Journal of Universal Computer Science, vol. 29, no. 1 (2023), 16-33
submitted: 22/6/2022, accepted: 21/11/2022, appeared: 28/1/2023 CC BY-ND 4.0

Behavioral and Psychophysiological Measures of Engagement During Dynamic Difficulty Adjustment in Immersive Virtual Reality

Oscar I. Caldas
(Universidad Militar Nueva Granada, Bogota, Colombia
https://orcid.org/ 0000-0002-3105-4656, oscar.caldas@unimilitar.edu.co)

Mauricio Mauledoux
(Universidad Militar Nueva Granada, Bogota, Colombia
https://orcid.org/ 0000-0002-9964-947X, mauricio.mauledoux@unimilitar.edu.co)

Oscar F. Aviles
(Universidad Militar Nueva Granada, Bogota, Colombia
https://orcid.org/ 0000-0001-8676-9926, oscar.aviles@unimilitar.edu.co)

Carlos Rodriguez-Guerrero
(Vrije Universitiet Brussel, Brussels, Belgium
https://orcid.org/ 0000-0002-0297-9748, carlos.rodriguez.guerrero@vub.be)

Abstract: Dynamically Difficulty Adjustment (DDA) has been widely used to preserve engagement in serious and entertaining games, reach better learning, and enhance user performance. A variety of studies suggests that in DDA, task performance (score) rises until hitting a plateau associated with the skill level. However, the sense of engagement is individual and context-dependent, and the effect of DDA on other engagement indicators for immersive virtual environments is still unclear. This study measured objective indicators of engagement while study subjects played an immersive virtual game with DDA to find evidence of dynamic response, similar to game performance. Participants were demanded to perform repetitive upper-limb motions while recording the following indicators: Response Latency as perceptive engagement (elapsed time after sensory stimulus), Exercise Intensity as motion engagement (hand velocity), and psychophysiological responses as emotional engagement (Heart Rate, Skin Conductance, and Respiratory Rate). In addition, 30 features were extracted from the signals to evaluate their variations between time windows. Results indicate that response latency, vertical hand velocity, and heart rate showed significant changes over time during DDA and grew until hitting a plateau, i.e., at the subject's maximum performance. Moreover, some of the features extracted from the signals showed significant differences between time windows, and having strong correlation with the mean of score: max Latency, min velocity on the Y-axis, and mean heart rate, which suggest a promising application for evaluating changes in engagement between different experimental conditions in VR.

Keywords: Engagement, Immersion, Difficulty, Presence, Psychophysiology, Virtual Reality
Categories: I.3.8, I.3.m, J.3, J.4, L.2.0, L.2.3, L.2.5, L.6.1

DOI: 10.3897/jucs.89412

1 Introduction

Immersive Virtual Reality (IVR) has been increasingly used to provide higher engagement to learning tasks, whether psychomotor, affective, or cognitive [Karpouzis 20, Lee 15, Makransky 19]. Traditionally, engagement and involvement have been used to define the psychological state that can be reached when focusing energy and attention on a coherent set of stimuli [Skarbez 17, Witmer 98]. Therefore, the effectiveness of IVR is usually linked to deeper states of engagement in such a way that the activity becomes autotelic, and the subject loses self-consciousness and sense of time, which is the foundation of Csikszentmihalyi's concept of Flow: when the individual is so engaged that nothing else seems to matter [Csikszentmihalyi 98]. Many authors agree that engagement and flow formation is strongly related to giving a clear goal, enhanced feedback, and matching challenge with the user's skills [Cheng 14, Hoffman 09]. Moreover, IVR-based games are helpful for defining and presenting goals, providing enhanced multisensory feedback, and allowing challenge/skill balance, and overall, it is believed that a IVR setup could be a means for objective real-time assessment of engagement.

Engagement is widely believed to be driven by motivation and revealed through active participation and invested interest. However, engagement is also context-dependent [Lequerica 10, O'Brien 16], and therefore, several human behaviors and attitudes that are unique for IVR should be analyzed to identify the subjective response to the virtual experience and thus measure the user engagement. A great deal of studies prefers using validated questionnaires to measure user engagement in terms of VR-relevant constructs (e.g., Immersion, Presence, Absorption, or Agency) [Brockmyer 09, Hamari 14, Makransky 17, Sailer 20; Turkay 15]; however, Slater had already warned that post-experience self-reports could not assure an accurate measure of the mental activity at the time of the experience. Therefore, this study focused on real-time measures, which could be out of the influence of the subject's interpretation, recalls, and time of exposure [van Baren 04)]. Response latency, exercise intensity, and psychophysiological signals are probably the most accepted indicators of behavioral and emotional engagement in interactive VR setups, and many authors have reported their use [Barreda 20, Darzi 19, Goršič 17, Kerous 15].

In addition, van Baren and IJsselsteijn pointed out that despite actively responding to objects and circumstances in IVR, just as in real life, individuals are also influenced by their task performance [van Baren 04]. It has been noticed that when users engage with an interactive task-based simulator, proper challenge leads to the desired state of flow, but the focused attention will only remain if obtaining the expected performance for the given skills [Cheng 14; Csikszentmihalyi 98]. Therefore, previous studies have moved from simply setting engaging setups for IVR to increasing performance by incorporating Dynamic Difficulty Adjustment (DDA) to reach the proper challenge/skill balance, i.e., not too high to avoid frustration but hard enough to encourage engagement" [Ozkul 19, Pinto 18, Rodriguez-Guerrero 17]. However, how behavioral and emotional indicators behave during DDA in IVR games is still unclear.

For instance, it is expected from DDA that in the absence of external assistance, all performance cues (e.g., score, error rate, completion time) rise to find a plateau at a given value where they will stay oscillating, which is related to the subject's maximum skill level [Rodriguez-Guerrero 12). We hypothesize that real-time measures of engagement, such as response latency (perceptive engagement [Li 16]), exercise intensity (motor [Goršič et al. 17]), and psychophysiological signals (emotional engagement [Knaepen 15]) might also reach their final value at the maximum level of activity, similar to performance's curve.

