Virtual Reality (2023) 27:1091–1115
https://doi.org/10.1007/s10055-022-00699-3

ORIGINAL ARTICLE

Investigating the effectiveness of immersive VR skill training and its link to physiological arousal

Unnikrishnan Radhakrishnan[1] · Francesco Chinello[1] · Konstantinos Koumaditis[1]

Received: 8 June 2022 / Accepted: 13 September 2022 / Published online: 14 November 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract

This paper details the motivations, design, and analysis of a study using a fine motor skill training task in both VR and physical conditions. The objective of this between-subjects study was to (a) investigate the effectiveness of immersive virtual reality for training participants in the 'buzz-wire' fine motor skill task compared to physical training and (b) investigate the link between participants' arousal with their improvements in task performance. Physiological arousal levels in the form of electro-dermal activity (EDA) and ECG (Electrocardiogram) data were collected from 87 participants, randomly distributed across the two conditions. Results indicated that VR training is as good as, or even slightly better than, training in physical training in improving task performance. Moreover, the participants in the VR condition reported an increase in self-efficacy and immersion, while marginally significant differences were observed in the presence and the temporal demand (retrieved from NASA-TLX measurements). Participants in the VR condition showed on average less arousal than those in the physical condition. Though correlation analyses between performance metrics and arousal levels did not depict any statistically significant results, a closer examination of EDA values revealed that participants with lower arousal levels during training, across conditions, demonstrated better improvements in performance than those with higher arousal. These findings demonstrate the effectiveness of VR in training and the potential of using arousal and training performance data for designing adaptive VR training systems. This paper also discusses implications for researchers who consider using biosensors and VR for motor skill experiments.

Keywords Immersive virtual reality · Skill training · Physiological arousal · Electro-dermal activity · Heart rate variability

1 Introduction

Virtual reality (VR)-based training is increasing in popularity and is being explored in recent years across domains like education (Radianti et al. 2020), rehabilitation (Howard 2017), and various industries targeting adult learners (Abich et al. 2021; Radhakrishnan et al. 2021b; Renganayagalu et al. 2021; Xie et al. 2021). VR-based skill training brings in several advantages like allowing learners to practice procedures safely and repeatedly with consistent feedback (Hamilton et al. 2021). For example, in a Cochrane meta-analysis of studies investigating the effectiveness of VR training in endoscopy skills, it was found that VR training was more effective than no training and as effective as physical training (Khan et al. 2019). The advantages of VR training are being further enhanced by the increasingly widespread availability of immersive VR (IVR) technologies which make use of CAVE (Cave Automatic Virtual Environment) technologies or head-mounted displays (HMDs), offering high-fidelity audiovisuals to the user (Makransky et al. 2019). The immersion and presence offered by IVR further enhance its effectiveness, particularly when the affordances of IVR are matched with the teaching/training method (Makransky and Petersen 2021). It must be noted that IVR still has limitations in comparison to physical reality, to name a few in particular: differences in visual acuity, field of view, and the presence of cybersickness, the latter possibly linked to differences in vestibular response (Ashiri et al. 2020). As the evidence for the effectiveness of IVR over other methods is mixed (Abich et al. 2021; Radhakrishnan et al. 2021b), one may ask: how can IVR training be improved?

✉ Unnikrishnan Radhakrishnan
unnik@btech.au.dk

Francesco Chinello
chinello@btech.au.dk

Konstantinos Koumaditis
kkoumaditis@btech.au.dk

[1] Department of Business Development and Technology, Aarhus University, Birk Centerpark 15, 7400 Herning, Denmark

IVR training primarily makes use of easily observable training/test performance metrics like task completion time and the number of errors (Abich et al. 2021; Radhakrishnan et al. 2021b). In addition to such objective measures, the literature on skill training outside of IVR has also investigated the links between arousal and performance (Storbeck and Clore 2008; Yerkes and Dodson 1908). The term *arousal* refers to many related phenomena like an increase in alertness, attention, emotion, or the ability to respond to stimuli through motor movements (Calderon et al. 2016). Arousal levels are measured using both subjective (questionnaires) and objective methods (sensors). Existing biosensing technologies can measure pupil dilation, heart rate, electro-dermal activity, brain activity, skin temperature, respiration rate, and other measures of the body's autonomic arousal. IVR literature provides several examples where arousal levels are incorporated into studies on social anxiety (Owens and Beidel 2015), treatment of phobias (Diemer et al. 2016), presence (Terkildsen and Makransky 2019), and other studies of emotions and behavior (Marín-Morales et al. 2018; Syrjämäki et al. 2020). However, there are only a few instances in immersive and non-immersive VR training literature where arousal levels are measured and then linked to performance (Parong and Mayer 2021; Wu et al. 2010). Such research would open up new avenues for advancing the state of the art, particularly aided by the increasing availability of cost-effective biosensors that can measure physiological arousal and their integration with commercial IVR technologies (e.g., HP Reverb, OpenBCI Galea). If such links can be established, IVR training itself may be further enhanced with adaptation (Zahabi and Abdul Razak 2020) by changing the parameters of the training environment to increase or decrease the trainee's arousal levels and performance.

This paper adds to the body of the literature on motor skill training in IVR with a between-subjects fine motor skill training experiment. With the aid of $N=87$ participants, we compared the effectiveness of IVR against physical training conditions with a focus on performance and arousal. The latter is achieved with the use of wearable biosensors which measure physiological arousal in the form of electro-dermal activity (EDA) and electrocardiogram (ECG) signals. These were recorded from all participants across the two conditions. Furthermore, the study investigated improvements in performance after training along with subjective measures of immersion, presence, enjoyment, self-efficacy, and task load.

2 Related works

2.1 Training in virtual reality

Virtual reality has been described as a collection of technologies that creates synthetic and interactive three-dimensional environments (Mikropoulos and Natsis 2011). These technologies range from highly immersive ones like head-mounted displays (HMDs) and CAVEs to devices providing a comparatively lower level of immersion like desktops and smartphone displays. Technological advances have resulted in HMDs becoming more popular in recent years, which in turn increased interest in their applications in education and training (Checa and Bustillo 2020; Makransky and Petersen 2021). However, research suggests that IVR training should not be just implemented as a one-size-fits-all solution, but instead works best when the design factors of the training environment complement the capabilities provided by the IVR hardware (Jensen and Konradsen 2018).

Learning/training in immersive virtual environments extends across many domains like school/university education, rehabilitation training for patients, professional training for doctors, and office/industrial workers, where it focuses on diverse kinds of cognitive, affective, and motor skills (Jensen and Konradsen 2018). For this study, we limit the discussion of training literature focusing on teaching various cognitive and motor skills to healthy individuals. The literature on cognitive skills taught in IVR primarily relates to school and college education (Hamilton et al. 2021), as well as teaching procedural and safety knowledge primarily for industrial training purposes (Feng et al. 2018; Patle et al. 2019). On the other hand, motor skill training literature in IVR has been dominated by medical use cases, particularly in the surgical and dental domains which require fine motor skills (Radhakrishnan et al. 2021b). IVR-based motor skill training researchers have investigated the relative advantages. IVR-based training has over other training media (physical training, video training, etc.) or variations within IVR, like different levels of visual/haptic fidelity (Huber et al. 2018; Jain et al. 2020), participant characteristics (Shakur et al. 2015), and training methods (Harvey et al. 2019). The results of these studies have been varied; for example, Pulijala et al. (2018) found IVR to be more effective than video/presentation training, Hooper et al. showed IVR to be more effective than physical training for hip arthroplasty surgery, Butt et al. observed the same advantage of IVR over physical training for catheter insertion training, but the advantage disappeared after a week (Butt et al. 2018). Huber et al. found IVR to be as effective as an 'augmented' VR condition (Huber et al. 2018). In a comparison of IVR to desktop VR training, Frederiksen et al. found that IVR was inferior in its effectiveness and caused more cognitive load among students of laparoscopic surgery (Frederiksen et al. 2020). Thus, whether IVR training can be as effective or more effective compared to other types of training is inconclusive so far and an open research topic (Checa and Bustillo 2020) and more so in the case of IVR-based motor skill training (Coban et al. 2022). This need inspired the first research question addressed in this work: *RQ 1—Is IVR training as effective as physical training in improving task performance?*

In order to answer this research question, it is important to include observable measures signifying training effectiveness (Magill and Anderson 2016); for example, performance metrics like time for task completion, and quality metrics like the number of mistakes/errors (Abich et al. 2021; Radhakrishnan et al. 2021b; Wulf et al. 2010). While measuring such performance metrics, trainees may be tested before and after training to measure their performance improvement (Magill and Anderson 2016, p. 269). When the tests are performed in a physical setting, they provide a measure of the transfer of skills from the virtual to the real environment, which has been argued in the literature to be crucial in establishing the effectiveness of IVR training (Jensen and Konradsen 2018; Levac et al. 2019).

Subjective measures have been linked to the effectiveness of learning/training in IVR environments in the Cognitive Affective Model of Immersive Learning (CAMIL) (Makransky and Petersen 2021). The CAMIL framework suggests that there are two affordances to learning in immersive VR, namely presence (arising from immersion) and agency (arising from interactivity) which affect six other factors, i.e., interest, motivation, self-efficacy, embodiment, cognitive load, and self-regulation, which in turn affect the effectiveness of IVR training. Popular subjective measures from IVR training literature include measures of cognitive load, like the NASA Task Load Index (NASA-TLX) (Hart and Staveland 1988), measures of immersion, like the Immersive Tendencies Questionnaire (ITQ), measures of presence, like the presence questionnaire (Witmer and Singer 1998), measures of usability, like the System Usability Scale (SUS) (Brooke 1996), measures of cyber/motion sickness, like the Simulator Sickness Questionnaire (SSQ) (Kennedy et al. 1993), and measures of self-efficacy (Lehikko 2021; Pintrich 1991). It should be noted that while 'Immersion' is an objective measure of how vivid the VR technology can be made (for example, IVR is more immersive than desktop VR), 'Presence' is understood to be a subjective measure of experience by users which arises from both immersion and interactivity in VR (Makransky and Petersen 2021). Cognitive load is also crucial to understanding the effectiveness of VR in comparison to other media, as it is negatively correlated with learning/training effectiveness (Koumaditis et al. 2020; Van Merriënboer and Sweller 2010). Another subjective measure of importance is 'self-efficacy,' defined as the subjective belief people have about their own ability to fulfill a task (Bandura 1986). Self-efficacy measures are gaining more attention in the literature, as it has been positively linked to the IVR modality and learning outcomes (Shu et al. 2019; Tai et al. 2022). Therefore, it is important to measure the subjective perception of trainees in different training modalities in order to investigate their relationship with training effectiveness. This need generates the second research question: *RQ 2—Is there a significant difference in the enjoyment, presence, immersion, task load, and changes in self-efficacy reported by participants in IVR compared to physical training?*

IVR training is used in various contexts of motor skills. These can be broadly categorized as context-specific or context-independent. Many examples of context-specific IVR training are found in the medical and surgical domains, where the procedure being trained can easily be used for the same procedure in the real world but rarely in other contexts. An example from the non-medical domain is Winther et al. (2020) who explored the effectiveness of IVR-based training vs conventional training for a pump maintenance task. Such context-specific explorations result in findings that can be applied in the real world easily but are limited by their limited external validity, i.e., they are hard to generalize to other contexts. An advantage of studies on employing context-independent scenarios is therefore that the result is often easier to generalize and transfer to related domains. Examples exist in the IVR motor skill training literature that use more context-independent scenarios like puzzle assembly (Carlson et al. 2015; Koumaditis et al. 2020; Murcia-Lopez and Steed 2018). Though such examples are not related to real-world tasks or scenarios, it can be argued that such studies and skill training scenarios may generate results that are more generalizable and transferable to related domains. Inspiration can be found in laparoscopy surgical training literature, where the use of box trainers is widespread, which are highly simplified representations of the tasks involved in laparoscopy (Aggarwal et al. 2004). In this paper, we identify a fine motor skill task (buzz-wire or wire loop game) inspired by the literature where it was previously investigated in ergonomics research (Shafti et al. 2016) and in the domain of motor control (Luvizutto et al. 2022; Read et al. 2013) and rehabilitation (Budini et al. 2014; Christou et al. 2018). In this task, the aim is to move a metallic loop across a wire without entering into contact. Immediate feedback is provided when a mistake is made in the form of a loud 'buzz' and, in some cases, a blinking red light in the background. The wire is bent at different locations which makes the task challenging to perform while maintaining a steady hand (Shafti et al. 2016). Read et al. (2013) found that a buzz-wire setup was effective in assessing the relation between manual dexterity and binocular vision. Budini et al. (2014) used buzz-wire training along with hand postural exercises for patients with hand tremors in their experiment and found improvements in goal-directed tasks. Christou et al. (2018)

present the only example of research using the buzz-wire setup in an IVR environment, designed as an exercise tool for patients who have suffered stroke and other brain trauma. Similar to Read et al. (2013), they found that the presence of binocular viewing is correlated with increased performance and also that they could distinguish between dominant and non-dominant hand performance. Furthermore, the details provided by Christou et al. (2018) on designing increasing levels of buzz-wire task complexity inspired the current study.

2.2 Arousal and learning

Though the terms 'arousal' and 'emotion' have been used interchangeably in the literature, arousal is one aspect of emotion, along with valence (ranging from negative to positive) according to dimensional models of emotion (Posner et al. 2005; Rubin and Talarico 2009). Similarly, the terms 'stress' and 'anxiety' have also been used to denote high arousal states with a negative valence (Janelle 2002; Pakarinen et al. 2019). Multiple methods have been used/utilized to measure arousal levels, using both subjective (Bradley and Lang 1994) and objective methods (Cacioppo et al. 2007). Among subjective techniques, subjects report their degree of arousal using instruments like the Self-Assessment Manikin (SAM) (Bradley and Lang 1994) and the Stress Arousal Checklist (Mackay et al. 1978). Such questionnaires are usually measured post-exposure and depend on the user's knowledge of their own arousal levels, their memory of the task, and comprehension of the questions. On the other hand, objective measures of arousal are a function of the body's autonomic nervous system, which produces measurable responses, reflecting the user's emotional and cognitive state. This includes changes in skin conductivity (electrodermal/EDA activity due to sweating), heart rate parameters (heart rate variability/HRV), respiration, skin temperature, pupil dilation, and brain activity (Cacioppo et al. 2007). These biosignals can be measured by sensors placed on the body (usually non-invasive) to provide measures of physiological arousal. Objective biosignal data also allow for a more fine-grained look at variations in the subject's arousal levels during a study using measures like Event-Related potentials (ERPs) in EEG, Skin Conductance Responses (SCRs) in EDA, Inter-beat Intervals (or R–R intervals) in heart rate variability data, among many others, where each signal can be used in isolation or be coupled with others in order to increase accuracy (Cacioppo et al. 2007).

