IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 August 2023, accepted 4 October 2023, date of publication 13 October 2023, date of current version 15 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3324554

RESEARCH ARTICLE

Influencer-Based Virtual Reality Stroop Task

EDOARDO BATTEGAZZORRE[1], FRANCESCO STRADA[1], LUCIA DE FRANCESCO[2], ALESSANDRO MAZZA[2], OLGA DAL MONTE[2], AND ANDREA BOTTINO[1], (Member, IEEE)

[1]DAUIN, Politecnico di Torino, 10129 Turin, Italy
[2]Department of Psychology, University of Turin, 10124 Turin, Italy

Corresponding author: Francesco Strada (francesco.strada@polito.it)

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART").

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Bioethical Committee of the University of Turin and performed in line with the Declaration of Helsinki (World Medical Association; 2013).

ABSTRACT Computerized Stroop tasks have proven to be valuable tools in experimental and clinical psychology, providing reliable and reproducible assessments of cognitive processes. This paper presents the Influencer-Based Virtual Reality Stroop Test (IB-VRST), an innovative application that allows users to perform an immersive Stroop test while exposed to various virtual influencers. These influencers are categorized as task-related and non-task-related distractors, and elements of social presence implemented through competitive and collaborative virtual avatars. To validate this application and analyze the effects of these influencers on performance and stress levels, we conducted two experiments where we collected quantitative data using application logs and biometric sensors, and qualitative data using pre- and post-experiment questionnaires and self-reported stress ratings. Results show that IB-VRST successfully creates an immersive and intuitive experience that prompts the Stroop effect. While distractors generally impair performance and increase stress levels, social presence elements generally improve performance and reduce stress, except when participants compete against a more skilled opponent. This study adds to the existing literature by providing a comprehensive examination of the effects of virtual influencing factors on immersive Stroop tasks, thereby supporting the development of more engaging, immersive, and effective virtual Stroop tests.

INDEX TERMS Social presence, stroop test, virtual reality, user experience.

I. INTRODUCTION

The Stroop interference effect, named after John Ridley Stroop, who first published a study on this phenomenon in 1935 [1], describes the discrepancy in reaction times between congruent and incongruent stimuli. The manifestation of this effect is leveraged by the Stroop test, where colored words are typically presented in incongruent ink (e.g., "red" in green ink) and participants are asked to name the ink color rather than the word, or vice versa. The Stroop test is used in both experimental and clinical psychology as it has been shown to be a reliable tool to discriminate between healthy and unhealthy populations affected by various mental deficits or disorders, such as traumatic brain injury [2], [3], attention deficit, hyperactivity disorder [4], and autistic spectrum disorders [5]. In addition, the Stroop test is a proven tool to elicit cognitive stress and increase physiological reactivity [6].

Regardless of its usage, extensive research has been dedicated to examining the impact of external factors on the Stroop test. These external factors can include various types of *distractors* and different forms of *social presence*. In what follows, we will group these two concepts under the umbrella term *influencers* (i.e., elements that have the power to influence or shape the scenario in which the Stroop task takes place and modify the subject's performance).

Distractors are events and occurrences introduced to divert the user's attention and induce stress while performing the

The associate editor coordinating the review of this manuscript and approving it for publication was Tai-Hoon Kim.

assigned task. Two types of distractors can be considered: (i) *task-related*, distractors that directly contribute to the Stroop interference effect (e.g., auditory stimuli for color words [7], [8]) or (ii) *non-task-related*, distractors that are not directly related to the color naming operation such as a passing car [9], [10], [11] or a ringing phone [11].

Social presence is the sense of ''being with another'' or ''being in a social situation''. Social interactions can affect attention and performance on executive tasks and can be either a distractor or facilitator, depending on the subject and task. In the Stroop test, social presence can be introduced in several forms: performing the test in the presence of an observer [12], [13], collaborating with a teammate [14], [15], or competing with an opponent [16], [17], [18], [19]. Studying the effects of social presence goes beyond analyzing its impact on Stroop performance. It provides valuable insights into the dynamics of social relationships and communication with practical implications for a variety of domains. For example, this knowledge can be used to improve collaboration in education or to create more engaging experiences in competitive games or sports activities [20].

The traditional Stroop test is administered in paper form, with the subject responding verbally. However, recent technological advances allowed the introduction of computerized versions of the Stroop test, which offer several advantages over the traditional method, including faster administration, complete control over stimulus presentation, and automatic calculation of results while maintaining reliability and validity. In addition, the computerized Stroop test could be used for remote patient ''self-assessment'' [21].

Among all computer-based methods, Virtual Reality (VR), is certainly the most advantageous option for conducting the Stroop test.[1] First, the stimuli can be presented in a more realistic and engaging way, which can increase the subjects' motivation and concentration [23]. Second, it offers greater control over the setting of the test environment, which in turn allows for greater ecological validity, leading to greater reliability of the results [9], [24]. Finally, Virtual Reality Stroop tests (VRSTs) can be used to create situations that are difficult to reproduce accurately and consistently in a laboratory setting [11].