This study was designed to test the effects of DDA in a physically demanding IVR game (requiring repetitive motions of the dominating upper limb) using an Increment/Decrement One Level algorithm (IDOL). Consequently, we were able to explore How objective measures of engagement with task-oriented exercises in IVR behave during DDA? Moreover, contribute to the research community by answering the following research questions (RQ):

RQ1: Which behavioral and psychophysiological indicators of engagement increase/decrease significantly in time as a response to DDA in IVR?

RQ2: Which of the defined indicators, either continuous signals or extracted features, correlate with in-game performance during DDA in IVR?

2 Methods

2.1 Hardware and Instrumentation

VR simulation was displayed via an Oculus Rift Head-Mounted Display (HMD), with integrated over-ear headphones and remote controllers suitable for visual, auditory, and haptic feedback. The entire system offered full immersion by position/orientation tracking (head and hands), latency < 10 ms, refreshment rate = 90 Hz, per-eye resolution = 1080 x 1200, field-of-view = 110º, 3D sound, and both-hands vibration. These technical attributes have been widely accepted to deliver the Place Illusion (PI) that Slater set as needed to promote Presence in VR, i.e., the sense of "being there" (Slater, 2009).

Regarding the psychophysiological signals, electrocardiogram (ECG), Galvanic Skin Response (GSR), and Respiration (RSP) were sampled using a Biosignals Plux Explorer (sampling rate = 1kHz, resolution = 16 bits).

As shown in Fig.1, ECG was acquired by two electrodes placed in the position of leads V2 and V6 (12-Lead ECG scheme), plus the ground electrode over the sternum. Likewise, two surface electrodes were placed on the intermediate phalanges of the index and middle fingers in the non-dominant hand to measure GSR. Finally, an adjustable elastic strap was worn around the thorax to measure RSP.

Electrode wires were carefully fixed to the body to avoid motion artifacts (e.g., by cable pulling) and discomfort capable of breaking the sense of Presence. However, none of the participants reported being aware of the sensors or cables during the task.

2.2 Study subjects

Ten members of the research group participated in this study (n = 10), 4 females (40%) and 6 males (60%), mean age 29.11 (SD = 9.99). All participants were healthy adults with no physical or cognitive conditions that could provide informed consent, follow instructions, and perform the upper limbs and head motions. There was no need to exclude participants due to undesired neurological conditions for VR simulations (such as vertigo, seizure history, or dizziness).

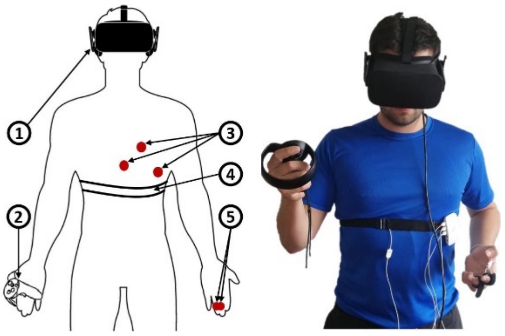


Figure 1: The VR device and the sensors placed on the subject's body: schematic (left) and photograph (right) Head Mounted Display (1), remote controller (2) and sensors: 3 ECG electrodes (3), respiration strap (4), and 2 GSR electrodes

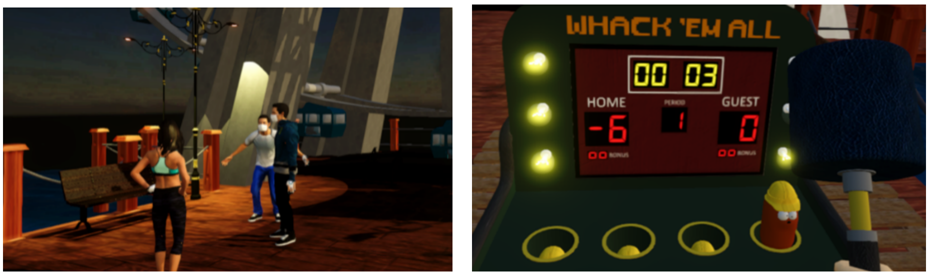


Figure 2: Virtual environment developed to provide multisensory interaction and difficulty adjustment in an upper-limb exercise. Left: General view of the VR scenario; right: First-person view of the VR version of the game Wack-a-mole.

2.3 Study protocol

Surface electrodes were fixed to the skin at least 5 minutes before sensor placement to avoid tonic drifts due to electrolyte penetration. After being briefly introduced to the instrumentation and procedure, each participant filled a survey to provide demographic data (age, gender, and hand laterality) and the first report of Virtual Reality Sickness Questionnaire (VRSQ) to assess changes in oculomotor and disorientation symptoms after the VR exposure, via the 4-points scale proposed by Kim et al. [Kim 18]. Then, while seated, a 1-minute-long screening of body signals was performed to obtain baseline data at resting conditions. Afterward, VR devices were provided, and the experiment started.

During the experiment, the participant was immersed in a custom-made VR simulation (Unity3D version 2019.3.12f1) of a funfair on a pier and asked to play the popular arcade machine game "whack-a-mole" (see Fig.2). The VR environment included the essential aspects that contribute to Plausibility Illusion (PI): coherent elements and actions providing the authentic experience expected by the user [Gilbert

16, Skarbez et al. 17]. Furthermore, the sense of coexistence was provided by including human-like animated avatars (for social Presence), along with an accurate representation of the real-life environment (for physical Presence) and precise bodily connectivity (for self-presence) [Caldas et al. 20, Makransky 17].

Participants were instructed to grab the mallet and hit the red button to start the arcade machine, and then play the 90-seconds game: to hit the moles as randomly released (one at a time) within a time limit here referred to as Time Out-of-the-Hole (TOH_0 = 1.6 s), that is the lapse between each mole pop-up and its automatic return. When the player hits the single mole target strongly enough to move it down (approached by the hand velocity) and within TOH, the score increases by 1 point, and difficulty increases by a given TOH amount reduction (-30 ms). On the contrary, if TOH runs out and the subject misses the mole target, the score decreases by 1 point, and TOH increases by the same amount. Moreover, the subsequent mole releases immediately after the previous one returns by either hit or miss, which avoids the risk of having multiple targets and any lapse between subsequent moles.