Arousal levels may have links to performance and learning outcomes, but limited empirical support is to be found. It has been hypothesized that an individual's experience of arousal affects attention, perception of time, and memory (Storbeck and Clore 2008), and that there is a non-linear 'inverted U-shaped' relationship between arousal levels and performance (Yerkes and Dodson 1908). However, the results have been inconclusive in validating this hypothesis (Storbeck and Clore 2008). Some examples from the literature point to a link between high arousal and better training performance (Homer et al. 2019; Matthews and Margetts 1991; Ünal et al. 2013). On the other hand, some explorations related to training have found that low arousal leads to better improvements in performance (Kuan et al. 2018; Pavlidis et al. 2019; Prabhu et al. 2010; Quick et al. 2017). The link between arousal and learning/training adds a further layer of complexity since the effectiveness of training is measured not by task performance alone but by changes in performance across different periods, usually as a change in performance before and after training (learning gain). Movahedi et al. (2007) illustrate this complexity in a sports training context where they found that participants performed worse during a retention test when their arousal levels during the test were mismatched with the arousal levels (either high or low) during training.

The use of physiological data to measure arousal levels in IVR literature is rare; however, some representative examples that use heart rate-related metrics for measuring arousal include Muñoz et al. (2019) where HRV metrics (along with EEG data) were used to detect calmness states among participants using an IVR target shooting simulator, Cebeci et al. (2019) where eye tracking and heart rate were used to measure the impact of different virtual environments on factors like cybersickness and emotions among study participants, and Larmuseau et al. (2020) where HRV along with EDA and skin temperature were used to measure cognitive load among students' learning statistics online. In the use of EDA data, some illustrative examples include understanding how soldiers respond to threatening stimuli during IVR training (Binsch et al. 2021), detecting student stress levels during a physics course (non-VR) (Pijeira-Díaz et al. 2018), and measuring EDA responses to insights made by participants in an IVR learning environment (Collins et al. 2019). There are currently only a few examples in IVR literature on the exploration of physiological arousal levels and their connection to fine motor skill training in virtual reality. One example is from a science education scenario where it was shown that learning in IVR leads to higher arousal and subsequently lower scores on a retention test (Parong and Mayer 2021). Another example is from non-immersive VR where a stroop interference task-induced arousal in participants during a virtual driving task and then found the optimal arousal levels related to increased performance (Wu et al. 2010). Therefore, a research gap exists in the literature for understanding the link between motor skill training in IVR, improvements in performance due to the training, and physiological arousal levels of the trainees. The following research questions were generated in order to address this gap: *RQ 3—Is there a significant difference between the physiological arousal levels of participants in IVR training compared to physical training? RQ 4—Is there a link between physiological arousal during training and improvements in performance after training?* In the next section, the design of the experiment is detailed which will help address these questions.

3 Methods

The experiment contains three phases as depicted in Figure 1: a pre-training phase common to all conditions where a pre-test of the motor skill is performed, a training phase in which the participants were randomly assigned to either VR or physical training conditions and a post-training phase where a post-test of the motor skill was performed for participants from both training conditions. The following sub-sections detail the motor skill task, the two experimental conditions, the pre-test and post-test tasks, the physiological and performance data measured during the experiment as well as the subjective data reported by the participants. The section ends with a detailed description of the experimental procedure shown in Fig. 1.

3.1 Motor skill task

In this study, the trainee is asked to grab the apparatus as shown in Fig. 2 and guide the metallic loop across a wire as fast as possible with the least amount of touching between the loop and the wire. There are two variations of the task, varying on the feedback provided when the loop touches

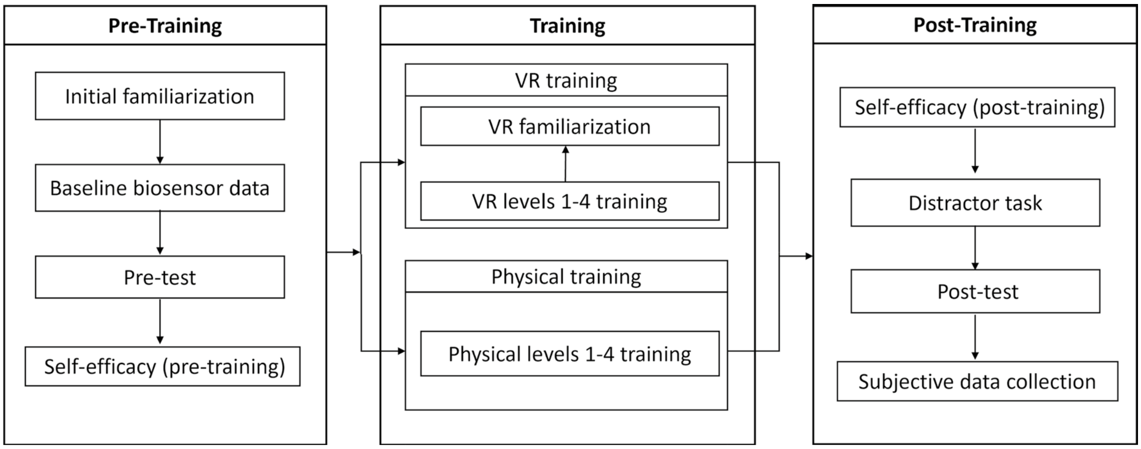


Fig. 1 Overview of experiment procedure

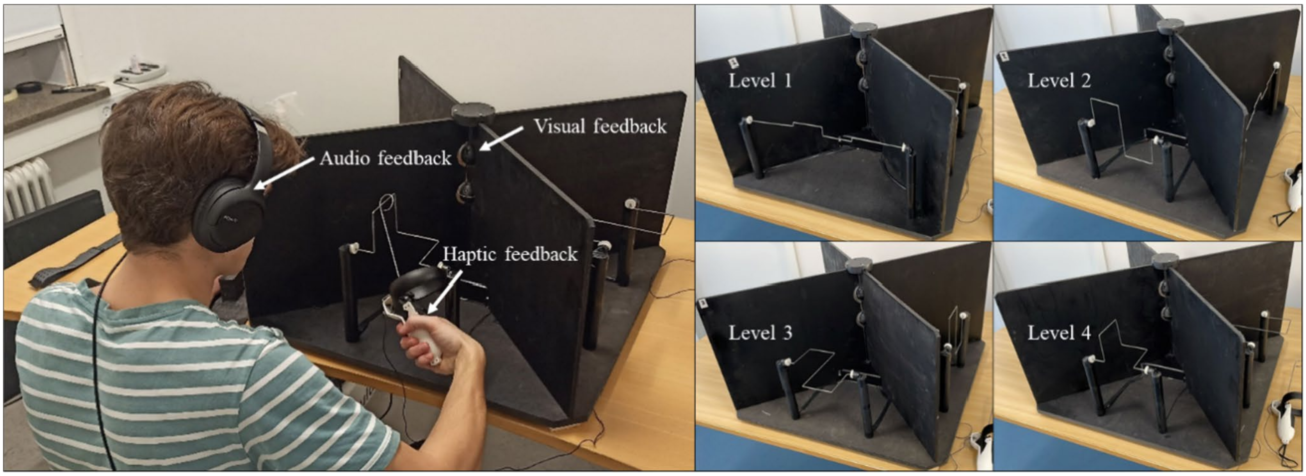


Fig. 2 Physical training condition. Left: participant moving loop across the wire in level 4. When the loop touches the wire, the participant receives audio, haptic, and visual feedback. Right: The four levels of training

the wire, i.e., when a mistake is made. In the training task, when the participant makes such a mistake, three kinds of feedback were provided simultaneously:

- *Haptic feedback* in the form of vibration in the Oculus Quest's Touch controller. Vibration is set to the maximum frequency and amplitude available in the Oculus SDK and delivered for 1/10th of a second.
- *Auditory feedback* was provided by playing a continuous 1000 Hz sine wave tone at 39 dB over the headphones worn by the participant (Sony WH-CH710N). Sound levels were verified and maintained across participants using the NIOSH iPhone app (National Institute for Occupational Safety and Health Sound Level Meter App).
- *Visual feedback* is provided by switching on a red LED (Fig. 2) placed at eye level behind the wire.

The training task in the physical and VR conditions is spread across four levels of increasing difficulty, with difficulty being specified as an increase in complexity of wrist movements needed to complete a level (see Table 1). For example, a wire with fewer bends requires less wrist movement, which in turn may produce fewer mistakes (i.e., the loop touching the wire) and the task may be completed (move from start to finish) quicker than a wire with more bends. This was verified in a previously published pilot study (Radhakrishnan et al. 2021a). These four levels were intended to help the participants train themselves, i.e., to develop the skills required to perform the test task more effectively. It should be noted that there were no instructions provided in either condition to facilitate the training by letting the participant construct their strategies for improving their skill level subject to the constraint of the environment.

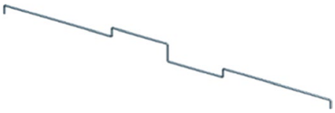
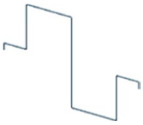
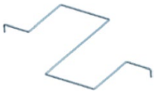
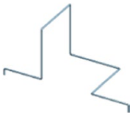
3.1.1 Training in physical condition

The wire in each training level rests on two 20-cm tall pillars to provide better task ergonomics for participants (verified in a pilot test). Two black vertical wooden panels are placed at right angles on the wooden base (Fig. 2), and the entire setup is painted black to reduce visual distractions. The start and end positions are shaped like cylinders with grooves inside for the loop to be placed. An Arduino Uno placed in a microcontroller box is used to detect contact between the loop and the wire (denoting mistakes) using a simple switch circuit. A 'mistake' signal is transmitted serially to the PC when the loop touches the wire. Two similar switch circuits are used to detect contact between the loop and the grooves on both the start and finish positions. When the participant lifts the loop off the start position, a 'start task' signal is transmitted by the contact circuit to the PC; similarly, an 'end task' signal is transmitted when the loop is placed in the end position. The loop is made by bending a 1-mm-thick metal wire with a diameter of 2.5 cm. The loop is then screwed to a 3D printed handle (adapted from Lagos (2019)) that houses an Oculus controller (Fig. 2) to provide haptic feedback.

3.1.2 Training in IVR condition

Participants in the VR condition wore an Oculus Quest (1st generation) head-mounted display (HMD) (Fig. 3) connected to a PC and running on Rift mode. The VR environment was developed in the Unity3D (version 2019.4) game engine to closely resemble the physical environment. The wires (for each training level) and loop were designed using the Blender3D design software. The participants were presented with

Table 1 Training levels

Difficulty level	Movement pattern	Level design
1	The first level (48 cm long) is almost straight across the x-axis with short deviations in the *y*-axis. The participant can complete the task with minimal twisting of the wrist	
2	The second level is 52 cm long and has bends in the *y*-axis. Participants may have to twist their wrists substantially compared to level 1	
3	The third level is 52 cm long (similar in proportions to level 2) with bends in the *z*-axis	
4	The last and most challenging level is 48 cm long with bends on all three axes	

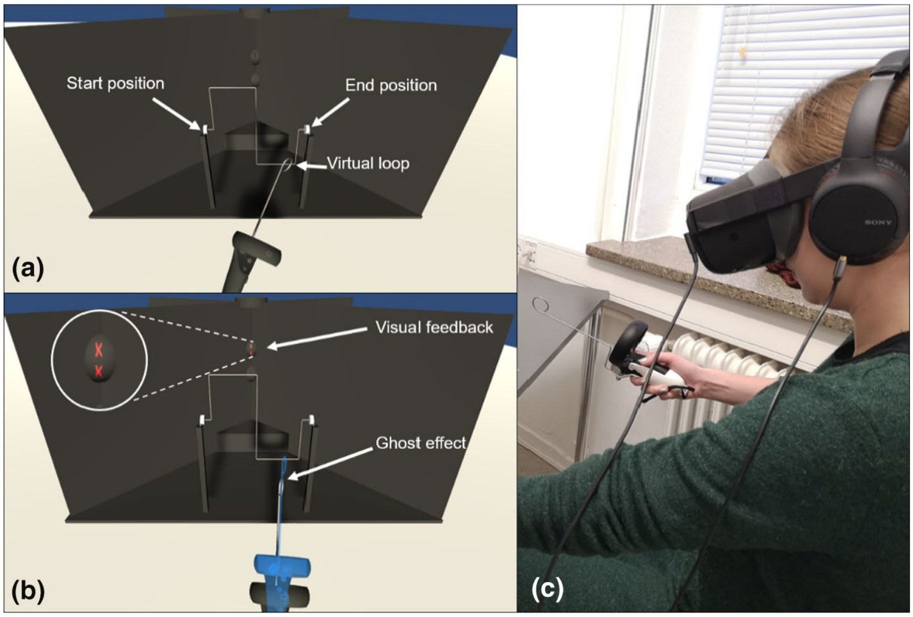


Fig. 3 **a** VR training environment. Virtual loop moving across level 2, **b** Ghost loop appears when contact is made, and the 'real' loop goes outside the wire. It disappears when the loop is placed back inside the wire. Visual feedback in form of a red 'X' mark in the background also turns on during contact. **c** Participant in VR condition wearing an Oculus Quest HMD (Rift mode)

the same four levels in VR as in physical condition. They hold a physical handle containing the loop and the Oculus controller (like those in the test and physical training tasks). The position and rotation of both the controller and the HMD are provided by the Oculus SDK which is then used to move the virtual loop and the participant's viewpoint in the three-dimensional space of the virtual environment (see Fig. 3).

The 'Measurements' asset from the Unity Asset Store was used to scale and position objects identically to their real-world counterparts (Vrchewal 2020). Both haptic and audio feedback modalities used the same parameters as the physical condition, and the visual feedback was in the form of a red 3D light behind each wire turning on during contact between the virtual wire and the virtual loop (Fig. 3b). Like the physical condition, 'start task,' 'end task,' and 'mistake' signals were sent to the data collection module (Fig. 5). Physics collision meshes were defined on the 3D models of the loop, the start and end positions, and the wires across the four levels. Collision tests were performed by Unity's inbuilt physics engine at 60 Hz.

