frontiers | Frontiers in Virtual Reality

TYPE Original Research
PUBLISHED 20 June 2025
DOI 10.3389/frvir.2025.1458191

OPEN ACCESS

EDITED BY
Ir. Ts. Dr. Norizam Sulaiman,
Universiti Malaysia Pahang, Malaysia

REVIEWED BY
Peter Eachus,
University of Salford, United Kingdom
Hai-Ning Liang,
The Hong Kong University of Science and Technology (Guangzhou), China

*CORRESPONDENCE
Daniel Archer,
daniel.archer.18@ucl.ac.uk

RECEIVED 02 July 2024
ACCEPTED 20 May 2025
PUBLISHED 20 June 2025

CITATION
Archer D, Li C, Chen G, Dai Y and Steed A (2025) Assessing the effect of arousal on performance in a virtual reality narrative scenario using biological signals.
Front. Virtual Real. 6:1458191.
doi: 10.3389/frvir.2025.1458191

Assessing the effect of arousal on performance in a virtual reality narrative scenario using biological signals

Daniel Archer[1]*, Chen Li[2], Guang Chen[2], Yixin Dai[2] and Anthony Steed[1]

[1]Computer Science Department, University College London, London, United Kingdom, [2]Department of Computer Science, Polytechnic University of Hong Kong, Kowloon, Hong Kong SAR, China

When analysing a user's experience of virtual reality (VR), factors such as their level of technical familiarity, proficiency with immersive technology and concomitant degree of physiological arousal inside a VR experience can have a significant impact on their performance, sense of presence and engagement. We have designed a modular narrative system to manipulate a user's levels of arousal in order to keep them within an optimal range for performance, which we hypothesise to be between not too stressed (high arousal) and not too bored (low arousal). We do so by instantiating an increasing number of simultaneous tests and environmental changes at different points during a VR experience. Changes in autonomic signals - such as heart rate, heart rate variability, galvanic skin response, and skin temperature - reveal changes in the levels of participant arousal. The user is embodied in a gender-specific out-group (Muslim) avatar that is subjected to an increasingly stressful event (a series of verbal Islamophobic attacks from a non-player character). We measure performance in a series of simultaneous multiple choice listening comprehension tasks (averaged to create a "narrative task score") undertaken as the scene unfolds, and a post-treatment recall task. As a pilot experiment, our primary objective is to validate the effectiveness of the system as a means of stress manipulation and thereby assess the impact and correlation that different levels of arousal have on task performance and biological signals. Results revealed a statistically significant difference in narrative task performance between stress levels, confirmed by a one-way ANOVA ($F(2, 45) = 5.06$, $p = 0.02$, $SE = 23.89$). The low stress group achieved the highest mean VR score ($M = 73.12$, $SD = 15.96$), followed by the high ($M = 63.25$, $SD = 18.23$) and medium stress groups ($M = 51.81$, $SD = 23.66$). Our hypothesis that the medium stress condition would produce the best performance was therefore rejected. Comparing heart rate variability (HRV) metrics, the Stress Index showed a statistically significant difference between conditions ($p = 0.043$, with significant within-condition changes also observed in the LF/HF ratio ($p = 0.005$ in low stress and $p = 0.008$ in high stress), further demonstrating the physiological changes between stress levels.

KEYWORDS

biological signals, virtual reality, avatar, performance, immersive, arousal, stress

1 Introduction

The breadth of virtual reality (VR) experiences available to the general public is growing rapidly, driven by the availability of cheap, capable devices. However, given that VR is a relatively new medium, there is a very wide range of levels of user familiarity and competence. Thus, almost all consumer VR experiences need to impart instructions to users, which have to assume very little familiarity with the medium (Steed et al., 2023). Previous studies have examined how the method of instruction determines a user's subsequent ability to operate effectively within a virtual space, but might also prove under or over-stimulating in the process, leading to poor performance or decreased retention of new information (Archer and Steed, 2022; Luong et al., 2020).

A key assumption of our research is that biological signals (including heart rate (HR), heart rate variability (HRV), galvanic skin response (GSR) and skin temperature) provide a suitable quantitative measure of arousal, and that each user has an optimal range for these measures during a virtual reality experience (VRE). This is based on the well-established relationship between pressure and performance first proposed by Yerkes-Dodson in 1908, the so-called "Law of Arousal" (Yerkes and Dodson, 1908). The premise is that there is a curvilinear relationship between motivation and performance, following an inverted U-model, in which performance is optimized by a moderate amount of arousal, but is reduced if that level of arousal becomes too high or too low. For more details on how these models were extended and the basis for our work, see Section 2. Optimal stress and performance models have already been applied to VR. Claude et al. (2015) studied stress-related effects of degrees of mental workload within immersive training scenarios. Parsons and Reinebold (2012) used biological signals to compare responsive virtual environments in neuropsychological evaluations.

In this study, we will build on the research outlined above to investigate whether significant correlations between arousal levels and task performance persist in an embodied, story-driven VR environment. Luong et al. (2020) adapted NASA's Multi-Attribute Task Battery (MATB-II) test (Comstock and Arnegard, 1992) into a VR cockpit to focus on real-time recognition of users' mental workload. We wanted to redesign elements of those performance tests while recontextualising the VRE from a cockpit to an everyday setting in order to assess whether a user's arousal levels affected their sense of affinity for the out-group avatar they embodied. In our case, we embodied the user in a Muslim male or female avatar corresponding to their self-reported gender and had them experience a stressful Islamophobic incident from a first-person perspective. Sensitivity around the topic was paramount given the confrontational nature of the stress event. The experiment was granted approval for data collection by Hong Kong Polytechnic University's Ethics Committee, under approval number HSEARS20221121002. A senior journalist from BBC's Asian Desk oversaw the script the treatment was based on, which also included Muslim interviewees' own experiences of Islamophobia in the United Kingdom. Audio recordings used in the treatment were taken directly from an Islamophobic incident in London that was reported in 2018.

For an overview of the difference between stress and arousal, see Section 2.1. We seek to manipulate arousal by controlling the number of simultaneous mental workload tasks the user is given at seven different story moments within the scene to be completed during a fixed 10-s time period. The study follows a body of work that involves the triggering and detection of stress responses in virtual scenarios in participants' biological signals (Martens et al., 2019; Cleworth et al., 2012; Archer and Finger, 2018).

We begin by assigning the participant's stress level. Specifically, the high stress condition has three tasks to be completed simultaneously, the medium stress condition has two, and the low stress condition has one. This was based on the premise that a high mental workload will result in higher arousal, producing stress and fatigue reactions (Gaillard, 2000).

Optimal results such as a higher performance score, greater post-treatment recall and raised variation in their biological signals from the baseline (such as HRV variables) are hypothesized to occur when a participant's arousal levels are high enough to maintain engagement and avoid boredom (eustress) without being so high that they provoke mental overload or panic (distress). We used eustress and distress as introduced by Selye (1974) to denote positive and negative manifestations of the stress response.

This study aims to validate the manipulation of stress conditions as an effective way to adjust the degree of arousal and an assessment of the concomitant impact on participants' physiological measurements (Hypothesis 1, 7, 8). A further Hypothesis 9 is to investigate whether the stress conditions will moderate the user's biological signals in a positive or negative way: the medium stress level is hypothesized to produce the highest performance metrics (narrative task score) while also not provoking discomfort or a feeling of simulator sickness. Stress levels have been shown to be independent of cybersickness in previous studies, such as the work of Servotte et al. (2020). This should be of broad interest because performance and engagement is an area that is relevant to a large number of VR designers.

2 Related work

We are interested in the relationship between task performance and arousal under stress in a virtual environment. The working assumption is that during sustained periods within an optimal arousal margin, the participant will feel more engaged and present in the virtual environment, improving their performance as a result. Broadhurst (1959) enhanced the original Yerkes and Dodson (1908) experiment by including four additional motivation levels and three difficulty levels. Nixon et al. (1979) work incorporated the Stress Response Curve and Klein (1982) linked improved cued recall to arousal or stress.

2.1 Stress vs. arousal

For the purposes of this paper, arousal shall be the term used to refer to the immediate response to a challenging task, and stress for when that response is insufficient to perform the task, resulting in the participant feeling overloaded.

Stress is a psychophysiological response to environmental stimuli when a situation is perceived as challenging or threatening, activating the sympathetic and parasympathetic

nervous systems, which together constitute the autonomic nervous system. Its purpose is to re-establish equilibrium when homeostasis is threatened through various physiological and behavioural adaptive responses (Chrousos, 2009). Lazarus (2006) distilled Hooke's etymological analysis of stress into three basic concepts using the metaphor of a load-bearing bridge: load, stress and strain. Load referred to the external force applied, stress to the area affected, and strain to the deformation of the structure as a result, analogous to the physiological response.

Arousal, on the other hand, has been described as a psychological trait that is considered a useful or appropriate aspect of a response to perceived demand (King et al., 1983). Cox's Stress/Arousal Adjective Checklist (SACL) (Mackay et al., 1978) was assessed by King et al. in a study involving 126 participants who were given a visual search/detection task involving slides of concealed men. These slides were shown at different speeds and responses were monitored among a group of parachutists, army clerks and psychiatric patients on medication. Twenty words (ten for stress and ten for arousal) were chosen for participants to describe their experience. Arousal was elevated in response to a high load cognitive demand. The demand of an uncomfortably-paced task was also found to mitigate against lowered arousal. Elevated arousal was associated with a coping response, while elevated stress appears to indicate the presence of fear or doubts about coping.