The above-mentioned dynamic difficulty adjustment is based on the commonly used Increment/Decrement One Level (IDOL) algorithm (Ozkul et al., 2019), and the flowchart further describes it in Fig.3. Therefore, participants' scores are expected to rise until they reach a plateau at an average maximum value, which Rodriguez-Guerrero named "skill limit" [Rodriguez-Guerrero 12].

Once participants ended the task, they were asked to give the second report of VRSQ, as well as to fill the User Engagement Questionnaire - Short Form (UES-SF), which has been proven to provide accurate reports of VR interactions in terms of aesthetic appeal, focused attention, perceived usability, and reward, via a 5-points Likert scale [O'Brien 18, Yu 19].

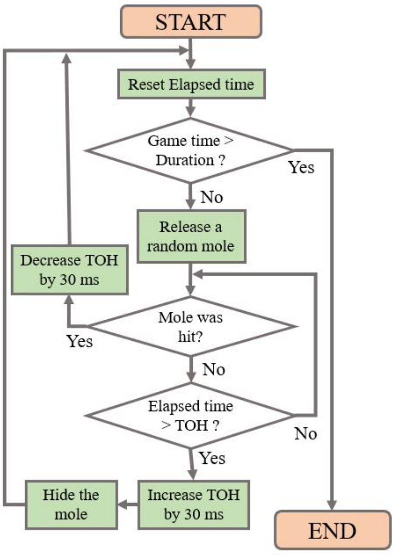


Figure 3: Flowchart describing the DDA algorithm by modifying the mole TOH (time out-of-the-hole).

2.4 Measurements and feature extraction

Performance was assessed via the in-game score (now called score), which measures the effectiveness of hitting the mole (sampled at every hit or miss). Moreover, the mean and the standard deviation of the score were also extracted from the time series.

In addition, two types of engagement indicators were considered for comparison: behavioral and emotional.

Behavioral engagement

Two indicators of behavioral engagement were measured: response latency (here called LAT) and exercise intensity. The first one is based on the model proposed by Li et al., where the time lag of responding to changes in interesting VR content is used as a measure of perceptive engagement since proving active participation [Li 16]. It was measured as the time between mole release and the corresponding hit or miss (see Eq.1). Notice that if the subject misses the mole, $LAT = TOH$

$$LAT(s) = \begin{cases} t_r - t_{cc} & : t_r - t_{cc} < TOH \\ TOH & : t_r - t_{cc} \geq TOH \end{cases} \tag{1}$$

where $LAT(s)$ is response latency, $t_{cc}(s)$ is the time at content change (mole pops up), $t_{cc}(s)$ is the time at responding (mole is hit), and TOH is the preset Time Out-of-the-Hole.

Finally, exercise intensity is measured as the root-mean-square (RMS) of the dominant hand linear velocity in the three axes ($VelX$, $VelY$, and $VelZ$), which has been proven to be appropriate to reveal energy expenditure (motor engagement) in upper-limb exercises [Tsurumi 02, Van Der Pas 11]. Hand velocity is screened directly from the controller's accelerometer, and the RMS of its value is agreed to provide a proper measure of the signal's power.

The mean and the standard deviation were extracted from each of the mentioned behavioral signals (e.g., μLAT, σLAT) for a total of eight behavioral features to be compared with the corresponding two Score features.

Emotional engagement: Psychophysiological signals

Three signals were used as psychophysiological indicators of the subject's engagement: Heart Rate (HR) obtained from ECG, Skin Conductance (SC) from GSR, and Respiratory Rate (RR) from RSP. Moreover, 25 features were extracted from these signals during the task: HR (8), SC (11), and RR (6). These signals were preprocessed by applying smoothing, notch, and band-pass filtering, normalized using baseline measures, and analyzed by morphology-based custom searching algorithms in Matlab (2018b).

The ECG raw signal was processed by using the Pan-Tompkins algorithm (band-pass filtering + differentiation + squaring + moving window integration [Pan 85] and adaptive heuristics to detect R-peaks and measure R-R intervals, and thus interpolate to get the HR time series. The mean (μHR) was extracted from the signal along with features indicating HR variability (HRV) in the time domain: standard deviation (σHR) and root mean square of successive differences (RMSSD); and in the frequency domain:

absolute Power Spectral Density for Low Frequencies [0.04-0.15 Hz] (aLF), for High Frequencies [0.15-0.4 Hz] (aHF) and ratio $\left({}^{LF}/_{HF}\right)$.

GSR analysis followed the state-machine approach proposed by Caldas et al., which analyses the GSR morphology to detect skin conductance responses (phasic component) and separates it from the tonic component (skin conductance level) [Caldas 20]. Features from SCL: mean and standard deviation (μSCL and σSCL), maximum-minimum range ($SCLrange$), and difference between final and initial values (ΔSCL); features extracted after identifying SCR waves: SCRs/time ($nSCR$), mean amplitude and rise time (μSCR_A and μSCR_{tr}) and their standard deviations (σSCR_A and σSCR_{tr}).

After peak detection (breaths), six features were extracted from the RR signal: mean and standard deviation (μRR and σRR), longest and shortest time between consecutive breaths ($maxTCB$ and $minTCB$), and the deepest and shallowest breath (*Deep, Shallow*).

2.5 Data analysis

All analyses were carried out using Matlab (2018b). Descriptive statistical analysis was firstly performed to graphically observe variations of the behavioral and psychophysiological signals along with performance, i.e., mean and standard deviation from all participants in terms of score, LAT, $VelX$, $VelY$, $VelZ$, HR, SC, and RR. All signals were windowed by 1-s intervals, resulting in 90 windows to assess inter-signal correlations. In contrast, features were extracted after windowing each signal by 15-s intervals (i.e., the shortest segment that showed to have enough samples for suitable signal smoothing), resulting in six windows to be compared for significant differences by using the null hypothesis: not showing significant differences between windows, using $\alpha = 0.05$, as in Eq.(2).

$$H_0\colon \mu_1 = \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 \tag{2}$$

where μ_i is the mean of the observed feature at each window.

Repeated Measures Analysis of Variance was performed to test significant differences in each signal and feature among all participants, with time windows as a within-subject factor and gender as a between-subject factor.