Though the VR condition mimics the physical, there are unavoidable differences between the two conditions:

- *Ghost effect during mistakes* When the participant makes a mistake, i.e., the loop touches the wire, there is nothing to physically restrict the participant's hand, unlike the physical condition where there is an actual wire to provide resistance. Though there is haptic vibration when contact is made, by the time the mistake is made, the loop would have passed through the wire creating an unrealistic effect for the participant which could potentially break their feeling of immersion (i.e., 'being there'). To solve this, a 'ghost effect' has been programmed to show a blue translucent loop at the contact position where the actual loop passes through the virtual wire (Fig. 3b). This helps the participant understand how to bring their loop back into the wire, at which point the blue translucent 'ghost' disappears.
- *VR familiarization* Participants were first exposed to a VR task to help them familiarize themselves with the movement of the virtual loop before starting the actual training. This is to avoid any negative outcomes from the novelty effect of using IVR among novice users (Hamilton et al. 2021). They were encouraged to intentionally make mistakes to learn the functionality of the ghost effect. The task is in the form of a straight wire which has no bends so that there is no unintended extra 'training effect' for participants in the VR condition.
- *Differences in media* In addition to the above two features which distinguishes VR from the physical, there are other differences arising from the nature of the VR medium itself, for example—the field of view and the visual acuity provided by the Quest HMD are lower compared to that provided by healthy human vision (Adhanom et al. 2021; Cuervo et al. 2018). Additionally, the weight of the HMD has not been replicated in the physical condition.

3.1.3 Test task

The wire in the test task is 52 cm long with eleven 90° bends in all three axes (x, y, and z) between the start and finish

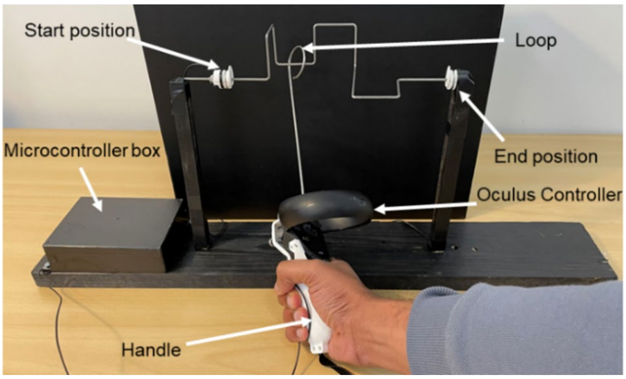


Fig. 4 Test task setup along with the loop attached to 3D printed handle containing a Quest controller

positions (see Fig. 4). Contact circuits like those used in the physical training setup are used here to detect contact between the loop and the wire as well as the corresponding start and end positions. The three contact signals 'start task,' 'mistake,' and 'end task' are serially transmitted to the PC similar to the training setup (see 3.1.2). There is no 'augmented' feedback provided when the participant makes a mistake in the test condition, i.e., there is no haptic, visual, or auditory feedback other than the natural feedback of two metal pieces touching each other. Like the training setup, all parts of the test setup are painted black to provide a consistent background with fewer visual distractions. An Oculus controller is placed inside the handle containing the loop to mimic the weight of the controller in the physical and VR training setups but provides no haptic feedback. The same test task is used before and after the VR/physical training task as an objective measure of training effectiveness.

3.2 Sensors and data collection

Data collected during the experiment come from three kinds of sources: the biosensors, the task-related signals coming from the test and training setups, and subjective data recorded in an online survey (at the end of the experiment). The first two types of data are facilitated by:

- *iMotions* iMotions is a commercial software platform that supports data collection from commercial biosensors across many modalities (iMotions A/S, Copenhagen, Denmark). In this study, iMotions was used as the endpoint for storing all data coming through the dataflow pipeline shown in Fig. 5, as it integrates timestamped data from the two biosensors alongside performance-related data coming from the data collection module.
- *Data collection module* A data collection module was developed in C# on the Unity3D game engine which collected task-related signals from the hardware setups (test and training) and the VR training software. Data from the hardware were read from two serial connections with a transmission rate of 9600 baud. The data collection module then transmitted in real-time the collected signals to the iMotions biosensor platform via a TCP socket connection (Fig. 5).

Subsequent sub-sections discuss the biosensors used for measuring electro-dermal and heart rate signals, associated arousal metrics (3.1.1), performance metrics for measuring the effectiveness of training (3.3.2), and survey data to measure the subjective experience of training using online questionnaires (3.3.3).

Fig. 5 Software architecture

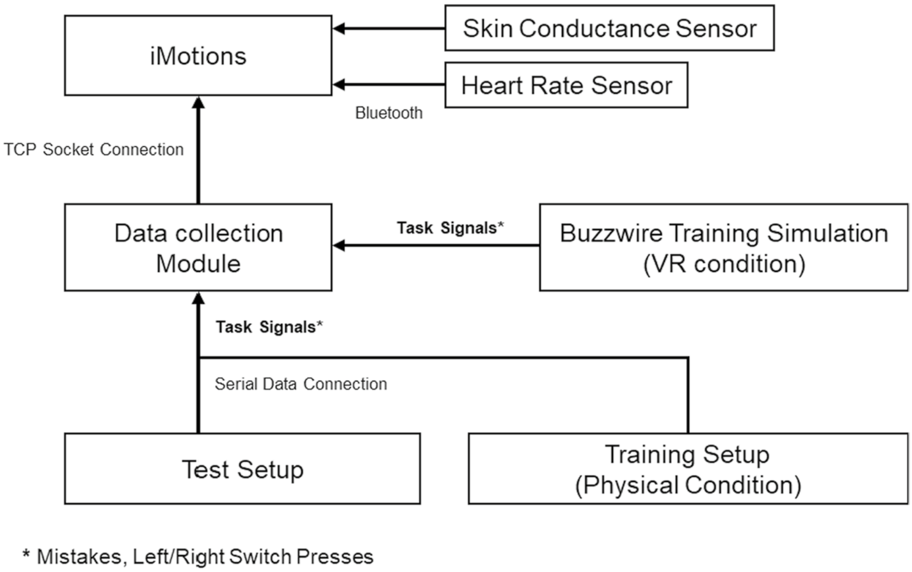


Table 2 Physiological metrics, their source, and their relation to changes in arousal

Physiological arousal metric	Signal source	Relation with arousal
Skin conductance (SC)	EDA	SC↑—Arousal ↑
Skin conductance response amplitude (SCRAmp)	EDA	SCRAmp↑—Arousal ↑
Skin conductance response peaks rate (SCRPeaks)	EDA	SCRPeaks↑—Arousal ↑
Heart rate (HR)	ECG	HR ↑—Arousal ↑
Inter-beat interval (IBI)	ECG	IBI ↓—Arousal ↑
Root mean square of successive difference (RMSSD)	ECG	RMSSD ↓—Arousal ↑
Standard deviation of NN intervals (SDNN)	ECG	SDNN ↓—Arousal ↑
Normalized high-frequency component (HFN)	ECG	HFN ↓—Arousal ↑
LF/HF (Low frequency/High frequency) ratio	ECG	LF/HF Ratio↑—Arousal ↑

3.2.1 Physiological sensing

For measuring the participant's physiological arousal levels, the Polar H10 (heart rate) and the Shimmer GSR + (skin conductance) sensors were used. Table 4 in the Appendix details all the physiological metrics used, their source, and their relationship with arousal according to the literature (Table 2).

3.2.1.1 Electrocardiogram (ECG) signals The Polar H10 (Polar Electro Oy, Kempele, Finland) is an electrocardiogram (ECG)-based heart rate (HR) monitor designed for athletes. It has been clinically validated to be as effective as medical-grade ECG hardware (Gilgen-Ammann et al. 2019) and has been used in recent VR literature (Muñoz et al. 2019; Ventura et al. 2021). It is worn around the chest with electrodes placed in contact with the skin. The data in the form of heart rate and Inter-beat Intervals (R–R intervals) are transmitted via a Bluetooth Low Energy (BLE) connection at a rate of 1–2 Hz to the iMotions application running on a PC. Measures of heart rate variability including time and frequency domain metrics have been calculated using the hrv-analysis Python library (Champseix 2021).

Increases in arousal are indicated by increases in heart rate (time-domain) and frequency-domain measures like LF/HF (Low Frequency/High Frequency) ratio (Orsila et al. 2008; Slater et al. 2006). On the other hand, decreases in time-domain HRV measures like IBI (Inter-Beat Interval), SDNN (Standard Deviation of NN Intervals), RMSSD (Root Mean Square of Successive Difference), and the frequency domain measure HFN (Normalized High-Frequency Component) indicate an increase in arousal (Shaffer and Ginsberg 2017). All HRV metrics have been baseline corrected by subtracting from them the corresponding mean baseline values (Healey and Picard 2005; Wulfert et al. 2005).

3.2.1.2 Electro-dermal activity (EDA) signals The Shimmer GSR + (Galvanic Skin Resistance) unit (Shimmer Research Ltd., Dublin, Ireland) measures EDA by passing a small current through electrodes placed in two locations on the body. The locations for the electrodes were verified in a pilot study where Shimmer electrodes were placed on the foot, the forehead, and the fingers of two participants, and the signals generated in response to stimuli were examined for signal quality and consistency. It was found that the index and middle fingers were the most reliable locations for sensing skin conductance which matched recommendations from the literature on skin conductance sensing (van Dooren and Janssen 2012). The index and middle fingers of the left hand were chosen to allow study participants to use their right hand alone for moving the loop across the wires.

Popular EDA measures include SC (Skin Conductance) measured in micro-siemens which increases in response to an increase in arousal (Collet et al. 2005). An increase in arousal also leads to a higher rate of skin conductance response peaks which are peaks in the SC amplitude lasting between 1 and 5 s after onset (Krogmeier et al. 2019; Terkildsen and Makransky 2019). The SCRPeaks measure is calculated as the number of skin conductance response peaks per minute. Similarly, the mean peak amplitude of all SCR peaks (SCRAmp) is also a positive measure of arousal (Khalfa et al. 2002; Krogmeier et al. 2019). SCL levels have been baseline corrected by subtracting from it the mean baseline values (Potter and Bolls 2012). All EDA signals were processed using the Neurokit2 Python library (Makowski et al. 2021).

3.2.2 Improvement in performance

The data collection module collects signals generated from both physical and IVR setups, namely the 'Start task,' 'End task,' and 'Mistake' signals. These are used to calculate the following two measures of performance:

- *Task completion time (TCT)* The time taken to move the loop from start to end.

- *Contact time (CT)* The total time the loop is in contact with the wire during the task which quantifies the number of mistakes by the participant.

These two measures are then used to calculate the following measures of performance improvement:

- *Improvement in task completion time (TCT-I)* This is calculated by subtracting the posttest TCT from the pre-test TCT for each participant. A positive value indicates an improvement in this performance metric.
- *Improvement in contact time (CT-I)* This is calculated by subtracting the posttest CT from the pre-test CT for each participant. A positive value indicates an improvement in this performance metric.
- *Improvement Score (IS)* Since the participants are asked to complete the test task by satisfying two potentially competing goals—to minimize both task completion time and contact time—participants may choose to prioritize one over the other. For example, a participant can choose to complete the task very slowly to minimize the chances of contact with the wire or vice versa. To balance out these two metrics, it is necessary to create a combined score metric that considers both improvements in task completion time (TCT-I) and contact time (CT-I). To calculate this measure, we first divide the two performance improvement measures, TCT-I and CT-I, into 10 equal-sized quartiles for all participants across both conditions, transforming the values into scores from 1 to 10 where 1 denotes the least improvement in performance and 10 the most. Subsequently, IS for a participant is defined as the sum of these two scores. A hypothetical participant who has improved the most in both TCT-I (score = 10) and CT-I (score = 10) metrics would then get a final improvement score (IS) of 20.

3.2.3 Subjective data

Subjective data were collected from all participants toward the end of the experiment using an online survey tool (Microsoft Forms) running on a laboratory PC. The different subjective metrics are listed below.

- *NASA Task Load Index (NASA-TLX)* NASA Task Load Index (Hart and Staveland 1988) is a validated measure of workload across six dimensions (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration). The 'raw' version of the NASA-TLX without weighted rankings was given to the participants where the answer to each measure was on a scale of range 1–21 (Hart 2006).
- *Immersion Questionnaire* The immersion questionnaire from Högberg et al. (2019) was adapted. Participants are asked to give answers on a Likert scale ranging from 1 to 7 (from strongly disagree to strongly agree). A combined Immersion Score is calculated by taking the average of all the responses to items. See Appendix for a list of all items in the questionnaire.
- *Presence Questionnaire* The presence questionnaire was adapted from the physical presence subscale of the Multimodal Presence Scale (Makransky et al. 2017) and the telepresence questionnaire (Kim and Biocca 1997). Participants are asked to give answers on a Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). A combined Presence Score is calculated by taking the average of all the responses to items (after reversing responses to inverse questions). See Appendix for a list of all items in the questionnaire.
- *Enjoyment* The participants are asked to rate their agreement with the question *'The training session was very enjoyable'* on a Likert scale from 1 (strongly disagree) to 7 (strongly agree).
- *Self-efficacy* The participants are asked *'How confident are you that you can perform a similar task effectively (go from start to finish as fast as you can with minimal mistakes) on a scale from 1 to 7?'* to measure self-efficacy, once before the training and once after training. Details are provided in the next section.

3.3 Study procedure

Ethics approval was obtained from the local research ethics committee for experimenting with human subjects. The study was conducted in two rooms, one dedicated to IVR training and the other to physical training. Participants signed up for the study using the lab's online participant recruitment system. The system automatically filtered the participants using the following criteria based on self-reported data (i.e., they were not medically certified or independently verified): (a) right-handed, (b) normal vision or corrected to normal vision with contact lenses, and (c) no mental illnesses or sensitivity to nausea. The requirement for right-handedness was added to eliminate variation in the setup. Participants signed up for 45 min timeslots of their choosing and were paid the equivalent of 15 Euros. Each condition/room was run by one researcher at a time. The researchers switched between them regularly to reduce investigator effects. The timeslots for both conditions were open from 9 a.m. to 5 p.m. on weekdays.