The resultant view is that mild stress tends to facilitate cognitive function, particularly in implicit memory or simple declarative tasks or when the cognitive load is not excessive (Sandi, 2013). An important distinction here is our focus on acute stress, which can be described as a recent, transient occurrence of a single stressor, as opposed to chronic or persistent stress, which refers to an ongoing difficulty facing an individual. A meta-analysis of acute stress effects on executive functions showed that stress contributes to a state of reactive and automatic cognitive processing while enhancing executive motor control, which should facilitate engagement with or escape from the current stressor (Shields et al., 2016).

While a moderate amount of stress is necessary for optimal engagement and focus during performance of a task, stress that exceeds an individual's coping resources can have deleterious effects on performance (Anton et al., 2021). Individual responses to stressful situations vary considerably, which, according to the Biopsychosocial Model (BPSM) of challenge and threat (Blascovich, 2008), can be explained by the evaluations of individuals of their personal coping resources and situational demands (e.g., skills, uncertainty, psychological danger). BPSM postulates that when people are engaged in a task, as evidenced by an increase in HR (Seery et al., 2009), and motivated to perform well, they enter into conscious, unconscious, and dynamic demand and resource evaluation processes. When task demands are deemed to outweigh personal coping resources, a threat state occurs (equivalent to our high stress state in this experiment), whereas when coping resources are judged to match or outweigh demands, a challenge state occurs (equivalent to our high arousal/optimal performance state). These states do not act as two dichotomous entities, but are instead two ends of a bipolar spectrum (Blascovich, 2008).

Mandrick et al. (2016) confirmed the psychophysiological cost of mental effort and stress in a n = 34 study using a new mental arithmetic n-back task (Toulouse N-back task) coupled with aversive audio samples, to assess the impact on task performance, pupil response, cardiovascular activity, and prefrontal cortex oxygenation. Stress probably triggered increased motivation and the recruitment of additional cognitive resources that minimize their aversive effects on task performance (effectiveness), but these compensatory efforts consumed resources that caused a loss of cognitive efficiency (ratio between performance effectiveness and mental effort).

2.2 Performance, arousal and stress in VR

Next, we consider external factors that affect arousal and performance in VR.Wu et al. (2010) embedded a VR version of the Stroop Task (VRST) inside an immersive military simulation, using the reaction time of the participants under varying degrees of stress as a measure of performance. They showed that when performance is rated on reaction time, moderate (as opposed to high or low) stress elicited the optimal level of arousal for the majority (11 of 18) of subjects. Furthermore, their findings suggested that high classification rates were achievable when psychophysiological responses (galvanic skin response, respiration, ECG, and EEG) of these three stimuli presentations were categorized into three levels of arousal.

Palmas et al. (2019) compared the completion time and error count during a VR training task - the assembly of a virtual drum set. The task was either gamified (indicating progress, score, audiovisual feedback) or not. Results showed that gamification produced beneficial effects, particularly for the participants who were VR noviced.

Avatar observation of user performance is another potential factor, as investigated by Hayes et al. (2010), and within the context of social facilitation by Sterna et al. (2019). Blascovich's research on the biopsychosocial model of challenge and threat from external observers is also of note. Novel task performance in the presence of others was impacted by both increased cardiac response and greater vascular resistance from baseline (Blascovich et al., 1999). Our study builds on the above examples, but also aims to fill the gap in the literature in relation to the combination of biological signals, performance and arousal in relation to embodied narrative-based VREs.

2.3 Body ownership and agency

Body ownership and the sense that your body has been replaced by a virtual avatar (Won et al., 2015) have emerged as a key component of user experience in immersive systems. Botvinick and Cohen (1998) revealed the three-way interaction between vision, touch, and proprioception (i.e., awareness of position and movement of the body) through the Rubber Hand Illusion (RHI). Yuan and Steed (2010), Zhang et al. (2015) and Sanchez-Vives and Slater (2005) demonstrated that a form of the RHI illusion occurs for a virtual hand experienced from a first-person point of view in a HMD. Agency is also an essential component for preventing the VR experience from feeling too passive: the actions of the virtual body are attributed to our physical self (Haggard and Chambon, 2012).

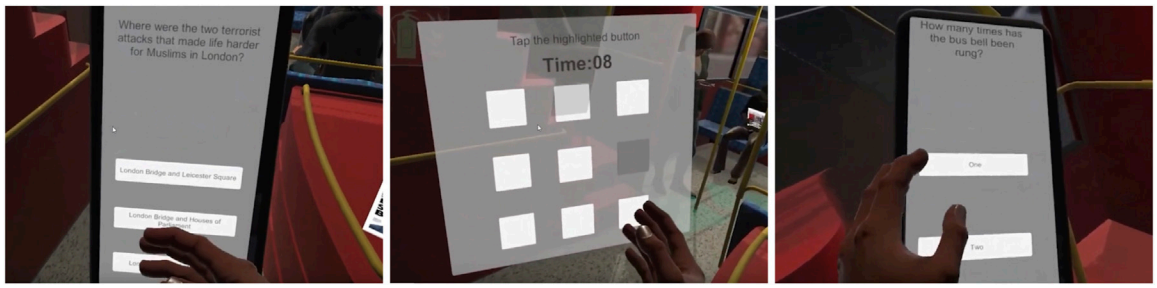

FIGURE 1
Narrative, movement and monitoring tasks comprising the low, medium and high stress conditions.

Extrapolating from these core tenets: how does perceived self-representation as a virtual avatar moderate performance in VREs? Embodiment in the first-person perspective of an avatar whose movements correspond to the user's may lighten a user's cognitive load, improving their memory performance after VR exposure (Steed et al., 2016). In that experiment, participants whose virtual avatar's hands were synchronised with their own physical hands had a significantly improved recall rate of pairs of memorised letters after performing a spatial recognition task than those who did so without a virtual avatar. Pan and Steed (2019) corroborated the same findings when they replicated the experiment using virtual hands alone. Peck and Tutar (2020) also deployed a VRST to compare user performance with or without virtually collocated hands, finding that proximal hands produced a significant increase in accuracy, despite recognizing that Stroop interference was not mediated by self-avatar or level of embodiment.

3 Materials and methods

In a between-subjects design, participants were randomly assigned one of three conditions at run-time: low, medium, or high stress. This condition dictated the number of simultaneous tasks they had to complete in a 10-s interval at seven different story moments during the immersive scenario shown in Figure 1. The three tasks were called the narrative task, the movement task, and the monitoring task. The narrative task was common to all three conditions and featured a multiple choice question related to the narrative voice-over played to the user prior to the quiz. The script was based on interviews with Muslims who had experienced discrimination in the United Kingdom. It was recorded in male and female voices to match the sex of the Muslim avatar assigned to the participant, thereby representing a form of internal monologue for the participant. The movement task was a "whack-a-mole"-style rapid response task for the user to tap one highlighted square within a 3 × 3 grid of nine squares. Failure to tap the correct square or not tap any square before the end of the countdown resulted in "failure" being logged. The monitoring task was a multiple choice question related to the number of times an audio sample of a bus bell had played since the last quiz, or the name of the previous bus stop, which was displayed in two locations in the scene (in a style identical to the layout of a London bus, which is what the scene was modelled on). Thus, the monitoring task was also tied to the embodied narrative of being inside a virtual bus that the player was experiencing. Low stress condition participants only had to complete the narrative task, medium stress participants had to complete the narrative and movement tasks, and *high stress* participants were given all three tasks to complete before the countdown of 10 seconds ended, as shown in Figure 1. A written non-VR version of the narrative task with the same multiple choice questions was repeated post-treatment to assess whether increased arousal led to improved post-treatment recall.

The tasks were inspired by NASA's Multi-Attribute Task Battery test (Comstock and Arnegard, 1992). They were designed to provoke increasing levels of arousal through assigning an increasing number of simultaneous mental workload tasks to the user within the fixed (10 s) time period. This is also based on the theory of cognitive load known as the split-attention effect, in which procedural instructions cause people to divide their attention between different types of information presented (e.g., textual information and graphical information), thus increasing cognitive load (Van Acker et al., 2018).

3.1 Demographics

Forty-eight participants completed the study (31 were female, 17 were male). The average age was 23.5 years. The ages ranged from 18 to 44 years, with a variance of 17.54 years 16 participants were assigned to each condition, and the mean age per condition was 22, 23, and 25 for low stress, medium stress and high stress respectively, with standard deviations of 2.52, 2.99, and 6.00. No inaccuracies were recorded in the sensor readings, although eight participants (3 in low stress, 1 in medium stress and 4 in high stress) ran into technical issues with the interactables in the bus scene, which meant they had to restart the experience. Recruitment was focused on the student population who could prove their level of English reading and listening comprehension was high enough. It was also open to post-graduate students within the Humanities faculty, which might account for the higher number of female participants.