Non-parametric statistics were used for correlation analysis considering the small data sample and the non-normal distribution of all eight signals and most features (Shapiro-Wilk test). Therefore, Kendall's Tau-b correlation (τ) was calculated to measure the strength of association between behavioral and psychophysiological features. The post hoc test to perform pairwise comparisons was Paired Samples T-Test with Bonferroni correction $\left(p < \frac{\alpha}{n};\ n = 6\right)$. In addition, this paired test was also used to evaluate the VRSQ questionnaire (comparing results from before and after exposure).

3 Results

All participants completed the study and were all included in the study analysis. However, one subject was found to have no skin conductance responses, which reduced the sample size in the corresponding features.

Results from the UES-SF questionnaire showed high engagement among all participants (4.76 ± 1.18). None of the participants reported to be affected in any of the symptoms assessed by the VRSQ test, and the null hypothesis was not rejected in the test results (p = 0:16), which suggests that this VR exposure was probably not promoting simulation sickness. However, caution must be taken due to the small study sample.

3.1 Demographic data

Only one individual was left-handed, which discarded laterality as a possible between-subject factor. Likewise, the small sample was not suitable for age range analysis, and therefore, gender was the only demographic factor in consideration. However, no evidence was found of gender being related to the responses from any signal nor feature extracted, i.e., no significant differences were detected by between-subject analysis.

3.2 Behavioral measures

From the charts shown in Fig.4, it can be seen that the Score and LAT changed significantly in time, the first growing $(p = 0.00, \eta_p^2 = 0.73)$ and the last decreasing $(p = 0.00, \eta_p^2 = 0.87)$. However, both charts reveal a damped trend to an average level, which in the case of the score is more evident when observing the two individual curves since they exhibit the level to which each subject's response starts to oscillate.

Fig.4 also indicates that the dispersion of data in both charts is considerably high if compared with the signal range, which explains the lack of significant and strong differences in standard deviations, which differs from the other extracted features. The differences between time windows for the feature with the largest effect size at each behavioral measure are displayed in Fig.5, where the damped trends are the most evident. In the case of $\mu Score$ (Fig.5-up), the post hoc test revealed significant differences between the second and sixth segment $(p = 0.04)$, between the third and the fourth segment $(p = 0.03)$, and as explicitly shown in the boxplots, between the first and all the other windows $(p = 0.00)$. In the case of μLAT (Fig.5-middle), it can be seen that the only significant differences were found between 1st and 3th $(p = 0.01)$, 1st and 4th $(p = 0.01)$, 1st and 5th $(p = 0.00)$, and 1st and 6th $(p = 0.00)$.

A significant negative correlation was found between Score and Latency $(p = 0{:}\,00, \tau = -0.68)$. Regardless of the statistically significant correlation, both indicators showed a very weak relationship with RMS of Hand Velocity in all three directions, i.e., $|\tau| < 0.01$.

Oppositely, Exercise Intensity showed no clear evidence of changes in the time since $VelX$ had no significant differences $(p = 0.01, \eta_p^2 = 0.34)$, $VelZ$ had a weak effect size $(p = 0.38, \eta_p^2 = 0.10)$, and only $VelY$ reported significant and moderate-strong relationship with the DDA changes along time $(p = 0.00, \eta_p^2 = 0.48)$. However, it is apparent from Fig.4 that $VelY$ reaches a higher final value and takes up to half of the session to become steady.

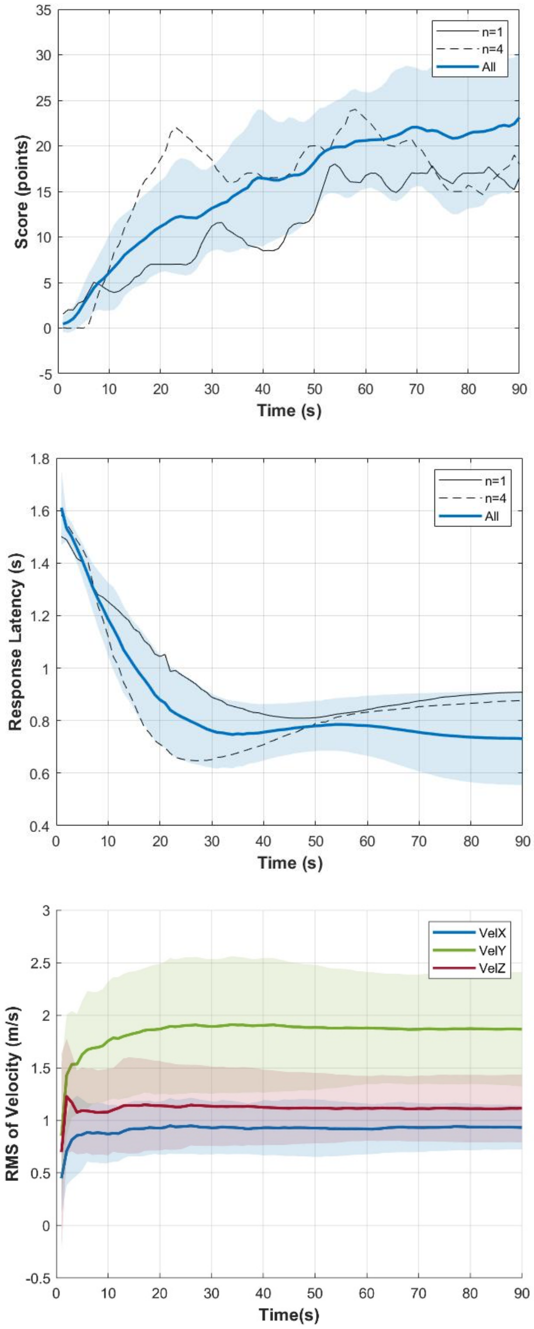


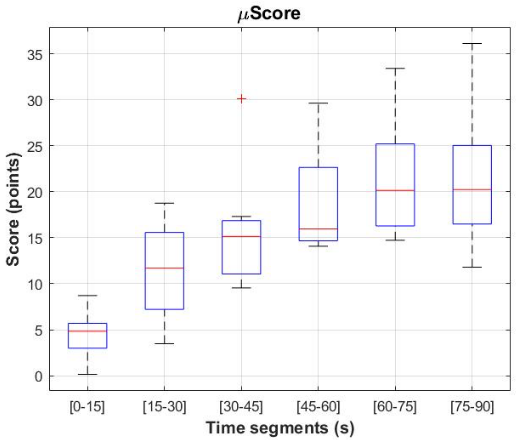
Figure 4: Performance and behavioral engagement measures in terms of the mean and standard deviation (shaded). Single measures from 2 random subjects $(n = 1$ and $n = 4)$ *were also included in the charts for further illustration.*