At the beginning of a session, the participants were asked to read and sign the consent form. They were then briefly familiarized with the experiment procedure by allowing them to practice on the first level of the physical training setup. Thus, all participants, independent of condition, were provided a chance to experience the physical setup (Fig. 2), the cue for starting each task (when they hear the word 'Go'), and the proper way to lift the handle from the start position and to rest it on the end position. Thereafter, they were given the privacy to wear the Polar H10 around their chest as the researchers left the room. After this, the Shimmer GSR electrodes were placed on the index and middle fingers of the participant's left hand. The participant places her/his left hand on a Styrofoam support pad placed toward the left side of the table with the palms facing upwards and the fingers kept relaxed. The participant was asked not to move or flex her/his hand to minimize the noise in the recorded signals. The signal quality for both sensors was checked and verified in the iMotions software before the experiment started.

Baseline biosensor data were then measured by asking the participants to remain seated quietly and still with their eyes closed, without heavy breathing. The baseline HR and GSR data were then used to normalize subsequent signals since the baseline HR and GSR values for each person varied considerably. The participants were then presented with the test task before training begins (detailed in Sect. 3.1.1). They start the test task after hearing the word 'Go' from the researcher. Upon completion, they were then asked the question on self-efficacy. Following this, they were trained on four levels of increasing complexity in either VR or physical conditions (depending on the random assignment at the beginning of the experiment). In the physical condition, after each level of training, the researcher would rotate the wooden base by 90° (Fig. 2) so that the next level is facing the participant. This process took 10 to 15 s, which was absent in the VR condition where the switch to the next level was instantaneous. At the beginning of each level, they were asked to relax for 30 s by resting their right hand on their lap and start the task only when they hear the word 'Go,' this time from the headphone. After the training, the participant was asked the self-efficacy question again. They were then presented with a distractor task in the form of a maze to reduce the recency effect (Carlson et al. 2015; Winther et al. 2020). They were asked to spend about a minute both visualizing the solution and then picking up the maze with their right hand to solve it, in order to minimize recency effects (Bjork and Whitten 1974). They were finally given the test task and again asked to perform it as quickly as possible with the least number of mistakes possible. Following this, the participant was asked to remove the sensors and to fill out an online questionnaire containing the NASA-TLX questionnaire and questions on enjoyment, presence, and immersion. When the participant started performing either the test or training task, the researcher steps behind a panel to reduce biases in performance due to the Hawthorne effect (Demetriou et al. 2019).

No personal information was recorded, except for those required for compensating the study participants, which were handled according to university data protection policies. The researchers followed COVID-19 safety protocols, including sanitizing the sensors, table, and buzz-wire handles after every participant completed the experiment.

4 Results

The statistical analysis was performed using the statistical methods available in SciPy (Scientific Python) and Pingouin packages (Vallat 2018; Virtanen et al. 2020), and plots were generated using the Seaborn and Matplotlib Python packages (Hunter 2007; Waskom 2021). 87 participants were part of the study, divided between the physical ($N=42$) and VR training ($N=45$) conditions. 48 participants identified themselves as male, 37 as female, and 2 as other. 46 participants indicated their age group in the 18–24 range, and 36 indicated theirs in the range 25–34. The majority of participants in the VR condition (69%, $N=31$) indicated that they had tried a VR head-mounted display 1–5 times, 1 reported trying IVR 5–10 times, and 6 reported trying IVR more than 10 times, whereas 7 had never tried VR before. Data from eight participants had to be excluded from the analysis of performance metrics because of data loss arising from VR headset tracking errors, and the biosignals from 15 participants had to be excluded from analysis due to sensor errors. Shapiro–Wilk tests for normality were applied to all the variables, and if a variable was found to violate assumptions of normality, non-parametric statistical tests were used: Wilcoxon Signed Rank (Wilcoxon 1945) for paired, and Mann–Whitney U tests for independent tests (Mann and Whitney 1947), and the related W and U statistics are reported. When the variables used for comparison, both followed normal distributions, Student's t test and Welch's t-test (for unequal variances) were used to test for independence, and related t-statistic and Cohen's d are reported. A significance level of 0.05 was selected while interpreting the results of the statistical tests. The datasets analyzed during the current study are available from the corresponding author on reasonable request.

4.1 Improvement in performance

Figure 6 depicts the three performance metrics for the VR and physical conditions: task completion time (Fig. 6a), contact time (Fig. 6b), and improvement score (Fig. 6c) (see Sect. 3.3.2 for definitions). The task completion time

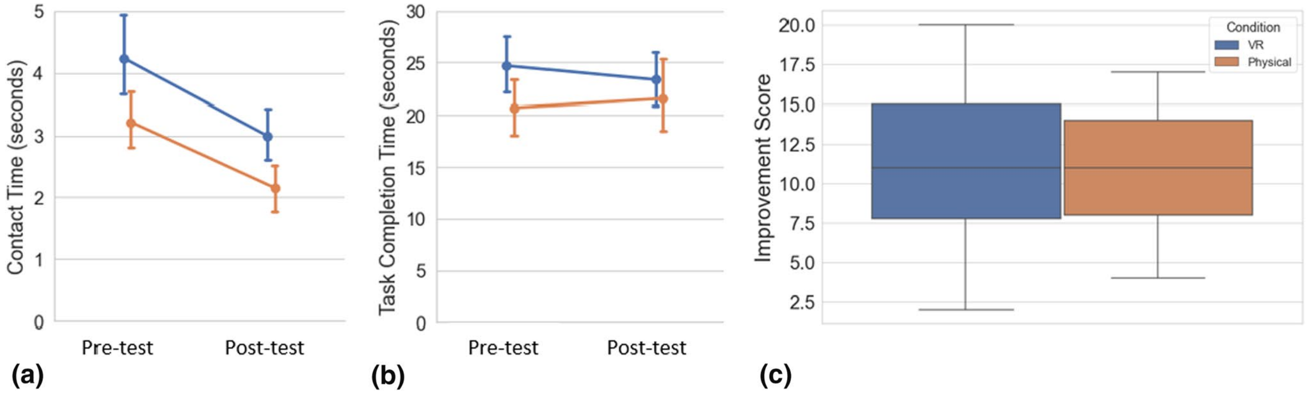


Fig. 6 Change in performance metrics within VR ($N=45$) and physical conditions ($N=42$) for **a** contact time, **b** task completion time, and **c** between the conditions for improvement score

and contact time metrics were analyzed to see if there were changes from the pre-training task to the post-training task. Analyses were also performed to see if there were statistically significant differences between the improvement scores of the two conditions.

4.1.1 Within-condition changes

In terms of contact time (CT), a statistically significant decrease of 1.21 s from pre- to post-training ($p<0.001$, $w=126.0$) was observed among participants in the VR condition ($N=40$). For the same group, a near statistically significant decrease of 1.33 s was observed in the task completion time (TCT) from pre-training to post-training phases ($p=0.062$, $w=352.0$). In the physical condition ($N=39$), there was a statistically significant decrease of 1.07 s in CT from pre- to post-training phases ($p<0.001$, $w=114.0$). On the other hand, though a slight deterioration of TCT may be observed in Fig. 6b for the physical condition from pre-training to post-training phases, this was not statistically significant (0.83 s, $p=0.412$, $w=387.0$).

4.1.2 Between conditions

To compare performance in VR ($N=40$) and physical ($N=39$) conditions, improvements in task completion time (TCT-I), contact time (CT-I), and improvement scores (IS) were calculated (see Sect. 3.3.2). Since the metrics from both these conditions were non-normally distributed, Mann–Whitney U independent samples tests were performed. The results showed no statistically significant differences between the improvement scores in the two conditions ($p=0.353$, $t(77)=-0.38$, $d=0.085$). Regarding

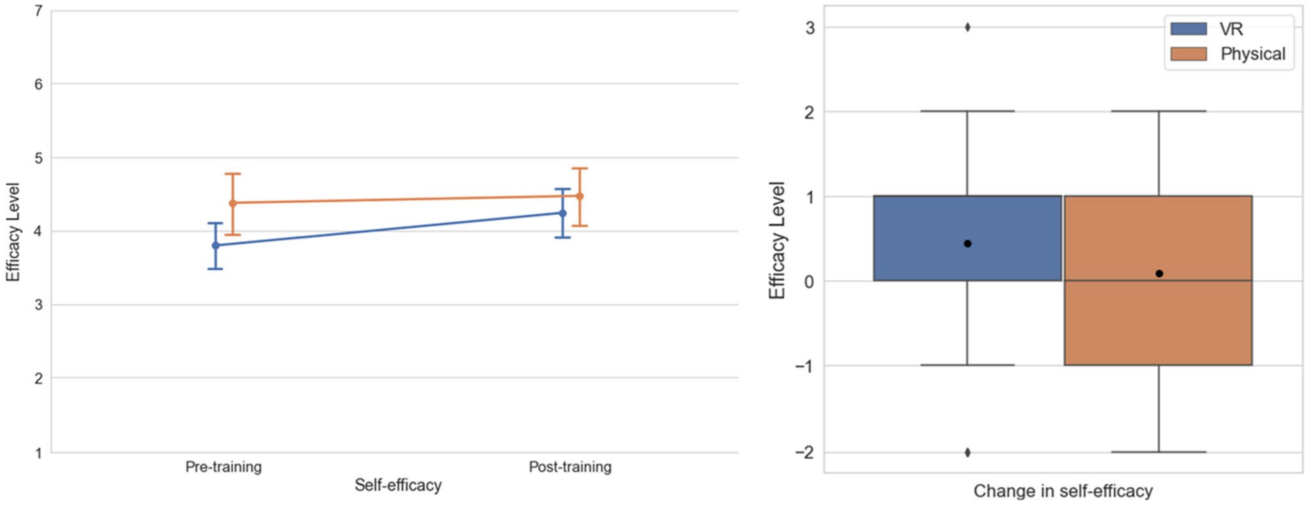


Fig. 7 Left: Self-efficacy levels from pre-training to post-training phases (on a scale of 1–7). Right: Change in self-efficacy levels across VR (N=45) and physical conditions ($N=42$)

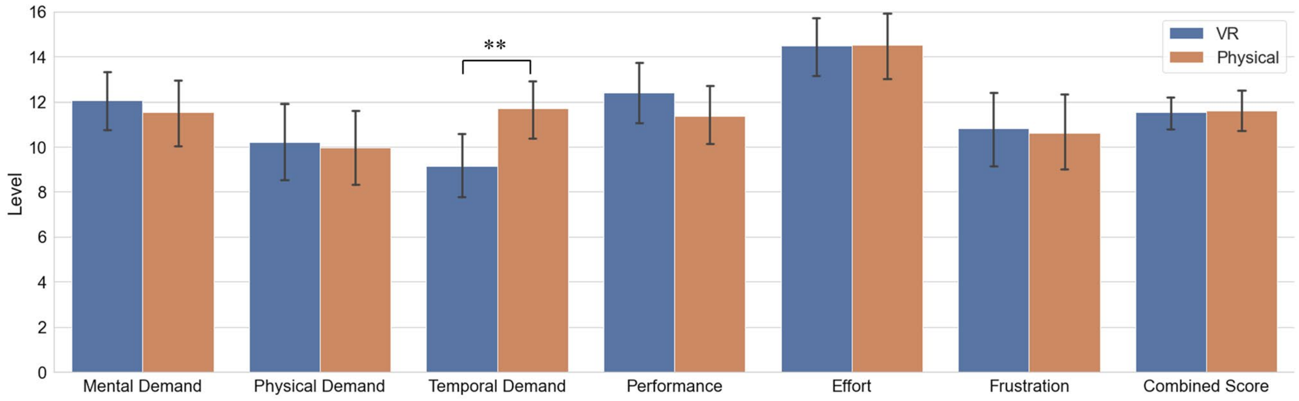


Fig. 8 NASA TLX Scores across VR ($N=45$) and physical conditions ($N=42$). ** denotes significant difference at $\alpha=0.05$

improvement in task completion time (TCT-I), though it can be seen from Fig. 6b that the task completion time for participants in the VR condition shows a visible improvement (i.e., decreases), this was not statistically significant ($p=0.2864$, $U=722$). CT-I also showed similar trends with participants in the VR condition showing no statistically significant differences with participants from the physical condition ($p=0.4746$, $U=773$).

4.2 Improvement in self-efficacy

As indicated in Fig. 7, in the VR condition ($N=45$), there is a statistically significant increase in the reported self-efficacy from the pre-training phase (3.8) to the post-training phase (4.24; $p=0.016$, $w=120.5$). Though a slight increase in reported self-efficacy in the physical condition (N=42) from the pre-training phase (4.38) to the post-training phase (4.48) can be observed in Fig. 7, this difference was found not to be statistically significant ($p=0.545$, $w=191.5$). It was also observed that the change in self-efficacy in the VR condition (0.44) was greater than the change in self-efficacy in the physical condition (0.095). This difference approaches statistical significance ($p=0.0585$, $U=767.5$).

4.3 Task load

Figure 8 shows the item-wise scores for NASA-TLX between the VR ($N=45$) and physical ($N=42$) conditions. Participants reported their perceived task load on six dimensions, i.e., mental, physical, and temporal demand, along with frustration, effort, and performance (Hart and Staveland 1988). Among these six dimensions, it can be observed that both VR and physical training result in similar task load values except for the temporal load parameter where participants in the physical condition report a mean score of 11.71 ± 2.87 (on a scale from 1 to 21) which is significantly higher than what participants in the VR condition reported (9.16 ± 2.49; $p=0.012$, $U=738.5$). There was no statistically significant difference in the combined NASA TLX Score between the physical (11.62 ± 2.87) and VR conditions (11.53 ± 2.49; $p=0.436$, $t(81.5)=0.161$, $d=0.03$).

4.4 Immersion, presence and enjoyment

Figure 9 shows the immersion, presence, and enjoyment scores between the VR (N=45) and physical (N=42) conditions. Cronbach's alpha coefficients were calculated for both questionnaires and found to be 0.88 for Immersion and 0.69 for Presence, indicating an acceptable internal consistency of the scales. An analysis of the Immersion Score (which is the mean of all items on the Immersion questionnaire) shows that participants in the VR condition report higher immersion on average (4.94 ± 0.99) as compared to participants in the physical condition (4.54 ± 0.98) and that this difference is statistically significant ($p=0.031$, $t(84.47)=-1.88$, $d=0.404$) with statistical significance also being observed for items I2, I4, and I9. Analysis of the combined Presence Score shows participants reporting a higher score on average for VR (4.61 ± 0.93) compared to physical (4.4 ± 0.79). This difference approaches statistical significance ($p=0.0736$, $U=774$) with statistical significance also being observed for items P5, P6, P10, and P14. See Tables 5, 6 in the Appendix for item-wise statistics for both Immersion and Presence questionnaires. Finally, participants report higher enjoyment for the VR condition (6.02 ± 1.23) as compared to physical condition (5.52 ± 1.15; $p=0.0175$, $U=696.5$).