In terms of VR experience, the mean score for prior VR usage was based on two factors: the number of VR experiences played and the type of VR headset ownership. The former was divided as follows on a 1-5 scale: zero experiences 1); less than five 2); less than ten 3); less than thirty 4); over fifty 5). The latter was divided by ascending order by the price of VR headsets, based on the rationale that the more expensive (and the greater the number - since users could also

select multiple options) headsets owned, the more a participant had invested in the medium. This was divided into: none 1); Google Cardboard 2); Oculus Go/Gear VR 3); Quest/Pico 4); Vive/Index/Rift 5). For example, a participant who has played a VR experience 4 times and owns a Google Cardboard headset would score 2 + 2/2 = 2/5 on the experience score. See Supplementary Section 7.1 for the full formula.

The mean participant experience score was 2/5 for each condition. It is accepted that an increased sample size could improve the aforementioned discrepancies in gender distribution and experience across conditions. Sagnier et al. (2019) found that gender had an effect on participant self-assessment and the ability to act during an assembly task in VR, while prior experience had an effect on pragmatic and hedonic quality stimulation as well as performance.

3.2 Protocol

After arriving at the lab, participants were briefed and given the information sheet. This gave sufficient time for their biological signals to establish a normal resting level. They then began the experience by putting an E4 Empatica wireless Bluetooth wearable on the wrist of their non-dominant hand followed by the Meta Quest 2 HMD. Participants were screened for being non-Muslim, in order to avoid triggering any negative experiences the participant might have experienced and provoking memories of historical stress as a result.

The baseline biological signals of the participant were then recorded for 1 min inside a VR experience of a minimally-decorated virtual room. They were then directed to complete an interactive survey inside VR relating to their VR consumption habits and whether they own a headset (*survey scene*), described in the demographics section above. Combining these measurements together constituted an ultra-short-term (UST, that is <5 minute) experimental norm for HR measurement (Shaffer and Ginsberg, 2017). This was to gauge their familiarity with navigating similar scenarios and to obtain a pre-treatment baseline. Then they were instructed to select a gender and colour closest to their skin tone.

Next, they entered a scene containing a virtual London double-decker bus, which we shall refer to henceforth as the *bus scene*. All participants were given the same gender-selectable avatar corresponding to their own gender with two interactable hands synced to their hand movements via the Meta Quest 2 controllers, the same colour as their selected skin tone. Their hands with articulated fingers were visible within short range and their entire bodies were visible in the bus windows surrounding them. An interactive tutorial explained the core mechanic to the participant by playing a sample piece of audio voiceover and prompting them to answer a multiple choice question related to what they had just heard. After successfully completing the trial, participants were prompted to pick up a virtual bus card and tap it on the virtual card reader to begin the main experience. They were then prompted to grab various other virtual interactive objects in the scene, such as a newspaper and different virtual handles, which were highlighted, a common interactive trope used in many VR experiences. Once grabbed, a different voiceover clip would play, and the story, movement, or monitoring tasks would appear after the clip had finished together with a 10-s countdown to complete them, depending on the participant's pre-assigned stress level condition, as shown in Figure 2.

After picking up four of the interactive objects, the participant made their way down the virtual bus and interacted with different virtual non-player characters (NPCs), who responded in an outwardly positive manner to the user's presence, nodding and smiling. Then, the main stress event occurred: an Islamophobic passenger (*the attacker*) began insulting the participant, shouting abuse at them in a clear confrontation (see Figure 3). The tasks continued during intervals in the main stress event. The attacker then stood up, continuing to shout and threaten the participant. Another passenger came to the participant's defence (*the defender*) before the driver stops the bus and the attacker walks away, signalling the end of the main stress event. The participant was then told the VR experience had ended, given assistance in removing the headset, and prompted to complete the post-treatment questionnaire on a laptop *post-treatment*. This concluded the experiment.

3.3 Measures

The VR study was comprised of two distinct phases in VR: the survey scene and the bus scene, followed by the post-treatment survey outside of VR. In terms of performance measures, a comparison was made between the number of correct answers per assigned task in the bus scene, as well as the number of correct answers in the post-treatment survey. Each of the three tasks (narrative, movement and monitoring) appeared seven times during the treatment and was scored in the same way. All seven of the story questions were repeated post-treatment to measure participant recall. The participant scores were summed from the total number of correct answers to derive a percentage score.

A post-treatment, non-VR survey was also given to participants to qualitatively assess their level of embodiment, perceived level of realism and qualitative stress levels, as well as five questions derived from an abridged version of the NASA Task Load Index (NASA-TLX) (Hart and Staveland, 1988) and the Simulator Sickness Questionnaire (SSQ) developed by Kennedy et al. (1993) see Tables 7, 8 for further details.

Stress measurements were obtained by calculating the absolute difference between the mean HR for a baseline measurement (taken during the survey scene, when the participant was instructed to stand idle) and for the bus scene. HR is the number of heartbeats per minute (bpm) and is derived from the raw Blood Volume Pulse (BVP) signal via Photophlethysmography (PPG) Allen (2007). The BVP signal is an optical detection of the pulsatile blood flow resulting from heart beats, sampled at a rate of 300 Hz and then averaged over a square window to create a 1 Hz HR reading. The data was manually cleaned of anomalies such as improbably clustered repeated values and unsustained (2 s or less) spikes or drops of more than 30 bpm, which indicated a temporary malfunction in the wearable, such as ambient light contamination of the PPG sensor. The same processes were repeated for the GSR (measured in microsiemens at 4 Hz) and skin temperature (measured in degrees Celsius) recordings.

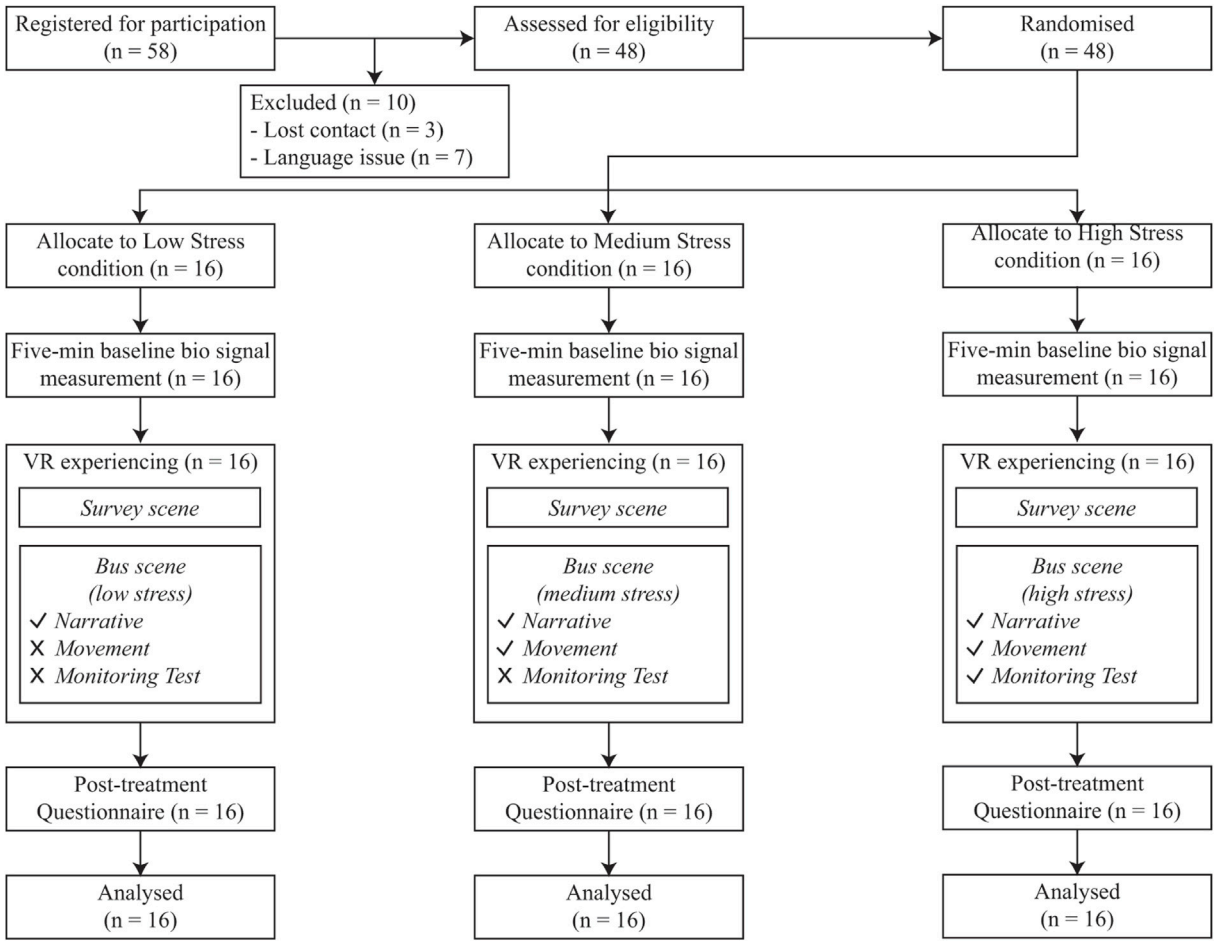

FIGURE 2
Operational Flow chart of the study.