Table1 describes the observed behavior of all features that reported significant changes between time windows, being the minimum value the one with the strongest effect size in $VelX$ $\left(\eta_p^2 = 0.91\right)$, $VelY$ $\left(\eta_p^2 = 0.88\right)$, and $VelZ$ $\left(\eta_p^2 = 0.73\right)$. In the case of $minVelY$, shown in the lower boxplot in Fig.5, posthoc tests found that all-time windows were significantly higher than 1st ($p = 0.00$). Moreover, all four score-related features showed to be significantly correlated with $minVelY$ ($p < 0.05$), with a moderate-strong relationship with $\mu Score$ ($\tau = 0.41$) and $minScore$ ($\tau = 0.44$).

Signal		Within-subject effect	
Name	Feature	H_0	Effect Size
Score	$\mu Score$	$p = 0.00$	${\eta_p}^2 = 0.76$
LAT	μLAT	$p = 0.00$	${\eta_p}^2 = 0.51$
RMS VelX	$\mu VelX$	$p = 0.00$	${\eta_p}^2 = 0.91$
RMS VelX	$\mu VelY$	$p = 0.03$	${\eta_p}^2 = 0.88$
RMS VelX	$\mu VelZ$	$p = 0.00$	${\eta_p}^2 = 0.79$
HR	μHR	$p = 0.04$	${\eta_p}^2 = 0.32$

HR: Heart Rate, LAT: Latency, RMS: Root Mean Square
p: Significance, ${\eta_p}^2$: Partial eta-squared
In bold: ${\eta_p}^2 > 0.5$ (strong effect size)

Table 1: Differences between extracted features in time windows. Only features with significant differences (p < 0:05) are reported.



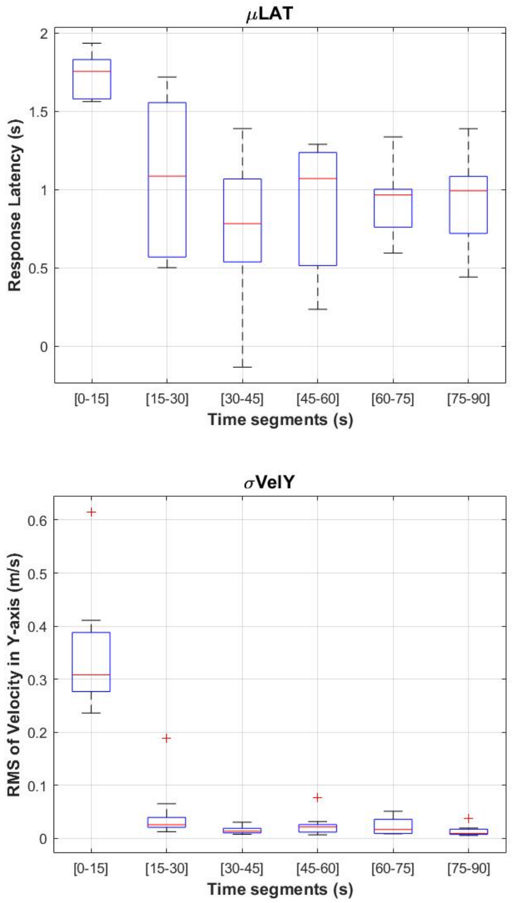


Figure 5: Inter-quartile range of score and the other engagement-related indicators. Only the feature with significant differences (i.e., p < 0:05) and the largest effect size is displayed per indicator (See Table1).

All three measures of exercise intensity show a significant positive correlation between them, but only $VelY$ and $VelZ$ showed a strong association $(p = 0.00, \tau = -0.52)$, and none of them seem to have a relationship with the behavioral measures $(p = 0{:}00, |\tau| < 0.01)$.

3.3 Psychophysiological measures

In contrast to behavioral engagement indicators, only one psychophysiological signal changed significantly in time, even with weak effect size $(p = 0.01, \eta_p^2 = 0.26)$, as presented in Fig.6. None of the features extracted from GSR and RSP were found to be significantly different in time windows, but only two features from HR with moderate effect sizes: μHR $(p = 0.05, \eta_p^2 = 0.32)$ and $medHR$ $(p = 0.05, \eta_p^2 = 0.41)$. In the case of μHR, post hoc tests found significant differences between windows in the following pairwise comparisons: second and fourth ($p = 0.02$), second and fifth ($p = 0.01$), and third and fourth ($p = 0.02$). Such differences are displayed in Fig.7, where a possible similarity between windows 4, 5, and 6 can be observed, which suggests a mild growing damped trend.

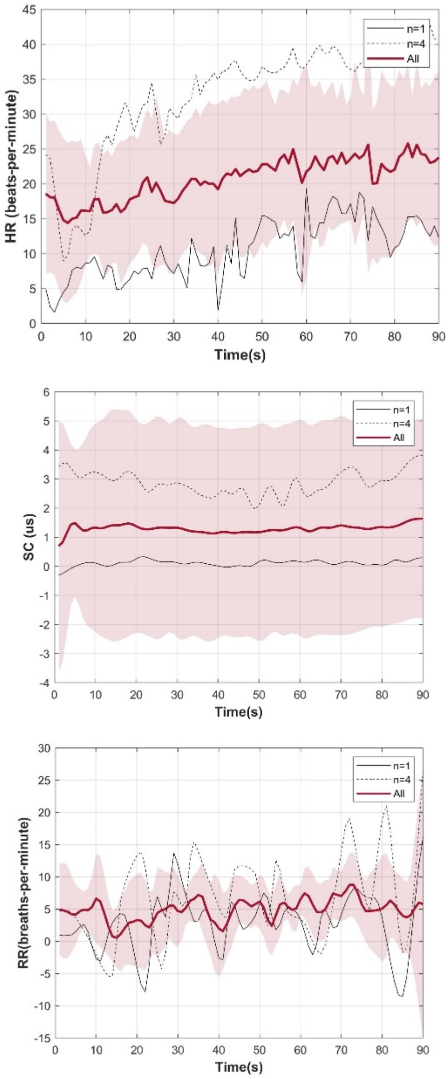


Figure 6: Average psychophysiological signals during the VR exposure: Heart Rate (HR), Skin Conductance (SC), and Respiratory Rate (RR). Signals from 2 random subjects were also included to illustrate single responses and normalized to the respective mean at rest [n = 1: μHR = 87.30 bpm, μSCL = 3.57 μS, μRR = 18.26 bpm], [n = 4: μHR$ = 89.16 bpm, μSCL = 3.85 μS, μRR = 15.14 bpm].