4.5 Physiological arousal

4.5.1 Arousal levels between conditions

Table 3 lists all the physiological arousal metrics recorded during the training session. Only data points recorded between the start and finish points for each training level

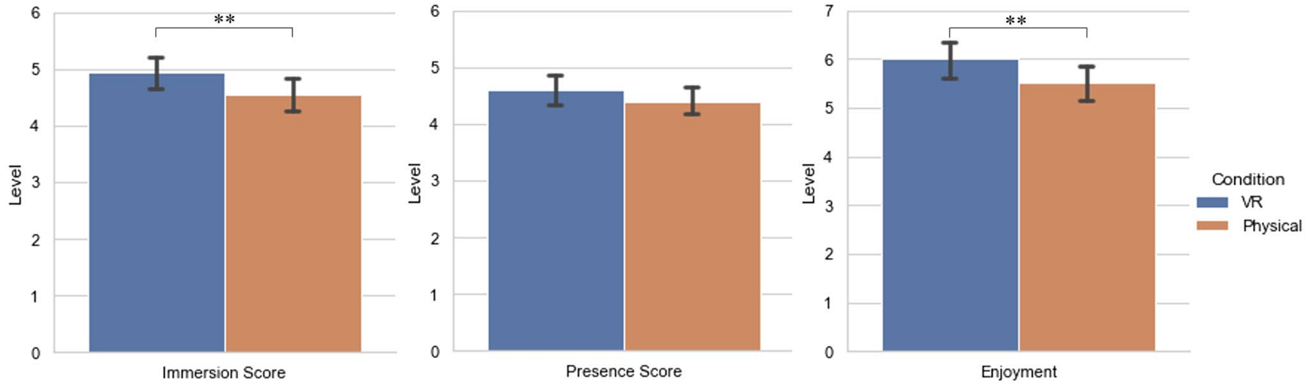


Fig. 9 Immersion, presence, and enjoyment scores across VR ($N=45$) and physical conditions ($N=42$). ** denotes significant difference at $\alpha=0.05$

have been considered and then averaged to generate arousal metrics that represent the whole training phase. The metrics listed have been adjusted to each participant's baseline where appropriate.

Among EDA metrics, in the VR condition ($N=33$), the mean SCRPeaks of 9.9 was found to be significantly lower than the SCRPeaks of 12.04 in the physical condition ($N=39$) denoting higher arousal among participants in the physical condition ($p=0.0032$, $U=885$). Among HRV measures, mean baseline-corrected HR was lower in VR (− 1.3) than physical (0.05) and the difference approaches statistical significance ($p=0.066$, $t(73.7)=1.52$, $d=0.34$). Showing similar trends, the mean baseline-corrected IBI was found to be higher in VR (16.86) than physical (1.33), but the difference is not statistically significant ($p=0.087$, $t(75.6)=-1.37$, $d=0.31$).

Comparisons between other EDA and HRV metrics showed no statistically significant differences though they mostly align with the findings in the SCRPeaks and IBI metrics with higher arousal in physical than in VR. Among EDA measures, the mean SC across all training levels in the VR condition ($N=33$) is 2.31, which is lower than the mean SC from the physical condition ($N=39$), 2.63. However, this difference is not statistically significant ($p=0.1221$, $U=747$). Mean SCRAmp for VR (0.19) is lower than physical (0.21), but the difference is not statistically significant ($p=0.445$, $U=676$). Among time-domain HRV measures, the mean baseline-corrected RMSSD for VR (− 22.22) is higher than physical (− 8.6) with no statistically significant difference ($p=0.614$, $U=732$) and the mean baseline-corrected SDNN in VR (− 18.34) is lower than physical (− 9.0) with no statistically significant difference ($p=0.3821$, $U=791$). Among frequency domain HRV metrics, mean baseline-corrected HFN in VR (5.83) is lower than physical (9.72) where the difference is not statistically significant ($p=0.195$, $t(71.8)=0.865$, $d=0.196$) and mean baseline-corrected LF/HF ratio in VR (− 1.02) is greater than physical with no statistically significant difference ($p=0.3086$, $U=710$).

4.6 Arousal level and performance

To assess the link between arousal levels and performance, data from both IVR and physical groups were combined, and Spearman rank correlation tests (for non-normal data) were performed between the physiological arousal metrics and

Table 3 Physiological arousal metrics across physical ($N=39$) and VR training conditions ($N=39$ for HRV, $N=33$ for EDA)

Physiological arousal metric		Physical (mean ± SD)	VR (mean ± SD)	p value
HRV	Mean HR#	0.05 ± 3.6	− 1.31 ± 4.29	0.066
	Mean IBI#	1.33 ± 48.15	16.86 ± 51.73	0.087
	Mean RMSSD#	− 8.6 ± 172.06	− 22.2 ± 127.24	0.614
	Mean SDNN#	− 9.0 ± 125.07	− 18.34 ± 79.84	0.3821
	Mean LF/HF ratio#	− 1.15 ± 3.52	− 1.02 ± 5.0	0.3086
	Mean HF Normalized#	9.72 ± 17.32	5.83 ± 22.19	0.1737
EDA	Mean SC#	2.63 ± 2.11	2.31 ± 2.21	0.1222
	Mean SCRAmp	0.21 ± 0.21	0.19 ± 0.16	0.445
	Mean SCRPeaks	12.04 ± 3.31	9.9 ± 2.29	0.0032**

**Denotes significant difference at $\alpha=0.05$, # denotes baseline-corrected metrics

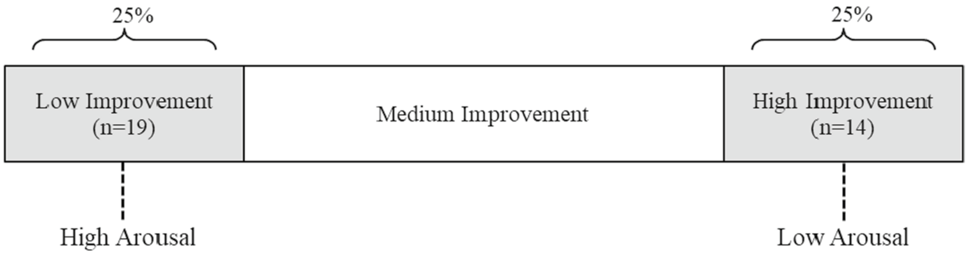


Fig. 10 The participants (from both conditions) were divided into high and low-performance groups. The high improvement group is in the upper 75th percentile of performance based on the improvement score. Similarly, the low improvement group is from the bottom 25th percentile. Participants who showed the highest improvement had lower arousal than those who had the lowest improvement

performance improvement metrics. The tests showed almost no correlation between arousal and improvement in performance with most ρ values between − 0.1 and 0.1. Notable statistically significant but weak correlations include the correlation between TCT-I and SCRAmp ($\rho = -0.24$, $p = 0.041$), TCT-I and SC ($\rho = -0.24$, $p = 0.0434$), and near statistically significant correlations include those between TCT-I and RMSSD ($\rho = -0.21$, $p = 0.068$) and IS and SCRAmp ($\rho = -0.19$, $p = 0.098$).

As part of a post hoc analysis to explore the relationship between arousal levels and performance, we defined two kinds of participants: high and low improvement groups in terms of their improvement score (IS) as denoted in Fig. 10. Those participants whose IS was greater than the upper bound of the IQR (inter-quartile range), i.e., the top 25%, were defined to be in the high improvement group ($N = 14$). Similarly, those participants whose IS was lesser than the lower bound of the IQR (the bottom 25%) were defined to be in the low improvement group ($N = 19$). Table 7 shows the results of Mann–Whitney U tests to compare the physiological arousal metrics between these two groups. Among statistically significant differences, the mean SCRAmp of the low improvement group (0.25) was greater than that of the high improvement group (0.12) ($p = 0.0298$, $U = 63$), and the mean SC of the low improvement group (3.49) was greater than that of the high improvement group (1.59) ($p = 0.0252$, $U = 55$).

5 Discussion

This section discusses the results and is structured around each of the four research questions formulated in the related works section.

5.1 Is IVR training as effective as physical training in improving task performance?

Both IVR and physical training result in statistically significant improvements in contact time (CT) from pre-training to post-training phases. This shows that participants from both training conditions achieved fewer mistakes while performing the task. Participants in the IVR group showed improvements in task completion time which neared statistical significance. However, for participants who underwent physical training, the task completion time did not show a statistically significant change. Overall, the results suggest that training in fine motor skills results in quantifiable performance improvements for participants in both IVR and physical training. This is expected and as per the literature on IVR-based skill training (Radhakrishnan et al. 2021b).

To compare the effectiveness of the two training modalities, three metrics to quantify improvement were defined: improvements in task completion time (TCT-I), improvements in contact time (CT-I), and an Improvement Score (IS) which combines the first two metrics. Statistical tests comparing these three metrics between IVR and physical conditions showed no statistically significant differences. Thus, the results indicate that IVR training is as effective as physical training for training in the buzz-wire task, thus supporting similar findings in other IVR skill training literature (Murcia-Lopez and Steed 2018; Schwarz et al. 2020). One can argue that the novelty effect of IVR might have played a role in its effectiveness as it was observed that 31 participants in the IVR condition had tried VR only 1–5 times before the study, and 7 had never tried VR before. In a review, Merchant et al. (2014) found a link between the novelty effect of desktop VR-based high school education and learning outcomes and that the latter may even decrease as the number of VR sessions increases. Thus, novelty in VR use can play a role yet as the current study utilized a short familiarization task prior to the actual experimental task, this effect can only be a small attribute of the observed effectiveness.

The current finding that IVR training is as good as physical training should also be considered in terms of the potential for further enhancement of this training modality. The literature suggests different methods to do this: the inclusion of haptic feedback (Frederiksen et al. 2020; Winther et al. 2020) and the inclusion of body representation and movements (other than the head and controllers) (Jensen and Konradsen 2018).

Inspirations for improving IVR training might also be taken from motor skill training literature which suggests techniques like decreasing the frequency of feedback as the skill level of the participant increases during training (Hebert and Coker 2021), allowing participants to choose whether they want to receive feedback or not (Chiviacowsky and Wulf 2005), or the IVR simulation adapting aspects of the training to the individual in real-time using physiological arousal levels and/or performance metrics (Zahabi and Abdul Razak 2020).

5.2 Is there a significant difference in the enjoyment, presence, immersion, task load, and changes in self-efficacy reported by participants in IVR compared to physical training?

Participants in the IVR condition reported on average significantly more enjoyment levels than participants in the physical condition. This finding is consistent with IVR literature (Makransky et al. 2019). One parameter that is typically associated with frustration and lack of enjoyment during a VR experience is cybersickness. Herein, there were no incidents of cybersickness reported by the participants, probably due to the seated arrangement. Participants in the IVR condition reported on average more immersion (with statistical significance) than those in the physical condition. Similar trends exist for the presence measure, with participants in IVR training reporting more presence than those in physical training, where the difference was found to approach statistical significance. Though the IVR condition showed higher presence and immersion scores compared to the physical condition, it should be kept in mind that results from such metrics gain more importance when all subjects experience the same environment (Usoh et al. 2000). Nevertheless, the results are encouraging and as expected, as participants did not feel less immersed or present in the IVR environment as compared to the physical.

The NASA-TLX results show that in all parameters except temporal demand, IVR training induces roughly the same workload on participants as physical training. This was expected, as all kinds of visual noise and other confounding variables were tightly controlled across both conditions. However, VR, if not designed properly, may cause more cognitive load due to the possible complexity and novelty of the VR interactions involved. The one task load parameter where IVR training shows a statistically significant advantage over physical training is temporal demand. However, one cannot draw clear conclusions from this finding and further research is needed, for example, to compare the total training time across both conditions (which was not part of the research questions) along with the perceived temporal demand. This opens up interesting possibilities, due to the presence of a 'time compression' effect in IVR as observed by Mullen and Davidenko (2021), where subjects experienced time to speed up while using VR compared to those in the control condition.

Participants in both physical and IVR training conditions reported an increase in self-efficacy, though a statistically significant increase was found only for the IVR group. Increases in self-efficacy levels have been found to correlate positively with learning outcomes (Makransky et al. 2019; Shu et al. 2019) and motor skill performance (Bandura 1986). However, both VR and physical training in the current study did not show different levels of improvement in performance. It is possible that the novelty effects of VR caused participants in the VR condition to start initially with a lower self-efficacy in spite of the VR familiarization, but they ended up with self-efficacy levels similar to the physical condition by the end of the training. Further research is required to understand the links between self-efficacy and familiarity with the IVR medium. Additionally, these participants in the VR condition were observed to both have lower physiological arousal along with their increased self-efficacy. According to Bandura (1986)'s model of self-efficacy, there is a possible interaction between self-efficacy and arousal which merits further research in the context of IVR skill training.

5.3 Is there a significant difference between the physiological arousal levels of participants in IVR training compared to physical training?

Analysis of EDA and HRV metrics from the physiological arousal data revealed that IVR training caused less arousal than physical training, with a significant difference found for the SCRPeaks (EDA) metric and a near significant difference found for the HRV metrics Heart Rate and Inter-Beat Intervals. However, the frequency domain HRV measures, i.e., HFN, LF/HF ratio, and the time domain HRV measures SDNN and RMSSD showed no statistically significant difference.

Though there is no literature on the comparison of arousal between IVR and non-IVR conditions for skill training, some indicative literature from other domains exists. Tian et al. (2021) found more physiological arousal (EDA, EEG measures) in participants being emotionally stimulated through videos in the IVR condition as compared to those in the 2D condition. Egan et al. (2016) in a comparative quality of experience study found greater HR in the IVR condition compared to the non-IVR 2D condition, while they found that EDA showed the opposite trend to our finding. We discuss possible causes for these seemingly contradictory trends toward the end of this section.

5.4 Is there a link between physiological arousal during training and improvements in performance after training?

A post hoc analysis was performed to compare the physiological data from participants with the highest improvement to those with the lowest improvement. This revealed greater arousal in two EDA measures (mean amplitude of skin conductance responses and mean skin conductance) for those participants who improved the least as compared to those who improved the most. This result is in alignment with findings from the literature; for example, in surgical simulation training (non-IVR), it has been found that lower performance is correlated with increased stress (higher arousal) levels (Prabhu et al. 2010; Quick et al. 2017). When correlation analysis was performed to compare the different arousal metrics with performance metrics for the whole study sample, we found statistically significant but weak correlations for improvement in task completion time (TCT-I) and among two EDA metrics: mean amplitude of skin conductance responses during training (SCRAmp) and mean skin conductance (SC). Further research should investigate the link between performance and arousal for participants across all levels of performance improvement.