FIGURE 3
Bus scene layout and the first and second parts of the main confrontation: the attacker is in black, defender in green.

In addition to real-time HR, we also analysed participant HRV, comparing an ultra-short measurement of their HRV (in both the time and frequency domains) during baseline (pre-survey scene) with the average of all of their ultra-short HRV readings during each of the seven stress events, using logged event markers in Unity to synchronise the timing. While we will now proceed to define each term in turn, they are summarised in Table 1 for easy reference.

As discussed in the work of Pinheiro et al. (2016), HRV assessment from PPG analysis via PRV (pulse rate variability) as opposed to ECG (electrocardiogram) can be used in healthy subjects across both time and frequency domains. HRV was calculated post-treatment using a fast-fourier transform in Kubios HRV Scientific (based on Matlab) on the inter-beat interval data gathered from the E4 sensor, with a medium-strength filter. The threshold-based artefact correction algorithm compares every Inter Beat Interval

TABLE 1 Heart Rate Variability (HRV) Variables and their response to stress.

Measure	Definition	Effect of stress
Time Domain		
IBI	Inter-beat Interval	Decreases
BVP	Blood Volume Pulse	Decreases
PNS	Parasympathetic Nervous System	Decreases
SNS	Sympathetic Nervous System	Increases
NN	Normal-Normal intervals between R-R peaks of a heartbeat waveform	Decreases
SDNN	Standard Deviation of Normal-to-Normal intervals	Decreases
RMSSD	Root Mean Square of Successive Differences	Decreases
pNN50	Percentage of NN Intervals >50 ms	Decreases
Frequency Domain		
LF	Low Frequency (mainly sympathetic (S) activity) 0.04–0.15 Hz	Increases
HF	High Frequency (mainly parasympathetic (P) activity), 0.15–0.4 Hz	Decreases
LF/HF	Ratio of Low to High Frequency - shows balance between S and PS	Increases

(IBI) value against a local average interval. Using a "Medium" correction level identifies all IBIs that are larger/smaller than 0.25 s compared to the local average. The correction is made by replacing the identified artifacts with interpolated values using a cubic spline interpolation (Standard and Premium, 2018). When the resulting HRV data produced a NaN error message, due to insufficient data caused by the limited time threshold, we extended the ultra-short window (starting from 10 s prior to the stress event, then including the 10 s stress event and 20 s immediately after) from 40 s up to a maximum of 60 s, by further extending the window either side of the stress event by 15 s (25 s prior to the test, 10 s of the test, 25 s post-test).

HRV analysis was conducted on both time and frequency-based domain measures. For the former, these included Standard Deviation of Normal-to-Normal interbeat intervals (SDNN), Root Mean Square of Successive Differences (RMSSD), and parasympathetic and sympathetic nerve system activity (PNS and SNS). Both PNS and SNS are multi-factor indexed measures comprised of multiple parameters. In the case of PNS, mean RR interval, root mean square of successive differences (RMSSD) and Poincare plot index SD1 (in turn linked to the RMSSD and the SD2/SD1 ratio and correlating with the Low-Frequency/High-Frequency (LF/HF) ratio) in normalized units. In the case of SNS, the parameters are mean HR, Baevsky's Stress Index (described below) and Poincare plot index SD2, which correlates with SDNN as well as the LF/HF ratio. In the case of both PNS and SNS, each parameter's values is standardized against normal population values, then scaled by standard deviations of the normal population and a proprietary weighting applied. Both PNS or SNS index values of zero indicate activity that is on average, equivalent to that of the normal population. Positive values reflect activity above the norm, and negative values below. During stress, activity in the autonomic nervous system is typically characterised by decreased parasympathetic (PNS) and increased sympathetic (SNS) modulation (Tarvainen et al., 2014). Sympathetic activity is commonly assessed using the relative power of LF components and the LF/HF ratio, which serve as indicators of sympathovagal balance (Usui and Nishida, 2017).

The Baevsky stress index, an index of regulatory tension based on Cardiac Autonomic Modulation (CAM) originally designed to assess astronauts during space flights (Baevsky and Chernikova, 2017), was also included since it also reflects the activity of the sympathetic part of the autonomic nervous system (Baevsky and Berseneva, 2008).

We also examined the effect of the main stress event (the Islamophobic confrontation) on the user's biological signals. This was calculated by taking the HR data 20 s before the start of the stress event, 10 s during the task, and then 20 s immediately after the stress event. The timing was calculated by the log files saved to the headset, which displayed the time code for when the player chose their quiz response. This data was compared to an ultra-short (1 min) baseline reading that was taken during the pre-survey scene.

The post-treatment survey presented the participant with the same narrative task, this time in a non-VR multiple choice format, as well as a self-reported SSQ and a truncated version of the NASA-TLX. Participants were also asked to rate their sense of embodiment, stress, realism and presence on a Likert scale of 1–5 as well as report any technical problems they encountered.

3.4 Stress assessment limitations

The research team was aware of the documented shortcomings of wrist-based heart rate HR measurements, particularly HRV, but decided to use the Empatica E4 wristband due to the considerable number of studies validating its usage McCarthy et al. (2016), its practicality and ease, and its reliability to record time-stamped data locally during a study with a high participant count (n = 48). Furthermore, the VR experience involved the use of hand controllers, which would have complicated the use of any wired sensors connected to the user's fingertips. Although a new generation of HMDs with integrated PPG sensors are being developed (Bernal et al., 2022; Gjoreski et al., 2021), they remain costly, not widely available, and have some limitations, including the longer preparation time for subjects, data synchronization, and levels of discomfort due to the additional device weight on the user's head during long VR experiences.

Motion artefacts are known to be a limiting factor in the use of PPG to estimate HR, with potential data loss occurring when the sensors fall out of contact with the skin. Several methods have been

implemented to address this problem, as demonstrated by Temko (2017). The use of electroencephalography (EEG) is also highly reputed as a non-invasive, accurate method to evaluate stress levels in real time, as shown by Vanitha and Krishnan (2016), but its use within a VR study was considered impractical given the room-scale nature of the experience.

Similarly, issues related to the low temporal resolution of PPG signals to detect HR were taken into account, but were outweighed by studies conducted by Stuyck et al. (2022), who concluded that HR, RMSSD, SDNN, and LF were validly estimated by the E4 wristband, with valid PR estimates from recording lengths as short as 10 s. Numerous other studies have used PPG to estimate HR in similar mobile VR experiences, such as Quintero et al. (2019); Chauhan et al. (2018).

3.5 Hypotheses

We hypothesise that a participant's autonomic responses provide a good indication of their level of arousal and engagement in the experience. Hypothesis 1 is that the level of stress experienced by a user can be moderated by our conditions and will be reflected in a statistically significant difference in the mean normalised interval time between heart beats at baseline and during stress events. Similar differences will be observed in other HRV and biological signal metrics in Hypothesis 3, 7.

Hypothesis 8 assess whether the participant's stress level will impact their performance on the narrative task both during and after the VR experience. Improved task performance will mean higher scores in the narrative task and post-treatment cued recall task. By contrast, if the participant is not sufficiently stressed, their task performance is hypothesised to be sub-optimal, resulting in lower scores and significantly reduced absolute difference in biological signals between baseline and stress events.

We hypothesise that the difference in a participant's stress levels will correspond to the respective low, medium or high conditions, due to the number of simultaneous tasks they are given per condition. Thus manipulating the stress level conditions will heighten or lessen the degree of stress, while also serving as a test of the physiological measurements.

Following on from the above, Hypothesis 9 is that the medium stress level will produce an optimal performance (the highest narrative task scores), since it encourages increased engagement without imposing too great a cognitive burden on the user. The narrative task score is a dependent variable as it was assigned to all participants, whereas the monitoring and movement task performance scores only served to add additional cognitive load to participants.

Hypothesis 1: That measurements of the HR and HRV will be affected by the stress condition, specifically.

Hypothesis 2: The mean NN (intervals between normal R-R peaks of a heartbeat waveform) measurement (in microseconds) taken at baseline and during the stress events will show a statistically significant increase from the low to high stress condition.

Hypothesis 3: The mean SDNN measurement taken at baseline and during the stress events will show a statistically significant decrease, reflecting greater stress, from the low to high stress condition.

Hypothesis 4: The mean RMSSD measurement taken at baseline and during the stress events will show a statistically significant decrease, reflecting greater stress, from the low to high stress condition.

Hypothesis 5: The mean Low Frequency/High Frequency (LF/HF) measurement taken at baseline and during the stress events will show a statistically significant increase, reflecting greater stress, from the low to high stress condition.

Hypothesis 6: There will be a statistically significant correlation between the HRV-derived stress index (SI) scores and the stress condition (low stress = lowest stress index, high stress = highest stress index).

Hypothesis 7: The mean GSR measurement taken at baseline and during stress events will show a statistically significant decrease, reflecting greater stress, from the low to high stress condition.

Hypothesis 8: There will be a statistically significant correlation between higher narrative task scores (both during and after the treatment) and participants' stress response, measured in PNS (H3.1) and SNS (H3.2) activity.

Hypothesis 9: Based on the Yerkes-Dodson theory of optimal performance margins, we hypothesise that participants given the medium condition will attain the highest VR narrative task scores.