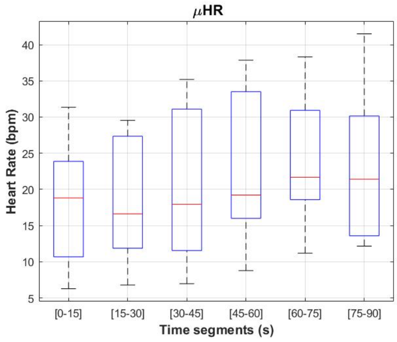


Figure 7: Inter-quartile range from the mean of Heart Rate (μHR). Data were normalized subjects' baseline (at rest).

Respiratory Rate was found to be positively correlated with score ($p = 0.00, \tau = 0.22$) and negatively correlated with Latency ($p = 0.00, \tau = -0.21$), which could suggest it to be a response from DDA. Likewise, μHR, was significantly correlated with $\mu Score$ ($p = 0.03, \tau = 0.27$), but the size of such relationship is too weak to be considered without caution, as well as with the correlation found between $\mu Score$ and LF/LH ($p = 0.00, \tau = 0.34$), and between $\mu Score$ and $nSCR$ ($p = 0.04, \tau = 0.23$).

Finally, Skin Conductance was found to be positively correlated with $VelX$ ($p = 0.00, \tau = 0.27$), in particular between $muVelX$ and $nSCR$ ($p = 0.04, \tau = 0.30$), whereas Heart Rate was found to be negatively correlated with $VelY$ ($p = 0.00, \tau = -0.20$).

4 Discussion

This study investigated the effects of dynamic difficulty adjustment (DDA) on behavioral and emotional engagement during immersive VR tasks. It was hypothesized that some common indicators of perceptive engagement, motor engagement, and emotional engagement behaved similarly to in-game performance (Score) during DDA.

As expected, DDA caused each participant's score to grow until they reached their apparent "skill limit," It started to oscillate to support the findings presented by Rodriguez-Guerrero [Rodriguez-Guerrero 12]. However, due to the high variability of responses (some subjects reached this level faster than others and differed in oscillations amplitude), this trend was not explicit in the chart with the mean of Score (Fig.4-up), but only observable when inspecting data dispersion or by the analysis of significance in features between time windows, such as in Fig.5.

Response Latency and Exercise Intensity are even less evident but describe a similar trend in time (negative in the first and positive for the last). The selected feature from each signal to be displayed in the boxplots ($muLAT$ and $muVelY$, seen in Fig.5) gave better evidence of this damped behavior, and both showed moderate-strong correlation with $\mu Score$. Findings confirmed that only $VelY$ changed significantly, consistent with the game dynamics that required the participant to perform repetitive vertical motions (moving the mallet to whack the moles), whereas $VelX$ and $VelZ$

remained steady almost from the beginning of the task. Given this and that both are subjective reactions to performance feedback during difficulty modulation (i.e., TOH dynamical adjustment), it is not surprising that findings confirmed the strong negative correlation between them.

In contrast to earlier findings, however, evidence of psychophysiological signals was rather disappointing. Only HR is likely to respond similarly to Score in DDA compared with the other measured signals, which can be suggested from the damped growing trend displayed by μHR in Fig6. These findings match those observed in earlier studies that said cardiac activity increases during stressful conditions, which can be noticed from HR growing with difficulty during DDA. Literature also suggests activation of the sympathetic system as higher perspiration (sweating) and faster respiration during challenging tasks as possible consequences of excitement and muscle tightness during higher physical effort [Goshvarpour 17, Leung 14]. However, the weak strength of the relationship of RR and Score and the lack of significant changes of SC in time suggest that they are not substantial representations of such effect.

Skin conductance seemed to start a growing trend in the last segment, possibly explained by the subject's cognitive or physical fatigue. However, this outcome is not conclusive and cannot be extrapolated to all VR users, and it will require longer sessions and more extensive study samples to be studied appropriately and confirmed as entirely out of the influence of other factors.

4.1 Study limitations and future work

The main limitation of the current study was the small sample size (n=10), mainly because of the pandemic-related lack of volunteers and risk of learning bias of performing repeated measures with the current subjects, which made it difficult to assure normal distribution and disallowed using parametric statistics. Therefore, some trends that were identified in measures signals had to be analyzed with caution to avoid confounding and underestimations. Likewise, some of the mentioned trends were observed in the last seconds of the exercise, suggesting the need for a more prolonged experiment to allow all subjects to reach and preserve their own "skill limit" and describe more significant changes in terms of the analyzed features. This shortening in time was also an issue during baseline measures since some participants take longer to reach the desired steady level, almost at the end of the fixed measuring time window.

Furthermore, some future studies should include a covariant factor related to experience in immersive VR, given the high variability in the time required to reach "skill limit" and its relationship with the ease to manipulate such devices.

Finally, it is also noticeable how the results from questionnaires cannot be conclusive due to the small study sample. Despite being effective, VRSQ could be reconsidered for more extensive studies to include nausea-related items. The subjective measurement of engagement (UES-S) helped prove levels among participants, but the study design (i.e., the single after-exercise measure) did not allow objective comparisons with the other sampled indicators. Further studies should include multiple within-subject trials with different experimental conditions, as is commonly seen in experimental designs that use this and other self-reports.

5 Conclusion

In this study, we have measured several engagement-related signals to evaluate their response during dynamic difficulty adjustment (DDA) in a 90-seconds-long immersive virtual reality game that involved repetitive upper-limb motions and was overall perceived as highly engaging (UES-SF tested). DDA was performed via an IDOL algorithm that increased or decreased game difficulty after any score variation, which produced that the game score described a damped growing trend towards a subject's maximum value according to their skill limit. Response Latency and hand velocity were the measures of behavioral engagement, whereas Heart Rate, Skin Conductance, and Respiratory Rate indicated emotional engagement.