For the last two research questions (links between arousal and training condition, arousal, and performance improvements), we found significant differences only in EDA metrics but not in HRV. This might be because EDA is purely a measure of sympathetic activity, as skin conductance levels are not counteracted by the parasympathetic nervous system. On the other hand, heart rate activity is controlled by both the sympathetic system (which causes heart activity to increase) and the parasympathetic system (which causes heart rate activity to decrease back to the baseline) (Cacioppo et al. 2007). Some literature finds EDA measures to be superior in terms of measuring changes in arousal (Dawson et al. 2016), even above HRV (Healey and Picard 2005).

6 Limitations

Motor skill learning literature indicates the possibility that short-term performance might misrepresent learning (Magill and Anderson 2016). Although a distractor task (see Sect. 3.4) was used in the current study to compensate for the short-term nature of the retention test, it may be necessary to perform the retention tests after longer intervals to give a more precise understanding of the relationship between training conditions and retention. IVR skill training literature points to many comparative studies where retention tests after long intervals show better or the same retention in performance for the IVR condition as compared to non-immersive VR and physical conditions (Butt et al. 2018; Buttussi and Chittaro 2018; Sakowitz et al. 2019). An illustrative example is in the burr-puzzle solving task by Carlson et al. (2015), where participants in a physical training condition initially outperformed those in IVR in terms of knowledge retention, but after two weeks, this effect was reversed. These examples suggest that such results may be expected in contexts similar to the current study; however, further research is still required.

The study was also purposefully limited in terms of the 'training' provided. Here, participants were not given instructions during or after the training (knowledge of results), but participants get only automated feedback during the training when mistakes were made (knowledge of performance). Further research may build upon the design of the experiment and incorporate different training strategies or instructions. Also, the study is limited only to people who self-reported to be right-handed, to better control the setup and minimize variations, but future research might consider designing buzz-wire arrangements that are compatible with left-handed participants.

Regarding considerations on arousal metrics, comparisons using HRV metrics in the current study showed a lack of significant results. This could potentially be explained if it is assumed that the main cause for arousal in the current task was contact feedback (audio-visual-haptic). Since the time spent by a participant in contact with the wire (i.e., committing mistakes) will only be a proportion of the total duration of the training, any short-term increases in HRV metrics (which is accompanied by a rapid return to normal) may get averaged out by variations in HRV metrics during the rest of the training where they do not make any mistakes. Another potential confounder, which could cause variation in HRV, is the physical aspect of the activity where the participant has the freedom to choose any possible configurations of hand-arm-shoulder movement to complete the task with their right hand. Controlling this was beyond the scope of the current setup. Future studies may require a more fine-grained analysis of the relation between different stimuli (feedback during mistakes, difficulty in navigating certain parts of the wires) and physiological signals. Inspirations from the literature include Liebold et al. (2017) where a post-stimuli window of 10 s was used for heart rate metrics and Boucsein (2012) which recommends a 1–5 s post-stimuli window to detect event-related skin conductance responses (ER-SCRs).

Regarding the choice of sensors used, the study is limited to only two measures of physiological signals (EDA and HRV). There is a multitude of physiological sensors which can be used to detect physiological arousal like electroencephalogram (EEG), skin temperature, and eye-tracking. Additional sensors were not used as they might have made the experimental procedure more complex and affected the behavior of the participants. However, additional sources of biosignals merit further exploration in IVR training research as there are indications that some signals may make others redundant, for example, pupil dilation (from eye-tracking sensors) has been found to be correlated with both EDA and HRV (Wang et al. 2018). It is known that melatonin (which is correlated with the time of day), and temperature affect HRV and EDA metrics (Boucsein 2012; Schachinger et al. 2008), but these factors were not controlled for in the experiment. On the other hand, these effects may have been reduced by the baseline correction applied to the various arousal metrics. Though arousal in this study is averaged across all the training sessions, the long recovery periods lasting several minutes for HRV signals to return to baseline levels (Moses et al. 2007) might potentially result in arousal from one level of training affecting the next. However, this issue may not affect EDA metrics, as a half recovery period from 2 to 10 s is found in the literature (Dawson et al. 2016), which is within the range of the 30 s rest interval between each level. The current study did not control for color blindness, and the self-reported normal vision of the participants was not medically certified, both of which might have caused differences in performance between the conditions.

A related factor affecting our study is the inherent difference between the haptic feedback available in the IVR and physical conditions. Though the vibration aspect is identical in both conditions, in the physical condition, there is the added feel of the physical wire though the vibration masks this feeling to a certain degree. We propose further experimentation in IVR modality alone, with conditions being varied for various haptic feedback modalities like portable, grounded, and wearable as observed by Radhakrishnan et al. (2021b) in their analysis of the use of haptics in industrial skills training. The investigation of possible links between haptic feedback modality, physiological arousal, and improvements in performance holds promise for improving the state of the art in IVR-based skills training.

7 Implications for researchers

Taking as a point of departure the findings and lessons learned from this study one may consider:

- IVR and other training modalities must be designed to minimize distractions. This study tries to achieve this by using black panels covering the peripheral view of the participant and using headphones which, in addition to providing audio feedback, also minimizes external noise. In their review of motor skill learning literature, Wulf et al. (2010) found that performance is increased when there is an 'external focus' directed at the effect of the movement itself instead of an 'internal focus' directed at the trainee's body movements. Therefore, it is recommended that such complexities be minimized unless there are reliable methods of representing hands, arms, and other relevant parts of the body realistically. The coherence principle from the Cognitive Theory of Multimedia Learning further supports this by stating that removing stimuli irrelevant to the training context can improve learning outcomes (Parong and Mayer 2021).
- *VR hardware* The use of the Oculus Quest often requires minor calibrations related to the setting of tracking boundaries. This may be avoided by making sure the study environment is consistent between sessions or by using external trackers.
- *Polar H10* This cost-effective yet highly accurate and reliable ECG heart monitor is a useful tool for measuring arousal levels (Polar Electro Oy, Kempele, Finland). Researchers should, however, take into consideration the time taken for setting up the device and for the study setup to give privacy and instructions to participants for properly wearing the device.
- *Shimmer GSR +* This is a cost-effective and reliable device for measuring electro-dermal activity (EDA) (Shimmer Research Ltd., Dublin, Ireland). The opportunity of measuring high-quality EDA signals from the fingers also restricts the training task from involving bimanual skills (use of both hands). Alternative but less accurate/convenient locations on the body can be considered if a training task demands the use of both hands (van Dooren and Janssen 2012).
- *Buzz-wire task* This task allows for one-hand use making it convenient for studies using EDA. The training task itself provides immediate feedback and allows for variations, for example, different types of audio, visual, or haptic feedback.

8 Conclusion

The study suggests that for the fine motor skill training presented, IVR training is as effective as physical training in improving task performance. Participants in the IVR condition reported an improvement in self-efficacy and significantly more enjoyment and immersion than physical training. Also, participants in the IVR condition on average displayed lower arousal than physical training. Though clear indications on the relationship between arousal and improvements in performance could not be found, EDA metrics hold potential for further investigation to answer this question by showing differences in arousal between high and low improvement groups. It is our understanding that such findings add to the IVR training field and can potentially pave the way to user-adaptive training systems (Zahabi and Abdul Razak 2020).

Future work could incorporate subjective measures of arousal (like the Self-Assessment Manikin) into the immersive VR training as an additional layer to confirm findings from the physiological arousal signals. Additional measures like EEG could be employed to investigate the effect of the different types of stimuli on different brain regions, resultant cognitive load, and their relationship with arousal and performance (Hofmann et al. 2021; Tian et al. 2021). However, this should be implemented in a manner that does not break immersion/presence. It should also be noted that the current study does not explore the origins of the physiological arousal observed during the study but only its effects on performance improvement. It is reasonable to assume that the arousal observed may have been primarily caused by the direct feedback provided (visual, audio, and haptic), but other factors may also play a role. The study tries to control such extraneous factors by features in the study design like providing an initial baseline phase for the users to relax and also rest periods between training levels. The present study does not go into a fine-grained analysis of the relationship between arousal and stimuli like feedback from mistakes or challenging parts like bends in the wire, but rather looks at arousal across the whole training phase. There could be merit in understanding the short-term changes in arousal for various kinds of stimuli; for example, haptic feedback which is increasingly becoming a major focus point for IVR research as it affects task performance and presence (Kreimeier et al. 2019) and is crucial for many fine motor skill training tasks in VR like surgery (Rangarajan et al. 2020). This study also considers averaged performance metrics across the entire training session to answer the primary research questions, but future work might consider variations during the motor skill training, particularly in understanding different control strategies and stages of learning (Sternad 2018). Future studies may also try to incorporate a cross-over study methodology in order to control for difference between groups, by exposing the same group of participants to counterbalanced exposures to VR and physical training with appropriate time intervals in between to reduce cross-over effects similar to Yin et al. (2019).

Appendix

See Fig. 11 and Tables 4, 5, 6, 7, 8, 9 and 10.

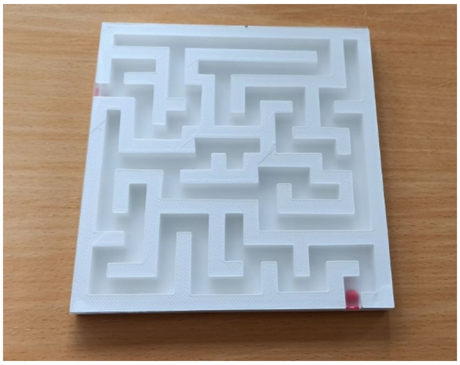


Fig. 11 Distractor maze task

Table 4 Physiological arousal metrics across high (*N* = 14 for HRV, *N* = 11 for EDA) and low improvement groups (*N* = 19 for HRV, *N* = 18 for EDA)

Physiological arousal metric		High improvement (N = 14) (Mean ± SD)	Low improvement (N = 19) (Mean ± SD)	*p* value
HRV	Mean HR#	− 0.11 ± 3.17	− 1.24 ± 3.76	0.196
	Mean IBI#	6.17 ± 46.41	14.11 ± 53.97	0.196
	Mean RMSSD#	38.69 ± 274.99	− 57.28 ± 120.41	0.6215
	Mean SDNN#	27.15 ± 201.54	− 38.22 ± 73.97	0.5935
	Mean LF/HF ratio#	− 2.02 ± 7.78	− 0.37 ± 1.41	0.7013
	Mean HF normalized#	0.58 ± 22.21	7.15 ± 19.15	0.2167
EDA	Mean SC#	1.59 ± 0.9	3.49 ± 2.74	0.0252**
	Mean SCRAmp	0.12 ± 0.09	0.25 ± 0.18	0.0298**
	Mean SCRPeaks	12.03 ± 2.29	11.92 ± 2.85	0.4198

**Denotes significant difference at $\alpha = 0.05$

#Denotes baseline-corrected metrics

Table 5 Presence questionnaire

1. During the training session, I forgot that I was in a lab
2. The training session totally filled my mind
3. During the training session, I was very captivated by what was presented to me
4. I felt like I really was present (‘was there’) during the training
5. When the training session was over, I felt like I was back from a journey
6. During the training session, I was not conscious of the room setup (assistant, speaker, chairs…)
7. During the training session, I lost the notion of time
8. During the training session, I was living what I was seeing as if it was happening to me for real
9. I lived the experience of performing the training intensely
10. During the training session, I often thought of something else. (Inverted Question)
11. I felt more like a participant than spectator of the training
12. I had to force myself to stay concentrated on the training session. (Inverted Question)
13. I always had in mind the fact that I was in a lab. (Inverted Question)
14. I was reacting to everything I was seeing as it was real

Table 6 Immersion questionnaire

The motor skill training experience …

1. Makes me feel immersed
2. Gives me the feeling that time passes quickly
3. Grabs all of my attention
4. Gives me a sense of being separated from the real world
5. Makes me lose myself in what I am doing
6. Makes my actions seem to come automatically
7. Causes me to stop noticing when I get tired
8. Causes me to forget about my everyday concerns
9. Makes me ignore everything around me
10. Gets me fully emotionally involved
11. Captivates me

Table 7 Performance metrics for the VR and physical conditions: IS (improvement score), TCT-I (improvement in task completion time), and CT-I (improvement in contact time)

Performance metric	VR (mean ± SD)	Physical (mean ± SD)	*p* value
IS	11.18 ± 5.11	10.79 ± 3.67	0.343
TCT-I	1.33 s ± 8.57 s	− 0.83 s ± 7.37 s	0.286
CT-I	1.24 s ± 2.04 s	1.06 s ± 1.14 s	0.474

Table 8 Presence scores

Presence item	Physical (mean ± SD)	IVR (mean ± SD)	*p* value
P1	3.19 ± 1.77	3.76 ± 1.63	0.0623
P2	4.88 ± 1.53	5.16 ± 1.22	0.2041
P3	5.31 ± 1.2	5.58 ± 1.27	0.1179
P4	5.43 ± 1.04	5.04 ± 1.45	0.2006
P5	3.45 ± 1.64	4.07 ± 1.54	0.0359**
P6	3.12 ± 1.6	4.27 ± 1.9	0.0018**
P7	4.62 ± 1.61	5.09 ± 1.5	0.0757
P8	4.93 ± 1.58	4.42 ± 1.86	0.1285
P9	4.79 ± 1.32	4.82 ± 1.4	0.4068
P10	3.86 ± 1.63	4.71 ± 1.18	0.0042**
P11	5.74 ± 1.13	5.53 ± 1.53	0.4438
P12	4.07 ± 1.63	4.58 ± 1.44	0.0512
P13	2.81 ± 1.67	2.84 ± 1.58	0.4398
P14	5.43 ± 1.35	4.67 ± 1.73	0.02348**
Combined	4.4 ± 0.79	4.61 ± 0.93	0.0736

**Denotes significant difference at α = 0.05

Table 9 Immersion scores

Immersion item	Physical (mean ± SD)	IVR (mean ± SD)	*p* value
I1	4.9 ± 1.39	5.02 ± 1.37	0.2899
I2	4.95 ± 1.34	5.36 ± 1.28	0.0439**
I3	5.6 ± 1.19	5.64 ± 1.26	0.3436
I4	3.64 ± 1.64	4.78 ± 1.51	0.0007**
I5	4.48 ± 1.78	4.64 ± 1.61	0.3528
I6	3.88 ± 1.71	4.33 ± 1.46	0.1225
I7	3.93 ± 1.79	4.53 ± 1.5	0.0501
I8	4.79 ± 1.69	4.96 ± 1.57	0.3224
I9	4.14 ± 1.57	4.87 ± 1.24	0.0114**
I10	4.38 ± 1.65	4.71 ± 1.41	0.1748
I11	5.29 ± 1.2	5.51 ± 1.2	0.1461
Combined	4.54 ± 0.99	4.94 ± 0.98	0.01751**

**Denotes significant difference at α = 0.05

Table 10 NASA-TLX scores

Item	Physical (mean ± SD)	IVR (mean ± SD)	*p* value
Mental demand	11.52 ± 4.55	12.1 ± 4.35	0.2738
Physical demand	9.95 ± 5.9	10.22 ± 5.76	0.4008
Temporal demand	11.71 ± 4.33	9.16 ± 4.83	0.006**
Performance	11.38 ± 4.38	12.4 ± 4.6	0.1123
Effort	14.52 ± 4.84	14.49 ± 4.35	0.4073
Frustration	10.62 ± 5.66	10.82 ± 5.84	0.4257
NASA-TLX score	11.62	11.53	0.379

**Denotes significant difference at α = 0.05

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10055-022-00699-3.

