Witmer and Singer presence questionnaire. *Presence: teleoperators and virtual environments* 8, 5 (1999), 560–565.
[225] Mel Slater, Domna Banakou, Alejandro Beacco, Jaime Gallego, Francisco Macia-Varela, and Ramon Oliva. 2022. A Separate Reality: An Update on Place Illusion and Plausibility in Virtual Reality. *Frontiers in Virtual Reality* 3 (2022). https://doi.org/10.3389/frvir.2022.914392
[226] Mel Slater, Andrea Brogni, and Anthony Steed. 2003. Physiological responses to breaks in presence: A pilot study. In *Presence 2003: The 6th annual international workshop on presence*, Vol. 157. Citeseer.
[227] Mel Slater, Carlos Cabriera, Gizem Senel, Domna Banakou, Alejandro Beacco, Ramon Oliva, and Jaime Gallego. 2023. The sentiment of a virtual rock concert. *Virtual Reality* 27, 2 (2023), 651–675.
[228] Mel Slater, Christoph Guger, Guenter Edlinger, Robert Leeb, Gert Pfurtscheller, Angus Antley, Maia Garau, Andrea Brogni, and Doron Friedman. 2006. Analysis of physiological responses to a social situation in an immersive virtual environment. *Presence* 15, 5 (2006), 553–569.
[229] Mel Slater and Anthony Steed. 2000. A virtual presence counter. *Presence* 9, 5 (2000), 413–434.
[230] Mel Slater and Sylvia Wilbur. 1997. A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments* 6, 6 (1997), 603–616.
[231] Craig A Smith and Leslie D Kirby. 2000. Consequences require antecedents: Toward a process model of emotion elicitation. *Feeling and thinking: The role of affect in social cognition* (2000), 83–106.
[232] 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.
[233] 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.
[234] 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
[235] Sony. 2024. PS VR2: Safety for Players. https://www.playstation.com/en-gb/support/hardware/ps-vr2-safety/
[236] Vinicius Souza, Anderson Maciel, Luciana Nedel, and Regis Kopper. 2021. Measuring Presence in Virtual Environments: A Survey. *ACM Comput. Surv.* 54, 8, Article 163 (oct 2021), 37 pages. https://doi.org/10.1145/3466817
[237] Bernhard Spanlang, Torsten Fröhlich, Vanessa F Descalzo, Angus Antley, and Mel Slater. 2007. The making of a presence experiment: Responses to virtual fire. In *Annual International Workshop on Presence*. 303–307.
[238] Anthony Steed, Andrea Brogni, and Vinoba Vinayagamoorthy. 2005. Breaks in presence as usability criteria. In *Proceedings of HCI international*.
[239] Sophia C Steinhaeusser, Sebastian Oberdörfer, Sebastian von Mammen, Marc Erich Latoschik, and Birgit Lugrin. 2022. Joyful adventures and frightening places–designing emotion-inducing virtual environments. *Frontiers in Virtual Reality* 3 (2022), 919163.
[240] Jonathan Steuer, Frank Biocca, Mark R Levy, et al. 1995. Defining virtual reality: Dimensions determining telepresence. *Communication in the age of virtual reality* 33 (1995), 37–39.
[241] Jonathan A Stevens and J Peter Kincaid. 2015. The relationship between presence and performance in virtual simulation training. *Open Journal of Modelling and Simulation* 3, 2 (2015), 41–48.
[242] Bethesda Game Studios. 2017. https://store.steampowered.com/app/611670/The_Elder_Scrolls_V_Skyrim_VR/
[243] 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).
[244] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jittrapol Intarasisrisawat, Maxine Glancy, and Chee Siang Ang. 2022. VREED: Virtual Reality Emotion Recognition Dataset Using Eye Tracking & Physiological Measures. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 178 (dec 2022), 20 pages. https://doi.org/10.1145/3495002
[245] 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.
[246] 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
[247] 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.
[248] Gustavo Tondello, Karina Arrambide, Giovanni Ribeiro, Andrew Cen, and Lennart Nacke. 2019. "I don't fit into a single type": A Trait Model and Scale of Game Playing Preferences.