Findings suggest that Response Latency (time between request and action) and hand velocity in the Y-axis (vertical motions of the hand) showed significant evidence of changes in response to DDA. Moreover, Latency showed strong negative correlation with the score, and notably, its mean and max values were strongly correlated with $muScore$. Likewise, the RMS of $VelY$ changed significantly and strongly related to the time-variant experimental conditions, i.e., difficulty adjustment, and its max value at each time segment showed a positive correlation with the mean of the score. However, the weak effect size in other features extracted from the RMS of $VelY$ suggests the need for further studies with more extensive study samples.

Heart Rate (HR) showed promising results through significant changes in time due to the difficulty adjustment and described a growing damped curve similar to the score. Despite showing no evidence of changes in time, the high variability of skin conductance also suggests the need for further studies with more extensive samples since it seems to describe a similar growing behavior that must be taken with caution.

Ethics Statement

The study protocol was followed in compliance with relevant laws and guidelines to be approved by the UMNG's Ethics committee. Authors ensure that informed consent was obtained from all volunteers after clarifying the purpose of the study and their right to withdraw.

Acknowledgments

This work was financially supported by Vicerrectoria de Investigaciones in Universidad Militar Nueva Granada (UMNG) under Grant ING-IMP-3124 (2020-2021). We would like to acknowledge the volunteers who participated in this study and the researchers of UMNG and VUB for their technical support and advisory.

References

[Barreda, 20] Barreda-Ángeles, M., Aleix-Guillaume, S., & Pereda-Baños, A. (2020). Users' psychophysiological, vocal, and self-reported responses to the apparent attitude of a virtual audience in stereoscopic 360°-video. *Virtual Reality*, *24*(2), 289–302. doi:10.1007/s10055-019-00400-1

[Brockmyer, 09] Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., & Pidruzny, J. N. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, *45*(4), 624–634. doi:10.1016/j.jesp.2009.02.016

[Caldas et al., 20] Caldas, O. I., Aviles, O. F., & Rodriguez-Guerrero, C. (2020). Effects of Presence and Challenge Variations on Emotional Engagement in Immersive Virtual Environments. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(05), 1109–1116. doi:10.1109/TNSRE.2020.2985308

[Caldas, 20] Caldas, O. I., Aviles, O. F., & Rodriguez-Guerrero, C. (2020). A simplified method for online extraction of skin conductance features: A pilot study on an immersive virtual-reality-based motor task. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *2020-July*, 3747–3750. doi:10.1109/EMBC44109.2020.9176424

[Cheng, 14] Cheng, L. K., Chieng, M. H., & Chieng, W. H. (2014). Measuring virtual experience in a three-dimensional virtual reality interactive simulator environment: A structural equation modeling approach. *Virtual Reality*, *18*(3), 173–188. doi:10.1007/s10055-014-0244-2

[Darzi, 19] Darzi, A., Wondra, T., McCrea, S., & Novak, D. (2019). Classification of multiple psychological dimensions in computer game players using physiology, performance, and personality characteristics. *Frontiers in Neuroscience*, *13*(November), 1–13. doi:10.3389/fnins.2019.01278

[Gilbert, 16] Gilbert, S. B. (2016). Perceived Realism of Virtual Environments Depends on Authenticity. *Presence*, *25*(4), 322–324. doi:10.1162/PRES_a_00276

[Gorsic et al., 17] Goršič, M., Cikajlo, I., Goljar, N., & Novak, D. (2017). A multisession evaluation of an adaptive competitive arm rehabilitation game. *Journal of NeuroEngineering and Rehabilitation*, *14*(1), 1–15. doi:10.1186/s12984-017-0336-9

[Gorsic, 17] Goršič, M., Cikajlo, I., & Novak, D. (2017). Competitive and cooperative arm rehabilitation games played by a patient and unimpaired person: effects on motivation and exercise intensity. *Journal of NeuroEngineering and Rehabilitation*, *14*(1), 1–18. doi:10.1186/s12984-017-0231-4

[Goshvarpour, 17] Goshvarpour, A., Abbasi, A., & Goshvarpour, A. (2017). An accurate emotion recognition system using ECG and GSR signals and matching pursuit method. *Biomedical Journal*, *40*(6), 355–368. doi:10.1016/j.bj.2017.11.001

[Hamari, 14] Hamari, J., & Koivisto, J. (2014). Measuring flow in gamification: Dispositional Flow Scale-2. *Computers in Human Behavior*, *40*, 133–143. doi:10.1016/j.chb.2014.07.048

[Hoffman, 09] Hoffman, D. L., & Novak, T. P. (2009). Flow Online: Lessons Learned and Future Prospects. *Journal of Interactive Marketing*, *23*(1), 23–34. doi:10.1016/j.intmar.2008.10.003

[Karpouzis, 20] Karpouzis, K., Liarokapis, F., Stoffregen, T. A., Grassini, S., Laumann, K., & Skogstad, M. R. (2020). The Use of Virtual Reality Alone Does Not Promote Training Performance (but Sense of Presence Does). *Article 1743. Psychol*, *11*(1743), 1–11. doi:10.3389/fpsyg.2020.01743

[Kerous, 20] Kerous, B., Barteček, R., Roman, R., Sojka, P., Bečev, O., & Liarokapis, F. (2020). Examination of electrodermal and cardio-vascular reactivity in virtual reality through a combined stress induction protocol. *Journal of Ambient Intelligence and Humanized Computing*, *11*(12), 6033–6042. doi:10.1007/s12652-020-01858-7

[Kim, 18] Kim, H. K., Park, J., Choi, Y., & Choe, M. (2018). Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Applied Ergonomics*, *69*, 66–73. doi:10.1016/j.apergo.2017.12.016

[Knaepen, 15] Knaepen, K., Marusic, U., Crea, S., Rodríguez Guerrero, C. D., Vitiello, N., Pattyn, N., … Meeusen, R. (2015). Psychophysiological response to cognitive workload during symmetrical, asymmetrical and dual-task walking. *Human Movement Science*, *40*(1), 248–263. doi:10.1016/j.humov.2015.01.001