Acknowledgments We thank colleagues at the Department of Business Development and Technology for their feedback, and staff at the Cognition and Behaviour lab (Aarhus University) and iMotions for providing support.

Declarations

Conflict of interest There are no conflicts of interest.

References

Abich J, Parker J, Murphy JS, Eudy M (2021) A review of the evidence for training effectiveness with virtual reality technology. Virtual Real 25:919–933. https://doi.org/10.1007/s10055-020-00498-8

Adhanom IB, Al-Zayer M, Macneilage P, Folmer E (2021) Field-of-view restriction to reduce VR sickness does not impede spatial learning in women. ACM Trans Appl Percept. https://doi.org/10.1145/3448304

Aggarwal R, Moorthy K, Darzi A (2004) Laparoscopic skills training and assessment. Br J Surg 91:1549–1558. https://doi.org/10.1002/bjs.4816

Ashiri M, Lithgow B, Suleiman A, Blakley B, Mansouri B, Moussavi Z (2020) Differences between physical vs. virtual evoked vestibular responses. Ann Biomed Eng 48:1241–1255. https://doi.org/10.1007/s10439-019-02446-3

Bandura A (1986) The explanatory and predictive scope of self-efficacy theory. J Soc Clin Psychol 4:359–373. https://doi.org/10.1521/jscp.1986.4.3.359

Binsch O, Bottenheft C, Landman A, Roijendijk L, Vermetten EHGJM (2021) Testing the applicability of a virtual reality simulation platform for stress training of first responders. Mil Psychol 33:182–196. https://doi.org/10.1080/08995605.2021.1897494

Bjork RA, Whitten WB (1974) Recency-sensitive retrieval processes in long-term free recall. Cogn Psychol 6:173–189. https://doi.org/10.1016/0010-0285(74)90009-7

Boucsein W (2012) Electrodermal activity. Springer, New York

Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Ther Exp Psychiatry 25:49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Brooke J (1996) SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B (eds) Usability evaluation in industry. https://doi.org/10.1201/9781498710411

Budini F, Lowery MM, Hutchinson M, Bradley D, Conroy L, De Vito G (2014) Dexterity training improves manual precision in patients affected by essential tremor. Arch Phys Med Rehabil 95:705–710

Butt AL, Kardong-Edgren S, Ellertson A (2018) Using game-based virtual reality with haptics for skill acquisition. Clin Simul Nurs 16:25–32. https://doi.org/10.1016/j.ecns.2017.09.010

Buttussi F, Chittaro L (2018) Effects of different types of virtual reality display on presence and learning in a safety training scenario. IEEE Trans Visual Comput Graph 24:1063–1076. https://doi.org/10.1109/TVCG.2017.2653117

Cacioppo JT, Tassinary LG, Berntson G (2007) Handbook of psychophysiology. Cambridge University Press, Cambridge

Calderon DP, Kilinc M, Maritan A, Banavar JR, Pfaff D (2016) Generalized CNS arousal: an elementary force within the vertebrate nervous system. Neurosci Biobehav Rev 68:167–176. https://doi.org/10.1016/j.neubiorev.2016.05.014

Carlson P, Peters A, Gilbert SB, Vance JM, Luse A (2015) Virtual training: learning transfer of assembly tasks. IEEE Trans vis Comput Graph 21:770–782. https://doi.org/10.1109/TVCG.2015.2393871

Cebeci B, Celikcan U, Capin TK (2019) A comprehensive study of the affective and physiological responses induced by dynamic virtual reality environments. Comput Animat Virtual Worlds 30:e1893. https://doi.org/10.1002/cav.1893

Champseix R, Ribiere L, Le Couedic C (2021) A python package for heart rate variability analysis and signal preprocessing. J Open Res Soft 9(1):28. http://doi.org/10.5334/jors.305

Checa D, Bustillo A (2020) A review of immersive virtual reality serious games to enhance learning and training. Multimed Tools Appl 79:5501–5527. https://doi.org/10.1007/s11042-019-08348-9

Chiviacowsky S, Wulf G (2005) Self-Controlled feedback is effective if it is based on the learner's performance. Res Q Exerc Sport 76:42–48. https://doi.org/10.1080/02701367.2005.10599260

Christou CG, Michael-Grigoriou D, Sokratous D, Tsiakoulia M BuzzwireVR (2018) An immersive game to supplement fine-motor movement therapy. In: ICAT-EGVE. pp 149–156

Coban M, Bolat YI, Goksu I (2022) The potential of immersive virtual reality to enhance learning: a meta-analysis. Educ Res Rev 36:100452. https://doi.org/10.1016/j.edurev.2022.100452

Collet C, Petit C, Priez A, Dittmar A (2005) Stroop color–word test, arousal, electrodermal activity and performance in a critical driving situation. Biol Psychol 69:195–203. https://doi.org/10.1016/j.biopsycho.2004.07.003

Collins J, Regenbrecht H, Langlotz T, Said Can Y, Ersoy C, Butson R (2019) Measuring cognitive load and insight: a methodology exemplified in a virtual reality learning context. In: 2019 IEEE international symposium on mixed and augmented reality (ISMAR), 10/2019. IEEE, Beijing. pp 351–362. doi:https://doi.org/10.1109/ISMAR.2019.00033

Cuervo E, Chintalapudi K, Kotaru M (2018) Creating the perfect illusion: What will it take to create life-like virtual reality headsets?. In: Paper presented at the proceedings of the 19th international workshop on mobile computing systems & applications, Tempe, Arizona.

Dawson ME, Schell AM, Filion DL (2016) The Electrodermal System. In: Berntson GG, Cacioppo JT, Tassinary LG (eds) Handbook of psychophysiology. Cambridge handbooks in psychology, 4th edn. Cambridge University Press, Cambridge, pp 217–243

Dawson ME, Schell AM, Filion DL (2016) The electrodermal system. In: The Handbook of psychophysiology. Cambridge University Press, pp 217–243. https://doi.org/10.1017/9781107415782.010

Demetriou C, Hu L, Smith TO, Hing CB (2019) Hawthorne effect on surgical studies. ANZ J Surgery 89:1567–1576. https://doi.org/10.1111/ans.15475

Diemer J, Lohkamp N, Mühlberger A, Zwanzger P (2016) Fear and physiological arousal during a virtual height challenge—effects in patients with acrophobia and healthy controls. J Anxiety Disorders 37:30–39. https://doi.org/10.1016/j.janxdis.2015.10.007

Egan D, Brennan S, Barrett J, Qiao Y, Timmerer C, Murray N (2016) An evaluation of Heart rate and electrodermal activity as an objective qoe evaluation method for immersive virtual reality environments. In: 2016 eighth international conference on quality of multimedia experience (QoMEX), 6–8 June 2016. pp 1–6. doi:https://doi.org/10.1109/QoMEX.2016.7498964

Feng Z, González VA, Amor R, Lovreglio R, Cabrera-Guerrero G (2018) Immersive virtual reality serious games for evacuation training and research: a systematic literature review. Comput Educ 127:252–266. https://doi.org/10.1016/j.compedu.2018.09.002

Frederiksen JG, Sørensen SMD, Konge L, Svendsen MBS, Nobel-Jørgensen M, Bjerrum F, Andersen SAW (2020) Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: a randomized trial. Surg Endosc 34:1244–1252. https://doi.org/10.1007/s00464-019-06887-8

Gilgen-Ammann R, Schweizer T, Wyss T (2019) RR interval signal quality of a heart rate monitor and an ECG Holter at rest and during exercise. Eur J Appl Physiol 119:1525–1532. https://doi.org/10.1007/s00421-019-04142-5

Hamilton D, McKechnie J, Edgerton E, Wilson C (2021) Immersive virtual reality as a pedagogical tool in education: a systematic literature review of quantitative learning outcomes and experimental design Journal of. Comput Educ 8:1–32. https://doi.org/10.1007/s40692-020-00169-2

Hart SG (2006) NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting, vol 9. Sage publications Sage CA, Los Angeles, CA. pp 904–908

Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Advances in psychology, vol 52. Elsevier, Amsterdam, pp 139–183

Harvey C, Selmanović E, O'Connor J, Chahin M (2019) A comparison between expert and beginner learning for motor skill development in a virtual reality serious game. Vis Comput. https://doi.org/10.1007/s00371-019-01702-w

Healey JA, Picard RW (2005) Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans Intell Transp Syst 6:156–166. https://doi.org/10.1109/TITS.2005.848368

Hebert EP, Coker C (2021) Optimizing feedback frequency in motor learning: self-controlled and moderate frequency KR enhance skill acquisition. Percept Mot Skills 128:2381–2397. https://doi.org/10.1177/00315125211036413

Hofmann SM, Klotzsche F, Mariola A, Nikulin V, Villringer A, Gaebler M (2021) Decoding subjective emotional arousal from EEG during an immersive virtual reality experience. Elife 10:e64812. https://doi.org/10.7554/eLife.64812

Högberg J, Hamari J, Wästlund E (2019) Gameful experience questionnaire (GAMEFULQUEST): an instrument for measuring the perceived gamefulness of system use. User Model User-Adapt Interact 29:619–660. https://doi.org/10.1007/s11257-019-09223-w

Homer BD, Plass JL, Rose MC, MacNamara AP, Pawar S, Ober TM (2019) Activating adolescents' "hot" executive functions in a digital game to train cognitive skills: The effects of age and prior abilities. Cogn Dev 49:20–32. https://doi.org/10.1016/j.cogdev.2018.11.005

Howard MC (2017) A meta-analysis and systematic literature review of virtual reality rehabilitation programs. Comput Hum Behav 70:317–327. https://doi.org/10.1016/j.chb.2017.01.013

Huber T, Wunderling T, Paschold M, Lang H, Kneist W, Hansen C (2018) Highly immersive virtual reality laparoscopy simulation: development and future aspects. Int J Comput Assist Radiol Surg 13:281–290. https://doi.org/10.1007/s11548-017-1686-2

Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95

Jain S, Lee S, Barber SR, Chang EH, Son Y-J (2020) Virtual reality based hybrid simulation for functional endoscopic sinus surgery IISE transactions on healthcare. Syst Eng 10:127–141. https://doi.org/10.1080/24725579.2019.1692263

Janelle CM (2002) Anxiety, arousal and visual attention: a mechanistic account of performance variability. J Sports Sci 20:237–251. https://doi.org/10.1080/026404102317284790

Jensen L, Konradsen F (2018) A review of the use of virtual reality head-mounted displays in education and training. Educ Inf Technol 23:1515–1529. https://doi.org/10.1007/s10639-017-9676-0

Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG (1993) Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. Int J Aviat Psychol 3:203–220. https://doi.org/10.1207/s15327108ijap0303_3

Khalfa S, Isabelle P, Jean-Pierre B, Manon R (2002) Event-related skin conductance responses to musical emotions in humans. Neurosci Lett 328:145–149. https://doi.org/10.1016/S0304-3940(02)00462-7

Khan R, Plahouras J, Johnston BC, Scaffidi MA, Grover SC, Walsh CM (2019) Virtual reality simulation training in endoscopy: a Cochrane review and meta-analysis. Endoscopy 51:653–664

Kim T, Biocca F (1997) Telepresence via television: Two dimensions of telepresence may have different connections to memory and persuasion. J Comput-Mediat Commun. https://doi.org/10.1111/j.1083-6101.1997.tb00073.x

Koumaditis K, Chinello F, Mitkidis P, Karg ST (2020) Effectiveness of virtual vs. physical training: the case of assembly tasks, trainer's verbal assistance and task complexity. IEEE Comput Graph Appl. https://doi.org/10.1109/MCG.2020.3006330

Kreimeier J, Hammer S, Friedmann D, Karg P, Bühner C, Bankel L, Götzelmann T (2019) Evaluation of different types of haptic feedback influencing the task-based presence and performance in virtual reality. In: Paper presented at the proceedings of the 12th ACM international conference on PErvasive technologies related to assistive environments, Rhodes, Greece

Krogmeier C, Mousas C, Whittinghill D (2019) Human–virtual character interaction: Toward understanding the influence of haptic feedback. Comput Anim Virtual Worlds 30:e1883. https://doi.org/10.1002/cav.1883

Kuan G, Morris T, Kueh YC, Terry PC (2018) Effects of relaxing and arousing music during imagery training on dart-throwing performance. Physiol Arousal Indices Compet State Anxiety Front Psychol. https://doi.org/10.3389/fpsyg.2018.00014

Lagos O (2019) Knuckles oculus quest and Rift S grip. Thingiverse. https://www.thingiverse.com/thing:3652161. 2020

Larmuseau C, Cornelis J, Lancieri L, Desmet P, Depaepe F (2020) Multimodal learning analytics to investigate cognitive load during online problem solving. Br J Edu Technol 51:1548–1562. https://doi.org/10.1111/bjet.12958

Lehikko A (2021) Measuring self-efficacy in immersive virtual learning environments: a systematic literature review. J Interact Learn Res 32:125–146

Levac DE, Huber ME, Sternad D (2019) Learning and transfer of complex motor skills in virtual reality: a perspective review. J Neuroeng Rehabil 16:121. https://doi.org/10.1186/s12984-019-0587-8