4 Results

Unless otherwise indicated, all of the data fitted the preconditions for the ANOVA, following Shapiro-Wilk4 tests for normal distribution and Levene's test for homogeneity of variances, as shown in Tables 10, 11. The complete results are shown in Table 9.

4.1 Biological signals

Comparisons of mean HRV differences between baseline and stress event scenes (see Section 3.3 for methods) highlighted some key differences in the participant pools by condition, as shown in Table 3. See Table 4 for statistically significant results and their effect sizes across the low and high stress conditions. Only the Stress Index showed significant between-condition variation (Low: 5.22 ± 3.3; High: 1.45 ± 5.48; p = 0.04, r = −0.11), confirming Hypothesis 1.5, with a strong effect size of −0.81.

While no statistically significant differences between conditions were found for SDNN (H1.2) or RMSSD (H1.3), both measures showed evidence of increased stress. SDNN values declined from low to high stress conditions (LS = 5.88, HS = 2.66), although the difference was not significant (p = 0.48). Similarly, RMSSD decreased under medium stress (MS = 0.48) but unexpectedly

TABLE 2 Pre-condition results by condition.

Condition	P-value	Test statistic (W)	Effect size	Standard deviation
Low Stress (LS) VR Score	0.2498	0.9307	0.177	16.4879
Mid Stress (MS) VR Score	0.2394	0.9295	0.2033	24.4328
High Stress (HS) VR Score	0.03803	0.8794	0.2222	18.8237
LS Post VR	0.04723	0.8855	0.2754	16.504
MS Post VR	0.03077	0.8735	0.1602	28.8557
HS Post VR	0.1299	0.913	0.2285	20.3466
LF/HF LS, Baseline	0.1749	0.921	0.1537	1.2532
LF/HF LS, Stress Events	0.3877	0.9404	0.1327	1.9148
LF/HF HS, Baseline	0.01869	0.8592	0.2299	1.5867
LF/HF HS, Stress Events	0.06079	0.8814	0.2169	1.9931
LS SI	0.5143	0.9494	0.1609	6.2865
MS SI	0.3348	0.936	0.1843	3.7207
HS SI	0.9789	0.9807	0.1317	4.0768
LS SI BL to SE	0.08146	0.8956	0.1609	3.3041
HS SI BL to SE	0.5249	0.9522	0.1218	5.2939

TABLE 3 HRV baseline to stress event comparison by condition.

HRV measure	Baseline	Stress events	Abs difference
Mean RR			
Low Stress	724.14	629.98	94.15
Medium Stress	697.53	666.7	30.83
High Stress	732.35	684.64	47.71
SDNN			
Low Stress	36.52	29.2	7.32
Medium Stress	31.78	34.19	−2.41
High Stress	38.34	36.45	1.89
RMSSD			
Low Stress	40.42	31.97	8.45
Medium Stress	32.99	33.44	−0.45
High Stress	43.94	38.11	5.83
Stress Index			
Low Stress	15.2	20.66	5.46
Medium Stress	17.57	17.82	0.25
High Stress	15.38	17.03	1.65
LF/HF Index			
Low Stress	1.67	3.39	1.72
Medium Stress	2.61	3.62	1.01
High Stress	1.87	3.56	1.69

TABLE 4 Heart Rate Variability Results within conditions between Baseline and Stress Events.

Condition	HRV metric	Significance	Effect
High Stress	LF/HF Ratio	0.02	0.2
High Stress	HF	0.03	0.49
Low Stress	Mean RR	0.06	0.3
Low Stress	RMSSD	0.09	0.83
Low Stress	SNS	0.03	0.93
Low Stress	Stress Index	0.02	0.86
Low Stress	LF/HF Ratio	0.01	0.41
Low Stress	LF	0.07	0.24

increased under high stress (HS = 5.81), suggesting either variability due to individual participant differences or technical difficulties that affected the medium stress group. Frequency-domain variables such as the LF/HF ratio (Hypothesis 5) and multi-factor variables such as the Stress Index (Hypothesis 6) and were the most reliable physiological indicators of a stress response, while time-domain HRV variables (SDNN, RMSSD) were less sensitive to the stress manipulation between groups.

Within conditions, statistically significant differences were found between the baseline and stress events in the low stress condition of the NN intervals (Hypothesis 2), the low and high stress of the Low Frequency and High Frequency levels (Hypothesis 5), the low stress of the Stress Index (Hypothesis 6), the low and high stress conditions of the GSR levels (Hypothesis 7), as well as the PNS/SNS activity (Hypothesis 8, 9).

A significant difference in GSR levels between baseline and stress events was found within low and high stress conditions ($p = 0.035$ and $p = 0.003$ respectively), but the difference was greater in the LS condition *versus* the HS condition, invalidating Hypothesis 7 (see Table 5). The average temperature increase was highest during

TABLE 5 Baseline (BL) to stress event (SE) comparisons across conditions.

H	Baseline (BL) to stress event (SE) comparison of low stress (LS) to high stress (HS)	Result	P Value between conditions	P value within conditions	Effect size	Conclusion H0 = null HA = Alternate
1.1	NN intervals (ms) will decrease	LS = 105.39 MS = 40.54 HS = 48.61	0.26	LS = 0.04 MS = 0.21 HS = 0.29	0.61	H0 between HA: LS
1.2	SDNN will decrease	LS = 5.88 HS = 2.66	0.48	LS = 0.23 MS = 0.8 HS = 0.53		H0
1.3	RMSSD will decrease	LS = 4.74 MS = 0.48 HS = 5.81	0.4	LS = 0.08 MS = 0.8 HS = 0.25		H0
1.4	LF/HF will increase	LS = 1.56 MS = 0.87 HS = 1.69	0.56	LS = 0.005 MS = 0.06 HS = 0.008	0.86 −0.67	HA: LS, HS close for MS
1.5	SI will increase	LS = 5.42 MS = 0.62 HS = 1.45	0.043 d = −0.81	LS = 0.01 MS = 0.52 HS = 0.35	−0.81	HA: all 3 HA for LS
2	GSR will decrease	LS = 1.15 MS = 1.32 HS = 0.87	0.83	LS = 0.035 MS = 0.15 HS = 0.003	−0.57 −0.49	HA within LS, HS
3.1	>Stress Condition=>Stress level = effect on VR score (PNS)	LS = −0.39 HS = −0.42	0.163	LS = 0.03 HS = 0.05	0.42 0.4	HA: LS, HS
3.2	>Stress Condition > Stress level = effect on VR score (SNS)	LS = 1.42 HS = 0.71	0.113	LS = 0.02 HS = 0.06	−1.84 −0.52	HA: LS H0 between
4	VR Score linked to Stress, with Medium stress highest	LS = 73.12 MS = 51.81 HS = 63.25	0.02		0.4	H0

TABLE 6 Performance scores by condition.

Performance	Low stress	Mid stress	High stress	Significance
Average VR story Score	73.13	51.81	63.25	p = 0.01
Average Post-Treatment story Score	63.38	44	58.63	p = 0.03
Combined story Scores	136.5	95.44	121.88	
Difference Pre-Post Scores	9.75	8.19	4.63	

the high stress condition - almost double the other two conditions (0.51 compared to 0.32 and 0.33 respectively), but this was also not statistically significant.

4.2 Performance

As shown by the data in Table 6, a one-way ANOVA was conducted on narrative task scores across stress levels. Significant differences emerged (p = 0.02, F = 5.06, SE = 23.89), with performance declining as stress increased Low (73.12 ± 15.96) > High (63.25 ± 18.23) > Medium (51.81 ± 23.66). Using Tukey's *post hoc* method for comparing group pairings, none of the q-scores (0.89, 0.48, 0.41 - low to medium, medium to high, low to high) were above the critical level of 3.42, meaning there was an effect (0.4) from low to high, but no pairwise difference.

4.3 Biological signals and task performance

Significant correlations were observed between Stress Index values and VR narrative task scores (Low p = < 0.001; High p = < 0.001, t = −0.2) validating hypothesis 1.5.

Another statistically significant relationship was found ($p < 0.05$) between the VR story scores and the difference in mean HR in a one-way ANOVA. This was true across all conditions with a weak correlation of r = 0.19, −0.57 and −0.08 for low, mid and high stress conditions. In a one-way ANCOVA, when comparing the change in HR (pre and post-stress event) to the VR score per condition, there was a statistically significant difference ($p = < 0.05$), although only in the naive/low-experience users (who scored less than 2/5 on the VR experience score described at the beginning of this section). This was true across all conditions with a correlation of r = −0.3, −0.02 and 0.23 for no, mid and high stress conditions, which is considered to be weak.

The fact that only the medium stress condition produced a slightly positive correlation between the stress index and the VR score warrants further research. Trend lines across all three conditions are plotted in the Supplementary Material. Medium stress participants displayed a higher baseline stress measurement in comparison to the two other conditions, which impacted the detection of a change in level during the stress events, see Supplementary Figure S7. HRV analysis confirmed the anomalies detected in the Medium Stress condition, which resulted in no statistically significant results across any of the time or frequency domains, disproving the fourth hypothesis, which we attribute to technical error. See Section 5 for further details. Despite the high stress condition provoking the largest change in mean HR from baseline to the bus scene (5.09 beats per minute), the mean difference was not statistically significant in a one-way ANOVA. The same was true for the absolute difference from mean HR during the bus scene (3.98 bpm).