[Lee, 15] Lee, H. Y., Kim, L., & Lee, S. M. (2015). Effects of virtual reality-based training and task-oriented training on balance performance in stroke patients. *J. Phys Ther. Sci*, *27*(6), 1883–1888. doi:10.1589/jpts.27.1883

[Lequerica, 10] Lequerica, A. H., & Kortte, K. (2010). Therapeutic Engagement. *American Journal of Physical Medicine & Rehabilitation*, *89*(5), 415–422. doi:10.1097/PHM.0b013e3181d8ceb2

[Leung, 14] Leung, B., & Chau, T. (2014). Autonomic responses to correct outcomes and interaction errors during single-switch scanning among children with severe spastic quadriplegic cerebral palsy. *Journal of NeuroEngineering and Rehabilitation*, *11*(1), 1–13. doi:10.1186/1743-0003-11-34

[Li, 16] Li, C., Rusák, Z., Horváth, I., & Ji, L. (2016). Development of engagement evaluation method and learning mechanism in an engagement enhancing rehabilitation system. *Engineering Applications of Artificial Intelligence*, *51*, 182–190. doi:10.1016/j.engappai.2016.01.021

[Makransky, 17] Makransky, G., Lilleholt, L., & Aaby, A. (2017). Development and validation of the Multimodal Presence Scale for virtual reality environments approach. *Computers in Human Behavior*, *72*(1), 276–285. doi:10.1016/j.chb.2017.02.066

[Makransky, 19] Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, *60*, 225–236. doi:10.1016/j.learninstruc.2017.12.007

[Csikszentmihalyi, 98] Csikszentmihalyi, M. (1998). *Finding Flow: The Psychology ofEngagement with Everyday Life* (2nd ed.). New York, USA: Basic Books.

[O'Brien, 16] O'Brien, H., & Cairns, P. (2016). *Why engagement matters: Cross-disciplinary perspectives of user engagement in digital media*. *Why Engagement Matters: Cross-Disciplinary Perspectives of User Engagement in Digital Media*. Springer International Publishing. doi:10.1007/978-3-319-27446-1

[O'Brien, 18] O'Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human Computer Studies*, *112*(December 2017), 28–39. doi:10.1016/j.ijhcs.2018.01.004

[Ozkul, 19] Ozkul, F., Palaska, Y., Masazade, E., & Erol-Barkana, D. (2019). Exploring dynamic difficulty adjustment mechanism for rehabilitation tasks using physiological measures and subjective ratings. *IET Signal Processing*, *13*(3), 378–386. doi:10.1049/iet-spr.2018.5241

[Pan, 85] Pan, J., & Tompkins, W. J. (1985). Pan Tomkins 1985 - QRS detection.pdf. *IEEE Transactions on Biomedical Engineering*, *32*(3), 230–236. doi:10.1109/TBME.1985.325532

[Pinto, 18] Pinto, J. F., Carvalho, H. R., Chambel, G. R. R., Ramiro, J., & Gonçalves, A. (2018). Adaptive gameplay and difficulty adjustment in a gamified upper-limb rehabilitation. *2018 IEEE 6th International Conference on Serious Games and Applications for Health, SeGAH 2018*, 1–8. doi:10.1109/SeGAH.2018.8401363

[Rodriguez-Guerrero, 12] Rodriguez-Guerrero, C. (2012). *Psychophysiological Feedback Control in Physical Human-Robot Interaction [dissertation]*. University of Valladolid.

[Rodriguez-Guerrero, 17] Rodriguez-Guerrero, C., Knaepen, K., Fraile-Marinero, J. C., Perez-Turiel, J., Gonzalez-de-Garibay, V., & Lefeber, D. (2017). Improving challenge/skill ratio in a multimodal interface by simultaneously adapting game difficulty and haptic assistance through psychophysiological and performance feedback. *Frontiers in Neuroscience*, *11*, 242. doi:10.3389/fnins.2017.00242

[Sailer, 17] Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, *69*, 371–380. doi:10.1016/j.chb.2016.12.033

[Skarbez, 17] Skarbez, R., Brooks, F. P., & Whitton, M. C. (2017). A survey of presence and related concepts. *ACM Computing Surveys*, *50*(6), 1–39. doi:10.1145/3134301

[Skarbez et al., 17] Skarbez, R., Neyret, S., Brooks, F. P., Slater, M., & Whitton, M. C. (2017). A psychophysical experiment regarding components of the plausibility illusion. *IEEE Transactions on Visualization and Computer Graphics*, *23*(4), 1322–1331. doi:10.1109/TVCG.2017.2657158

[Slater, 09] Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Phil. Trans. R. Soc. B*, *364*, 3549–3557. doi:10.1098/rstb.2009.0138

[Tsurumi, 02] Tsurumi, K., Itani, T., Tachi, N., Takanishi, T., Suzumura, H., & Takeyama, H. (2002). Estimation of energy expenditure during sedentary work with upper limb movement. *Journal of Occupational Health*, *44*(6), 408–413. doi:10.1539/joh.44.408

[Turkay, 15] Turkay, S., & Adinolf, S. (2015). The effects of customization on motivation in an extended study with a massively multiplayer online roleplaying game. *Cyberpsychology*, *9*(3). doi:10.5817/CP2015-3-2

[van Baren, 04] van Baren, J., & IJsselsteijn, W. (2004). *Measuring Presence : A Guide to Current Measurement Approaches*. *Measurement* (Vol. 1).

[van der Pas, 11] Van Der Pas, S. C., Verbunt, J. A., Breukelaar, D. E., Van Woerden, R., & Seelen, H. A. (2011). Assessment of arm activity using triaxial accelerometry in patients with a stroke. *Archives of Physical Medicine and Rehabilitation*, *92*(9), 1437–1442. doi:10.1016/j.apmr.2011.02.021

[Witmer, 98] Witmer, B. G., & Singer, M. J. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence*, *7*(3), 225–240. doi:10.1162/105474698565686

[Yu, 19] Yu, R. (2019). *Designing Coherent Interactions for Virtual Reality [dissertation]*. Virginia Polytechnic Institute and State University.