[249] Anestis Touloumis. 2014. R package multgee: a generalized estimating equations solver for multinomial responses. *arXiv preprint arXiv:1410.5232* (2014).
[250] Anestis Touloumis, Alan Agresti, and Maria Kateri. 2013. GEE for multinomial responses using a local odds ratios parameterization. *Biometrics* 69, 3 (2013), 633–640.
[251] Tanh Quang Tran, Tobias Langlotz, and Holger Regenbrecht. 2024. A Survey On Measuring Presence in Mixed Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 543, 38 pages. https://doi.org/10.1145/3613904.3642383
[252] Polar UK. 2024. Polar H10 Heart rate sensor. url=https://www.polar.com/uk-en/sensors/h10-heart-rate-sensor.
[253] Heather L. Urry and James J. Gross. 2010. Emotion Regulation in Older Age. *Current Directions in Psychological Science* 19, 6 (2010), 352–357. https://doi.org/10.1177/0963721410388395 arXiv:https://doi.org/10.1177/0963721410388395
[254] Valve. 2020. Half-Life: Alyx on Steam. https://store.steampowered.com/app/546560/HalfLife_Alyx/
[255] Thea F Van de Mortel. 2008. Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, The* 25, 4 (2008), 40–48.
[256] William N Venables and Brian D Ripley. 2013. *Modern applied statistics with S-PLUS*. Springer Science & Business Media.
[257] Mikko Vesisenaho, Merja Juntunen, Päivi Häkkinen, Johanna Pöysä-Tarhonen, Janne Fagerlund, Iryna Miakush, and Tiina Parviainen. 2019. Virtual reality in education: Focus on the role of emotions and physiological reactivity. *Journal of Virtual Worlds Research* 12, 1 (2019).
[258] HTC Vive. 2023. *Eye and Facial Tracking SDK - Developer Resources*. https://developer.vive.com/resources/vive-sense/eye-and-facial-tracking-sdk/
[259] 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.
[260] 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.
[261] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.
[262] Eric D. Wesselmann, James H. Wirth, Daniel K. Mroczek, and Kipling D. Williams. 2012. Dial a feeling: Detecting moderation of affect decline during ostracism. *Personality and Individual Differences* 53, 5 (2012), 580–586. https://doi.org/10.1016/j.paid.2012.04.039
[263] J. Christopher Westland. 2022. Information loss and bias in likert survey responses. *PLOS ONE* 17, 7 (07 2022), 1–17. https://doi.org/10.1371/journal.pone.0271949
[264] Ladd Wheeler and Harry T Reis. 1991. Self-recording of everyday life events: Origins, types, and uses. *Journal of personality* 59, 3 (1991), 339–354.
[265] Graham Wilson and Mark McGill. 2018. Violent video games in virtual reality: Re-evaluating the impact and rating of interactive experiences. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. 535–548.
[266] Bob G Witmer and Michael J Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence* 7, 3 (1998), 225–240.
[267] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. https:

//doi.org/10.1145/1978942.1978963

[268] 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.

[269] Tong Xue, Abdallah El Ali, Tianyi Zhang, Gangyi Ding, and Pablo Cesar. 2023. CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360° VR Videos. *IEEE Transactions on Multimedia* 25 (2023), 243–255. https://doi.org/10.1109/TMM.2021.3124080

[270] Tong Xue, Abdallah El Ali, Tianyi Zhang, Gangyi Ding, and Pablo Cesar. 2021. RCEA-360VR: Real-time, Continuous Emotion Annotation in 360° VR Videos for Collecting Precise Viewport-dependent Ground Truth Labels. 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 513, 15 pages. https://doi.org/10.1145/3411764.3445487

[271] 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.

[272] Pavel Zahorik and Rick L Jenison. 1998. Presence as being-in-the-world. *Presence* 7, 1 (1998), 78–89.

[273] 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.