4.4 Predictors of VR performance across different conditions

To investigate the nature of what variables affect narrative task performance outcomes, the intercept of an Ordinary Least Squares (OLS) regression model was calculated. It represents the expected value of the dependent variables when the value of all independent variables is zero. The regression model is represented by the following:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

where β_0 is the intercept - this is the baseline VR score when all predictors are excluded. The OLS method estimates β_0 by minimizing the sum of squared residuals (i.e., the sum of squared differences between observed outcomes: $\hat{Y}_i$ and predicted outcomes Y_i:

$$\sum \left(Y_i - \hat{Y}_i\right)^2$$

Specifically, the intercept is derived from the equation:

$$\beta_0 = \bar{Y} - \sum_{i=1}^{n} \beta_i \bar{X}_i$$

where $\bar{Y}$ and $\bar{X}_i$ are the means of the dependent and independent variables, respectively. This formula ensures that the regression line passes through the mean of the recorded data, making the intercept a critical baseline for interpreting the regression results. The intercept's statistical significance is evaluated using the null hypothesis test, with a significant p-value indicating that the baseline value is reliably different from zero.

The intercept coefficients for the VR and post-VR scores (i.e., the mean scores independent of any predictors, such as changes in SNS or PNS) were 62.99 and 65.65, respectively, with p values of < 0.001. The t-statistic (which measures the size of the difference relative to the variation in the data) score change from VR to Post-VR was $t = 0.98$, but no statistical significance was found ($p = 0.335$), indicating no significant difference in score improvement between the low and high stress groups in the two separate tests. For all statistical analysis equations used in this study, see the appendix.

However, when comparing the PNS and SNS interactions with the score change from VR to post-VR across low and high stress conditions using a multiple linear regression, the p-value was 0.036, indicating a statistically significant relationship between changes detected in the autonomic system and performance gains. This method was chosen to assess the relationship between the VR score change (the dependent variable) and multiple independent variables (changes in SNS and PNS, as well as their interaction):

$$\begin{aligned}\text{Score Change} = \beta_0 + \beta_1 (\text{SNS Change}) + \beta_2 (\text{PNS Change}) \\ + \beta_3 (\text{SNS Change} \times \text{PNS Change}) + \epsilon\end{aligned}$$

Where:

- β_0: Intercept (baseline level of Score Change when all predictors are zero).
- β_1, β_2: Main effects of SNS Change and PNS Change, respectively.
- β_3: Coefficient for the interaction term (SNS Change × PNS Change).
- ϵ: Error term, representing unexplained variance.

By including β_3 we can use the model to test whether the combined effect of SNS and PNS changes is different from the sum of their individual effects. The statistically significant interaction result suggests that the effect of a change in SNS on a change in the VR Score does indeed depend on a change in PNS. While SNS or PNS might not independently affect VR Score Change (as shown by the null hypothesis for H3.1 and H3.2), there is a combined effect that is meaningful.

The coefficient for the interaction between SNS and PNS in the multiple linear regression model predicting VR score change was 11.14, including a statistically signification moderation effect ($p = 0.036$). The positive interaction effect can be seen in Supplementary Figure S8: as SNS increases, the score change trends upwards, especially when considered in the context of changes in PNS. The effect isn't completely linear, and a degree of balance between SNS and PNS responses is key for this improvement, but it does highlight the interplay between SNS (stress response) and PNS modulation (recovery response).

As Supplementary Figure S8 demonstrates, a negative relationship between SNS change and VR score change suggests that SNS hyperactivation leads to a decreased VR score when PNS remains constant. This negative effect appears to be moderated, the higher the PNS change, whereas low PNS change only amplifies the negative relationship. This hints at the fact that excessive SNS activity may hinder cognitive performance and cued recall, leading to poorer sustained performance (VR to post-VR score).

4.5 Post-treatment survey

In the post-treatment story scores, there was a statistically significant result between conditions of $p = 0.03$ and the same pattern of the low stress condition producing the highest recall (63.38 ±16.5), followed by high stress (58.63 ±20.35) and lastly

TABLE 7 Post-treatment questions with answers split by stress level condition.

Post-treatment questions (out of 5)	Low stress	Mid stress	High stress	Significance
Realism rating	3.44	3.69	3.38	p = 0.607
Embodiment rating	3.19	4	3.31	**p= 0.059**
Stress rating	2.81	4.63	3.06	p = 0.592
Abridged NASA-TLX Survey				
How mentally demanding was the task?	4.69	4.63	4.81	p = 0.889
How physically demanding was the task?	3.44	2.69	3.19	p = 0.304
How hurried or rushed was the pace of the task?	3.81	4.31	4.81	**p = 0.066**
How hard did you have to work on the task?	4.13	4.31	4.75	p = 0.212
How insecure, discouraged, irritated were you?	4.06	4.38	3.88	p = 0.726
Mean - all users	4.03	4.06	4.29	p = 0.803
Mean - only naive users	3.75	4	4.42	p = 0.372

medium stress (44±28.86). While all cued recall scores decreased post-VR, the difference between pre and post VR scores was greatest in the Low Stress (9.75), followed by Medium Stress (8.19) and High Stress (4.63). This could be partly due to the fact that the post-treatment scores assessed participants' recall of their original story responses: if these were originally incorrect, they would naturally be incorrect post-treatment too. However, it could suggest a better sense of recall in the higher stress group. Further research is required with a larger participant pool to verify.

In the Abridged NASA-TLX Survey, see Table 7, the responses for "how hurried or rushed was the pace of the task" showed a close to statistically significant (p = 0.066) increase by stress condition (3.81, 4.31, and 4.81 for low, medium and high stress), which can be attributed to the increasing number of tasks that the participants were required to complete. The difference in reported workload was greater between conditions among naive users (3.75, 4 and 4.42 for no, medium and high stress respectively) with a lower yet still not statistically significant difference (p = 0.372), but this can be explained by the uneven and smaller sample sizes. The simulator sickness (SS) questionnaire results show in Table 8 showed that although there was a slight increase in SS by condition, this difference was not found to be statistically significant. The internal consistency of the SSQ analysis was good, demonstrated by the high value of Cronbach's Alpha for all items on the survey $\alpha = 0.851$.

5 Discussion and limitations

Statistically significant differences were found in the mean HR and several HRV time and frequency domain values (most notably the LF/HF ratio and Stress Index, supporting Hypothesis 5, 6) between baseline and stress events. No statistically significant differences were found in the changes between conditions in the NN intervals, SDNN and RMSSD (although there were within conditions for the former), therefore, Hypothesis 2, 3, 4 were rejected.

Furthermore, we found that we could indeed moderate the stress level (as measured by changes in the SI, PNS or SNS from baseline to stress event in Low and High Stress conditions) by varying the number of simultaneous tasks per condition, supporting Hypothesis 6, 8. However, further work is clearly needed to determine why the medium stress condition produced the highest stress response among participants, refuting Hypothesis 9, as reported in the preceding section.

Promising results were found in terms of the statistically significant difference in performance on the narrative task between conditions, partially supporting Hypothesis 8, 9, showing that the stress level does indeed affect performance, and that we can indeed moderate that level by varying our conditions. Further analysis of the variance between task performance scores revealed meaningful distinctions, particularly in the low and high stress conditions. The one-way ANOVA confirmed a statistically significant difference in narrative task scores between conditions ($p = 0.02$), with a moderate effect size ($\eta^2 = 0.4$).

Participants in the medium stress condition displayed unusually high baseline stress index values, potentially exacerbated by technical issues that affected that group. We attribute both factors to the absence of statistical significance in medium-stress HRV measures, in turn masking the stress moderation intended by the manipulation. Although certain HRV measures such as SDNN and RMSSD lacked significant between-condition differences, these were found in a number of within-group changes in both the low and high stress conditions. This highlights the sensitivities between individuals that researchers need to be mindful of in relation to HRV-based indices. It also reinforces the need for further participant segmentation and integration of normalization procedures in future studies.

The recruited sample did not evenly reflect the general population, leaning towards a higher than average number of female participants and being sampled from a university population. This could be one area where the stress inducers might be tailor-made to the individual, although this would need to be assessed by targeted sampling of the population. Since the main focus of this study was to validate the stress conditions' ability to manipulate arousal and quantitatively measure it via biological signals, the NASA-TLX test and post-treatment questions pertaining to embodiment were truncated. Future versions of this study will include the full NASA-TLX as well as the embodiment questionnaire presented by Peck and Gonzalez-Franco (2021).

Importantly, the regression analysis revealed a significant interaction effect between PNS and SNS changes and

TABLE 8 Simulator Sickness Questionnaire results split by stress level condition.

Condition	Nausea	Oculomotor	Disorientation	TS
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Low Stress (n = 16)	35.18 (25.768)	47.38 (28.022)	67.86 (50.542)	55.40 (34.743)
Medium Stress (n = 16)	41.14 (35.986)	44.53 (36.706)	67.86 (70.956)	56.33 (48.041)
High Stress (n = 16)	32.79 (19.696)	53.06 (24.288)	67.86 (30.866)	57.27 (25.282)
All (n = 48)	36.37 (27.597)	48.32 (29.694)	67.86 (52.212)	56.33 (36.420)

TABLE 9 Biological signal results by stress level condition.