Liebold B, Brill M, Pietschmann D, Schwab F, Ohler P (2017) Continuous measurement of breaks in presence: psychophysiology and orienting responses. Media Psychol 20:477–501. https://doi.org/10.1080/15213269.2016.1206829

Luvizutto G, Bruno A, Oliveira S, Silva M, Souza L (2022) Development and application of an electrical buzz wire to evaluate eye-hand coordination and object control skill in children: a feasibility study. Hum Mov 23:138–144

Mackay C, Cox T, Burrows G, Lazzerini T (1978) An inventory for the measurement of self-reported stress and arousal. Br J Soc Clin Psychol 17:283–284. https://doi.org/10.1111/j.2044-8260.1978.tb00280.x

Magill R, Anderson D (2016) Motor learning and control. McGraw-Hill Publishing, New York

Makowski D et al (2021) NeuroKit2: a python toolbox for neurophysiological signal processing. Behav Res Methods 53:1689–1696. https://doi.org/10.3758/s13428-020-01516-y

Makransky G, Lilleholt L, Aaby A (2017) Development and validation of the multimodal presence scale for virtual reality environments: a confirmatory factor analysis and item response theory approach. Comput Hum Behav 72:276–285. https://doi.org/10.1016/j.chb.2017.02.066

Makransky G, Borre-Gude S, Mayer RE (2019) Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. J Comput Assist Learn 35:691–707

Makransky G, Petersen GB (2021) The cognitive affective model of immersive learning (CAMIL): a theoretical research-based model of learning in immersive virtual reality. Educ Psychol Rev 1–22

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18(50–60):11

Marín-Morales J et al (2018) Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. Sci Rep 8:13657. https://doi.org/10.1038/s41598-018-32063-4

Matthews G, Margetts I (1991) Self-Report arousal and divided attention: A study of performance operating characteristics. Hum Perform 4:107–125. https://doi.org/10.1207/s15327043hup0402_2

Merchant Z, Goetz ET, Cifuentes L, Keeney-Kennicutt W, Davis TJ (2014) Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: a meta-analysis. Comput Educ 70:29–40. https://doi.org/10.1016/j.compedu.2013.07.033

Mikropoulos TA, Natsis A (2011) Educational virtual environments: a ten-year review of empirical research (1999–2009). Comput Educ 56:769–780. https://doi.org/10.1016/j.compedu.2010.10.020

Moses ZB, Luecken LJ, Eason JC (2007) Measuring task-related changes in heart rate variability. In: 2007 29th annual international conference of the IEEE engineering in medicine and biology society. pp 644–647. doi:https://doi.org/10.1109/IEMBS.2007.4352372

Movahedi A, Sheikh M, Bagherzadeh F, Hemayattalab R, Ashayeri H (2007) A practice-specificity-based model of arousal for achieving peak performance. J Mot Behav 39:457–462. https://doi.org/10.3200/JMBR.39.6.457-462

Mullen G, Davidenko N (2021) Time compression in virtual reality timing & time. Perception 9:377–392. https://doi.org/10.1163/22134468-bja10034

Muñoz JE, Pope AT, Velez LE (2019) Integrating biocybernetic adaptation in virtual reality training concentration and calmness in target shooting. In: Holzinger A, Pope A, da Plácido SH (eds) Physiological computing systems, vol 10057. Springer, Cham, pp 218–237

Murcia-Lopez M, Steed A (2018) A comparison of virtual and physical training transfer of bimanual assembly tasks. IEEE Trans vis Comput Graph 24:1574–1583. https://doi.org/10.1109/TVCG.2018.2793638

Orsila R et al (2008) Perceived mental stress and reactions in heart rate variability—a pilot study among employees of an electronics company. Int J Occup Saf Ergon 14:275–283

Owens ME, Beidel DC (2015) Can virtual reality effectively elicit distress associated with social anxiety disorder? J Psychopathol Behav Assess 37:296–305. https://doi.org/10.1007/s10862-014-9454-x

Pakarinen T, Pietilä J, Nieminen H (2019) Prediction of self-perceived stress and arousal based on electrodermal activity*. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), 23–27 July 2019. pp 2191–2195. doi:https://doi.org/10.1109/EMBC.2019.8857621

Parong J, Mayer RE (2021) Cognitive and affective processes for learning science in immersive virtual reality. J Comput Assist Learn 37:226–241. https://doi.org/10.1111/jcal.12482

Patle DS, Manca D, Nazir S, Sharma S (2019) Operator training simulators in virtual reality environment for process operators:

a review. Virtual Real 23:293–311. https://doi.org/10.1007/s10055-018-0354-3

Pavlidis I, Zavlin D, Khatri AR, Wesley A, Panagopoulos G, Echo A (2019) Absence of stressful conditions accelerates dexterous skill acquisition in surgery. Sci Rep 9:1747. https://doi.org/10.1038/s41598-019-38727-z

Pijeira-Díaz HJ, Drachsler H, Kirschner PA, Järvelä S (2018) Profiling sympathetic arousal in a physics course: How active are students? J Comput Assist Learn 34:397–408. https://doi.org/10.1111/jcal.12271

Pintrich PR (1991) A manual for the use of the motivated strategies for learning questionnaire (MSLQ). https://eric.ed.gov/?id=ED338122

Posner J, Russell JA, Peterson BS (2005) The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol 17:715–734

Potter RF, Bolls P (2012) Psychophysiological measurement and meaning: cognitive and emotional processing of media. Routledge, London

Prabhu A, Smith W, Yurko Y, Acker C, Stefanidis D (2010) Increased stress levels may explain the incomplete transfer of simulator-acquired skill to the operating room. Surgery 147:640–645. https://doi.org/10.1016/j.surg.2010.01.007

Pulijala Y, Ma M, Pears M, Peebles D, Ayoub A (2018) Effectiveness of immersive virtual reality in surgical training—a randomized control trial. J Oral Maxillofac Surgery 76:1065–1072. https://doi.org/10.1016/j.joms.2017.10.002

Quick JA, Bukoski AD, Doty J, Bennett BJ, Crane M, Barnes SL (2017) Objective measurement of clinical competency in surgical education using electrodermal activity. J Surg Educ 74:674–680

Radhakrishnan U, Blindu A, Chinello F, Koumaditis K (2021a) Investigating motor skill training and user arousal levels in VR:pilot study and observations. In: 2021 IEEE conference on virtual reality and 3d user interfaces abstracts and workshops(VRW), pp 625–626. https://doi.org/10.1109/VRW52623.2021.00195

Radhakrishnan U, Koumaditis K, Chinello F (2021b) A systematic review of immersive virtual reality for industrial skills training. Behav Inf Technol 40:1310–1339

Radianti J, Majchrzak TA, Fromm J, Wohlgenannt I (2020) A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. Comput Educ 147:103778

Rangarajan K, Davis H, Pucher PH (2020) Systematic review of virtual haptics in surgical simulation: A valid educational tool? J Surg Educ 77:337–347. https://doi.org/10.1016/j.jsurg.2019.09.006

Read JC, Begum SF, McDonald A, Trowbridge J (2013) The binocular advantage in visuomotor tasks involving tools. i-Perception 4:101–110

Rubin DC, Talarico JM (2009) A comparison of dimensional models of emotion: evidence from emotions, prototypical events, autobiographical memories, and words. Memory 17:802–808. https://doi.org/10.1080/09658210903130764

Sakowitz SM, Inglehart MR, Ramaswamy V, Edwards S, Shoukri B, Sachs S, Kim-Berman H (2019) A comparison of two-dimensional prediction tracing and a virtual reality patient methods for diagnosis and treatment planning of orthognathic cases in dental students: a randomized preliminary study. Virtual Real. https://doi.org/10.1007/s10055-019-00413-w

Schachinger H, Blumenthal TD, Richter S, Savaskan E, Wirz-Justice A, Kräuchi K (2008) Melatonin reduces arousal and startle responsiveness without influencing startle habituation or affective startle modulation in young women. Horm Behav 54:258–262. https://doi.org/10.1016/j.yhbeh.2008.03.013

Schwarz S, Regal G, Kempf M, Schatz R (2020) Learning success in immersive virtual reality training environments: practical evidence from automotive assembly. In: Proceedings of the 11th nordic conference on human-computer interaction: shaping experiences, shaping society. pp 1–11

Shaffer F, Ginsberg JP (2017) An overview of heart rate variability metrics and norms. Front Public Health. https://doi.org/10.3389/fpubh.2017.00258

Shafti A, Lazpita BU, Elhage O, Wurdemann HA, Althoefer K (2016) Analysis of comfort and ergonomics for clinical work environments. In: 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE. pp 1894–1897

Shakur SF et al (2015) Usefulness of a virtual reality percutaneous trigeminal rhizotomy simulator in neurosurgical training operative. Neurosurgery 11:420–425. https://doi.org/10.1227/NEU.0000000000000853

Shu Y, Huang Y-Z, Chang S-H, Chen M-Y (2019) Do virtual reality head-mounted displays make a difference? A comparison of presence and self-efficacy between head-mounted displays and desktop computer-facilitated virtual environments. Virtual Real 23:437–446. https://doi.org/10.1007/s10055-018-0376-x

Sk R, Mallam SC, Nazir S (2021) Effectiveness of VR head mounted displays in professional training: a systematic review technology. Knowl Learn 26:999–1041. https://doi.org/10.1007/s10758-020-09489-9

Slater M et al (2006) Analysis of physiological responses to a social situation in an immersive virtual environment. Presence 15:553–569

Sternad D (2018) It's not (only) the mean that matters: variability, noise and exploration in skill learning. Curr Opin Behav Sci 20:183–195. https://doi.org/10.1016/j.cobeha.2018.01.004

Storbeck J, Clore GL (2008) Affective arousal as information: how affective arousal influences judgments. Learn Mem Soc Pers Psychol Compass 2:1824–1843. https://doi.org/10.1111/j.1751-9004.2008.00138.x

Syrjämäki AH, Isokoski P, Surakka V, Pasanen TP, Hietanen JK (2020) Eye contact in virtual reality – a psychophysiological study. Comput Hum Beh 112:106454. https://doi.org/10.1016/j.chb.2020.106454

Tai K-H, Hong J-C, Tsai C-R, Lin C-Z, Hung Y-H (2022) Virtual reality for car-detailing skill development: learning outcomes of procedural accuracy and performance quality predicted by VR self-efficacy, VR using anxiety, VR learning interest and flow experience. Comput Educ 182:104458. https://doi.org/10.1016/j.compedu.2022.104458

Terkildsen T, Makransky G (2019) Measuring presence in video games: an investigation of the potential use of physiological measures as indicators of presence. Int J Hum Comput Stud 126:64–80. https://doi.org/10.1016/j.ijhcs.2019.02.006

Tian F, Hua M, Zhang W, Li Y, Yang X (2021) Emotional arousal in 2D versus 3D virtual reality environments. PLoS ONE 16:e0256211. https://doi.org/10.1371/journal.pone.0256211

Ünal AB, de Waard D, Epstude K, Steg L (2013) Driving with music: effects on arousal and performance. Transp Res F Traffic Psychol Behav 21:52–65. https://doi.org/10.1016/j.trf.2013.09.004

Usoh M, Catena E, Arman S, Slater M (2000) Using Presence questionnaires in reality. Presence Teleoper Virtual Environ 9:497–503. https://doi.org/10.1162/105474600566989

Vallat R (2018) Pingouin: statistics in Python. J Open Source Softw 3:1026

van Dooren M, Janssen JH (2012) Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. Physiol Behav 106:298–304

Van Merriënboer JJG, Sweller J (2010) Cognitive load theory in health professional education: Design principles and strategies. Med Educ 44:85–93. https://doi.org/10.1111/j.1365-2923.2009.03498.x

Ventura S, Cebolla A, Latorre J, Escrivá-Martínez T, Llorens R, Baños R (2021) The benchmark framework and exploratory study to investigate the feasibility of 360-degree video-based virtual reality to induce a full body illusion. Virtual Real. https://doi.org/10.1007/s10055-021-00567-6

Virtanen P et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2

Vrchewal (2020) Measurements. Unity. https://assetstore.unity.com/packages/tools/utilities/measurements-111690. 2021

Wang C-A, Baird T, Huang J, Coutinho JD, Brien DC, Munoz DP (2018) Arousal effects on pupil size, heart rate, and skin conductance in an emotional face task. Front Neurol. https://doi.org/10.3389/fneur.2018.01029

Waskom ML (2021) Seaborn: statistical data visualization. J Open Source Softw 6:3021

Wilcoxon F (1945) Individual comparisons by ranking methods. Biom Bull 1:80–83. https://doi.org/10.2307/3001968

Winther F, Ravindran L, Svendsen KP, Feuchtner T (2020) Design and evaluation of a VR training simulation for pump maintenance based on a use case at grundfos. In: 2020 IEEE conference on virtual reality and 3D user interfaces (VR), 3/2020. IEEE, Atlanta, GA, USA, pp 738–746. doi:https://doi.org/10.1109/VR46266.2020.00097

Witmer BG, Singer MJ (1998) Measuring presence in virtual environments: a presence questionnaire. Presence Teleop Virt 7:225–240. https://doi.org/10.1162/105474698565686

Wu D, Courtney CG, Lance BJ, Narayanan SS, Dawson ME, Oie KS, Parsons TD (2010) Optimal arousal identification and classification for affective computing using physiological signals: virtual reality stroop task. IEEE Trans Affect Comput 1:109–118. https://doi.org/10.1109/T-AFFC.2010.12

Wulf G, Shea C, Lewthwaite R (2010) Motor skill learning and performance: a review of influential factors. Med Educ 44:75–84

Wulfert E, Roland BD, Hartley J, Wang N, Franco C (2005) Heart rate arousal and excitement in gambling: winners versus losers. Psychol Addict Behav 19:311

Xie B et al (2021) A review on virtual reality skill training applications. Front Virtual Real 2:49

Yerkes RM, Dodson JD (1908) The relation of strength of stimulus to rapidity of habit-formation. J Comp Neurol Psychol 18:459–482. https://doi.org/10.1002/cne.920180503

Yin J, Arfaei N, MacNaughton P, Catalano PJ, Allen JG, Spengler JD (2019) Effects of biophilic interventions in office on stress reaction and cognitive function: a randomized crossover study in virtual reality. Indoor Air 29:1028–1039. https://doi.org/10.1111/ina.12593

Zahabi M, Abdul Razak AM (2020) Adaptive virtual reality-based training: a systematic literature review and framework. Virtual Real. https://doi.org/10.1007/s10055-020-00434-w

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.