Biological signals	Low	Medium	High	Significance
Average HR from baseline during bus scene	4.55	3.96	5.09	p = 0.76
Average abs difference from mean HR during Bus	2.81	2.34	3.98	p = 0.63
Average GSR from baseline during Bus	1.15	1.32	0.87	p = 0.64
Average Temp from baseline during Bus	0.33	0.32	0.51	p = 0.27

TABLE 10 VR score levene test results.

Source	DF	Sum of square	Mean square	F statistic	P-value
VR Score (between groups)	2	251.5417	125.7708	0.5585	0.576
Error (within groups)	45	10,133.9375	225.1986		
Total	47	10,385.4791	220.9676		

TABLE 11 Post-VR score levene test results.

Source	DF	Sum of square	Mean square	F statistic	P-value
Post-VR Score (between groups)	2	932.6667	466.3333	2.1925	0.1234
Error (within groups)	45	9571.2497	212.6944		
Total	47	10,503.9164	223.4876		

improvements in the narrative task score. While neither in isolation predicted performance gains, the interaction between the two yielded a statistically significant result (p = 0.036), with a positive coefficient of 11.14. This suggests that a dynamic interplay of sympathetic activation and parasympathetic regulation may underpin optimal task execution in VR, indicating that successful task performance is not only a matter of arousal intensity, but of finding the optimal homeostatic balance between opposing physiological systems. Future applications may benefit from dynamically adaptive systems that monitor this interplay to modulate task difficulty in real-time.

The reliability and noise of the E4 sensors given their wrist placement is acknowledged as a shortcoming compared to finger-bound sensors as discussed in Section 3.4, despite its proven utility and validity (McCarthy et al., 2016). The manual process (via button pressing on the device) of syncing the biological signal data with the interaction log data in VR is also highlighted as a potential area in which data asynchrony could occur. To mitigate this Unix timestamps were cross-referenced with the log times inside the headset, which was synced prior to the experiment.

6 Conclusion and future work

In conclusion, our study demonstrated that manipulating mental workload in a narrative VRE can influence both task performance and autonomic indicators of arousal, although further refinement is necessary to moderate this effect more precisely. A one-way ANOVA revealed a statistically significant effect of the stress condition on VR narrative scores ($F(2, 45) = 5.06$, $p = 0.02$, $SE = 23.89$), with the low stress group outperforming both medium and high stress groups.

Statistically significant differences were found in the Stress Index (p = 0.043, LF/HF ratio (p = 0.005 in the low stress condition, p = 0.008 in the high stress condition), PNS and SNS activity, with the

most pronounced effects observed within the low and high stress groups.

There was also a statistically significant negative correlation between the SI and VR performance scores in both the low and high stress conditions ($p < 0.001$; $t = -0.2$), indicating that elevated physiological stress was associated with decreased task accuracy, as shown in Supplementary Figures S9, 10 of the appendix. Although the difference between baseline and stress event in the HRV-derived SI between conditions was statistically significant, it was the opposite of our hypothesis (the low stress group had the highest mean SI difference and the medium stress group the lowest).

Also contrary to our hypotheses, the medium stress produced anomalous results, likely attributable to the higher number of technical issues in that participant group, which raised baseline stress scores due to repeated runs, in turn reducing the difference between the baseline and stress event data. The medium stress level group scored the highest in the "how insecure, discouraged, irritated" post-treatment question (see Table 7), perhaps alluding to a negative sense of stress born of frustration inside the experience, as opposed to the sense of eustress tied to higher engagement that the experiment was designed to create.

Post-hoc regression revealed that PNS-SNS interplay had a moderating effect on some improvements in VR score, highlighting a nuanced physiological dynamic that underpins cognitive performance. Future work will extend this regression analysis, investigating the predictive relationship between changes to the PNS and SNS and task performance, as well as place greater emphasis on the stress index metric and less emphasis on body temperature, which was found to be noisier and less prone to phasic changes than expected. We will also add a user interface prompt inside the user's heads-up display (HUD) to guide them to the next interactive object after 10 s of inactivity, in an attempt to improve the user experience. The monitoring task (high stress condition) will also be altered, since many participants complained that they struggled with this task. The subsequent study will prioritise real-time adaptive stress manipulation using biofeedback from a Bluetooth wearable that will refine task difficulty and procedural variation based on a custom optimal arousal margin per participant. Incorporating personalized stress baselines may also help control for variability between individuals that affected metrics such as RMSSD and SDNN. Additionally, a validated embodiment scale and the full NASA-TLX will be integrated, as will the Implicit Association Test (Greenwald et al., 1998) to further contextualise findings within intergroup empathy, while continuing to explore how stress and embodiment may contribute to behavioural change in perspective-taking.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Hong Kong PolyU Ethics Committee. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

DA: Writing – original draft, Writing – review and editing, Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Validation, Visualization. CL: Formal analysis, Funding acquisition, Methodology, Resources, Supervision, Writing – review & editing. GC: Data curation, Formal Analysis, Investigation, Writing – review and editing. YD: Investigation, Writing – review and editing. AS: Supervision, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. DA's work was supported by UCL's Doctoral Training Partnership grant from the United Kingdom Engineering and Physical Sciences Research Council (EPSRC). The study was also partially supported by The Hong Kong Polytechnic University (Project ID: P0035264).

Acknowledgments

The authors wish to thank Sebastian Friston, Ben Congdon and Eddy Schaeffer for their contributions to the Unity project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frvir.2025.1458191/full#supplementary-material

References

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiol. Meas.* 28, R1–R39. doi:10.1088/0967-3334/28/3/r01

Anton, N. E., Athanasiadis, D. I., Karipidis, T., Keen, A. Y., Karim, A., Cha, J., et al. (2021). Surgeon stress negatively affects their non-technical skills in the operating room. *Am. J. Surg.* 222, 1154–1157. doi:10.1016/j.amjsurg.2021.01.035

Archer, D., and Finger, K. (2018). Walking in another's virtual shoes: do 360-degree video news stories generate empathy in viewers?

Archer, D., and Steed, A. (2022). "Optimizing performance through stress and induction levels in virtual reality using autonomic responses," in *2022 IEEE international symposium on mixed and augmented reality adjunct (ISMAR-Adjunct)* (IEEE), 622–627.

Baevsky, R., and Berseneva, A. (2008). Methodical recommendations use kardivar system for determination of the stress level and estimation of the body adaptability standards of measurements and physiological interpretation. Moscow. [Google Scholar]

Baevsky, R. M., and Chernikova, A. G. (2017). *Heart rate variability analysis: physiological foundations and main methods*. Moscow, Russia: Cardiometry.

Bernal, G., Hidalgo, N., Russomanno, C., and Maes, P. (2022). "Galea: a physiological sensing system for behavioral research in virtual environments," in *2022 IEEE conference on virtual reality and 3D user interfaces (VR)* (IEEE), 66–76.

Blascovich, J. (2008). Challenge, threat, and health.

Blascovich, J., Mendes, W. B., Hunter, S. B., and Salomon, K. (1999). Social facilitation as challenge and threat. *J. personality Soc. Psychol.* 77, 68–77. doi:10.1037//0022-3514.77.1.68

Botvinick, M., and Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature* 391, 756. doi:10.1038/35784

Broadhurst, P. (1959). The interaction of task difficulty and motivation: the yerkes dodson law revived. *Acta Psychol.* 16, 321–338. doi:10.1016/0001-6918(59)90105-2

Chauhan, U., Reithinger, N., and Mackey, J. R. (2018). "Real-time stress assessment through ppg sensor for vr biofeedback," in *Proceedings of the 20th international conference on multimodal interaction: adjunct*, 1–5.

Chrousos, G. P. (2009). Stress and disorders of the stress system. *Nat. Rev. Endocrinol.* 5, 374–381. doi:10.1038/nrendo.2009.106

Claude, G., Gouranton, V., and Arnaldi, B. (2015). Versatile scenario guidance for collaborative virtual environments. *GRAPP* 15, 6.

Cleworth, T. W., Horslen, B. C., and Carpenter, M. G. (2012). Influence of real and virtual heights on standing balance. *Gait and posture* 36, 172–176. doi:10.1016/j.gaitpost.2012.02.010

Comstock, J. R., and Arnegard, R. J. (1992). The multi-attribute task battery for human operator workload and strategic behavior research. *Tech. Rep.*

Gaillard, A. W. (2000). Stress, workload, and fatigue as three biobehavioral states: a general overview. *Stress, workload, fatigue*, 623–639. doi:10.1201/b12791-3.12

Gjoreski, H., I. Mavridou, I., Fatoorechi, M., Kiprijanovska, I., Gjoreski, M., Cox, G., et al. (2021). "Emteqpro: face-mounted mask for emotion recognition and affective computing," in *Adjunct proceedings of the 2021 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2021 ACM international symposium on wearable computers*, 23–25.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *J. personality Soc. Psychol.* 74, 1464–1480. doi:10.1037/0022-3514.74.6.1464

Haggard, P., and Chambon, V. (2012). Sense of agency. *Curr. Biol.* 22, R390–R392. doi:10.1016/j.cub.2012.02.040

Hart, S. G., and Staveland, L. E. (1988). Development of nasa-tlx (task load index): results of empirical and theoretical research. *Adv. Psychol.* 52, 139–183. doi:10.1016/s0166-4115(08)62386-9

Hayes, A. L., Ulinski, A. C., and Hodges, L. F. (2010). "That avatar is looking at me! social inhibition in virtual worlds," in *International conference on intelligent virtual agents* (Springer), 454–467.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Lilienthal, M. G. (1993). Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. *Int. J. Aviat. Psychol.* 3, 203–220. doi:10.1207/s15327108ijap0303_3

King, M., Burrows, G., and Stanley, G. (1983). Measurement of stress and arousal: Validation of the stress/arousal adjective checklist. *Br. J. Psychol.* 74, 473–479. doi:10.1111/j.2044-8295.1983.tb01880.x

Klein, S. B. (1982). *Motivation: biosocial approaches*. McGraw-Hill College.

Lazarus, R. S. (2006). *Stress and emotion: a new synthesis*. Springer publishing company.

Luong, T., Argelaguet, F., Martin, N., and Lécuyer, A. (2020). "Introducing mental workload assessment for the design of virtual reality training scenarios," in *2020 IEEE conference on virtual reality and 3D user interfaces (VR)* (IEEE), 662–671.

Mackay, C., Cox, T., Burrows, G., and Lazzerini, T. (1978). *An inventory for the measurement of self-reported stress and arousal*. Leicester, United Kingdom: British Journal of Social and Clinical Psychology.

Mandrick, K., Peysakhovich, V., Rémy, F., Lepron, E., and Causse, M. (2016). Neural and psychophysiological correlates of human performance under stress and high mental workload. *Biol. Psychol.* 121, 62–73. doi:10.1016/j.biopsycho.2016.10.002

Martens, M. A., Antley, A., Freeman, D., Slater, M., Harrison, P. J., and Tunbridge, E. M. (2019). It feels real: physiological responses to a stressful virtual reality environment and its impact on working memory. *J. Psychopharmacol.* 33, 1264–1273. doi:10.1177/0269881119860156

McCarthy, C., Pradhan, N., Redpath, C., and Adler, A. (2016). "Validation of the empatica e4 wristband," in *2016 IEEE EMBS international student conference (ISC)* (IEEE), 1–4.

Nixon, P., Murray, R., and Bryant, C. (1979). Stress response curve. *Tech. Rep.* Available online at: https://explorable.com/how-does-stress-affect.

Palmas, F., Labode, D., Plecher, D. A., and Klinker, G. (2019). "Comparison of a gamified and non-gamified virtual reality training assembly task," in *2019 11th international conference on virtual worlds and games for serious applications VS-games* (IEEE), 1–8.

Pan, Y., and Steed, A. (2019). "Avatar type affects performance of cognitive tasks in virtual reality," in *25th ACM symposium on virtual reality software and technology*, 1–4.

Parsons, T. D., and Reinebold, J. L. (2012). Adaptive virtual environments for neuropsychological assessment in serious games. *IEEE Trans. Consumer Electron.* 58, 197–204. doi:10.1109/tce.2012.6227413

Peck, T. C., and Gonzalez-Franco, M. (2021). Avatar embodiment. a standardized questionnaire. *Front. Virtual Real.* 1, 575943. doi:10.3389/frvir.2020.575943

Peck, T. C., and Tutar, A. (2020). The impact of a self-avatar, hand collocation, and hand proximity on embodiment and stroop interference. *IEEE Trans. Vis. Comput. Graph.* 26, 1964–1971. doi:10.1109/TVCG.2020.2973061

Pinheiro, N., Couceiro, R., Henriques, J., Muehlsteff, J., Quintal, I., Goncalves, L., et al. (2016). "Can ppg be used for hrv analysis?," in *2016 38th annual international conference of the IEEE engineering in medicine and biology society EMBC* (IEEE), 2945–2949.

Quintero, L., Papapetrou, P., and Munoz, J. E. (2019). "Open-source physiological computing framework using heart rate variability in mobile virtual reality applications," in *2019 IEEE international conference on artificial intelligence and virtual reality AIVR* (IEEE), 126–1267.

Sagnier, C., Loup-Escande, E., and Valléry, G. (2019). "Effects of gender and prior experience in immersive user experience with virtual reality," in *International conference on applied human factors and ergonomics* (Springer), 305–314.

Sanchez-Vives, M. V., and Slater, M. (2005). From presence to consciousness through virtual reality. *Nat. Rev. Neurosci.* 6, 332–339. doi:10.1038/nrn1651

Sandi, C. (2013). Stress and cognition. *Wiley Interdiscip. Rev. Cognitive Sci.* 4, 245–261. doi:10.1002/wcs.1222

Seery, M. D., Weisbuch, M., and Blascovich, J. (2009). Something to gain, something to lose: the cardiovascular consequences of outcome framing. *Int. J. Psychophysiol.* 73, 308–312. doi:10.1016/j.ijpsycho.2009.05.006

Selye, H. (1974). "Stress without distress," in *Psychopathology of human adaptation* (Springer), 137–146.

Servotte, J.-C., Goosse, M., Campbell, S. H., Dardenne, N., Pilote, B., Simoneau, I. L., et al. (2020). Virtual reality experience: immersion, sense of presence, and cybersickness. *Clin. Simul. Nurs.* 38, 35–43. doi:10.1016/j.ecns.2019.09.006

Shaffer, F., and Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Front. public health* 258, 258. doi:10.3389/fpubh.2017.00258

Shields, G. S., Sazma, M. A., and Yonelinas, A. P. (2016). The effects of acute stress on core executive functions: a meta-analysis and comparison with cortisol. *Neurosci. and Biobehav. Rev.* 68, 651–668. doi:10.1016/j.neubiorev.2016.06.038

Standard, H., and Premium, H. (2018). Kubios hrv

Steed, A., Archer, D., Izzouzi, L., Numan, N., Shapiro, K., Swapp, D., et al. (2023). Immersive competence and immersive literacy: exploring how users learn about immersive experiences. *Front. Virtual Real.* 4, 22. doi:10.3389/frvir.2023.1129242

Steed, A., Pan, Y., Zisch, F., and Steptoe, W. (2016). "The impact of a self-avatar on cognitive load in immersive virtual reality," in 2016 IEEE virtual reality (VR) *(greenville, SC, USA: ieee)*, 67–76. doi:10.1109/VR.2016.7504689

Sterna, R., Strojny, P., and Rkebilas, K. (2019). Can virtual observers affect our behavior? *Soc. Psychol. Bull.* 14, 1–18. doi:10.32872/spb.v14i3.30091

Stuyck, H., Dalla Costa, L., Cleeremans, A., and Van den Bussche, E. (2022). Validity of the empatica e4 wristband to estimate resting-state heart rate variability in a lab-based context. *Int. J. Psychophysiol.* 182, 105–118. doi:10.1016/j.ijpsycho.2022.10.003

Tarvainen, M. P., Niskanen, J.-P., Lipponen, J. A., Ranta-Aho, P. O., and Karjalainen, P. A. (2014). Kubios hrv–heart rate variability analysis software. *Comput. methods programs Biomed.* 113, 210–220. doi:10.1016/j.cmpb.2013.07.024

Temko, A. (2017). Accurate heart rate monitoring during physical exercises using ppg. *IEEE Trans. Biomed. Eng.* 64, 2016–2024. doi:10.1109/tbme.2017.2676243

Usui, H., and Nishida, Y. (2017). The very low-frequency band of heart rate variability represents the slow recovery component after a mental stress task. *PloS one* 12, e0182611. doi:10.1371/journal.pone.0182611

Van Acker, B. B., Parmentier, D. D., Vlerick, P., and Saldien, J. (2018). Understanding mental workload: from a clarifying concept analysis toward an implementable framework. *Cognition, Technol. and work* 20, 351–365. doi:10.1007/s10111-018-0481-3

Vanitha, V., and Krishnan, P. (2016). Real time stress detection system based on eeg signals. *Biomed. Research-India* 27, S271–S275.

Won, A. S., Bailenson, J., Lee, J., and Lanier, J. (2015). Homuncular flexibility in virtual reality. *J. Computer-Mediated Commun.* 20, 241–259. doi:10.1111/jcc4.12107

Wu, D., Courtney, C. G., Lance, B. J., Narayanan, S. S., Dawson, M. E., Oie, K. S., et al. (2010). Optimal arousal identification and classification for affective computing using physiological signals: virtual reality stroop task. *IEEE Trans. Affect. Comput.* 1, 109–118. doi:10.1109/t-affc.2010.12

Yerkes, R. M., and Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *J. Comp. Neurol. Psychol.* 18, 459–482. doi:10.1002/cne.920180503

Yuan, Y., and Steed, A. (2010). "Is the rubber hand illusion induced by immersive virtual reality?," in *2010 IEEE virtual reality conference*, 95–102. doi:10.1109/VR.2010.5444807

Zhang, J., Ma, K., and Hommel, B. (2015). The virtual hand illusion is moderated by context-induced spatial reference frames. *Front. Psychol.* 6, 1659. doi:10.3389/fpsyg.2015.01659