Despite the literature shows that VRSTs are able to effectively replicate the traditional paper-and-pencil version of the test [10], the current research on VRSTs still has some limitations. First, few efforts have been made to take advantage of the intuitive interaction metaphors offered by immersive VR which can improve immersion and presence. Second, despite the relevance of these elements in any VR application, there is still a lack of thorough evaluation of the impact of usability, user experience, immersion, and presence of current VRSTs. Third, there are no experiments comparing the relative impact of environmental distractors and social presence on performance and induced stress levels. Finally, while numerous studies have been conducted on the effects of social presence in traditional or computer-based Stroop or Stroop-like tasks [14], [16], [17], [18], [19], [25], [26], no study used immersive VR and virtual avatars to analyze the effects of cooperation and competition.

To address these issues, in this paper, we present the influencer-based VR Stroop test (IB-VRST), which aims to compare the effects of different influencers while experiencing an immersive VRST leveraging an HMD and intuitive interaction metaphors. To validate the proposed system, we conducted experiments with 102 volunteers and collected quantitative (psychophysiological stress and performance metrics) and qualitative data (self-reported stress levels and usability, immersion, and presence rating questionnaires to gain insight into participants' subjective experiences in the VR environment). In particular, our experiments seek to answer the following research questions:

- **RQ1:** How do immersive VR affect user engagement, immersion, and presence in VRSTs?
- **RQ2:** What is the impact of different types of distractors (task-related vs. non-task-related) on user performance and induced stress in VRSTs?
- **RQ3:** What is the impact of social presence (competing with an opponent vs. collaborating with a teammate) on user performance and induced stress in VRSTs?
- **RQ4:** Which is the differential impact of various types of influencers (i.e., distractors and social presence), in terms of user performance and induced stress?

The experimental results show that the proposed VRST is effective in creating an immersive and user-friendly experience that successfully elicits the Stroop effect. Regarding the different influencers investigated, the presence of distractors generally had a negative impact on performance and increased stress levels. On the other hand, social presence had always a positive influence on performance and had a stress-reducing effect when collaborating with a teammate or competing against a less skilled opponent. However, when competing against a more skilled opponent, better performance was associated with higher stress levels.

In conclusion, the main contributions of this work can be summarized as follows:

- we are the first to conduct a thorough investigation of the usability, user experience, and sense of immersion and presence of a VRST;
- we present a comparison of the effects of distractors (either task-related or not) and elements of social presence on user performance and stress levels with immersive VR technologies;
- we introduce the first immersive VRST application which leverages virtual avatars to investigate the effects of social presence;

[1] In the following, according to [22], we will refer to non-immersive VR when the simulation is viewed on a desktop monitor, and immersive VR when either head-mounted displays (HMDs) or other immersive technologies, such as CAVEs, are used.

- we compare for the first time the psychophysiological and performance effects of collaborative and competitive social influencers on Stroop task performance.

II. RELATED WORKS

This section starts with an overview of the current research advances in VRST and the use of distractors in these environments. Then, it discusses research on social presence influencers, focusing on examples taking into account competition and cooperation (and, thus, excluding settings where subjects are completing the task in front of an external observer [12], [13]).

As a note, most computerized versions of the Stroop task reviewed (including VRSTs) are based on the Color Word Interference test included in the Delis Kaplan Executive Function System [27], which we follow in our work. The main measures used by researchers to gain a comprehensive understanding of the cognitive processes involved in a Stroop task are performance (e.g., the number of correct and timely answers, response time), the Stroop interference effect (i.e., the difference in response times between incongruent and congruent stimuli), and psychophysiological stress levels.

A. DISTRACTORS IN IMMERSIVE VRSTS

As discussed by Parsons [11], the use of realistic virtual environments (VEs) and ''diegetic'' distractors (i.e., distractors that are fully and believably integrated into the VE) can help overcome the shortcomings of traditional implementations of the Stroop task while providing the same level of control as a laboratory environment.

The Virtual Classroom [9], the Stroop Apartment [24], and the simulation presented in [28] are some notable immersive VRSTs that incorporate multiple environmental distractors divided by sensory channel into auditory, visual, and mixed audiovisual. Some of these VRSTs have been extended in subsequent studies [3], [10], [11], [29], [30], [31] and have also been used to assess autism spectrum disorders [5]. However, in all these studies interaction relies on standard desktop metaphors and peripherals (e.g., mouse click in [9], [24], a colored keyboard in [28] or buttons on a controller [31]).

The first immersive VRTS leveraging fully-fledged 3D interactions, is the Stroop Room [32], which provides a VE consisting of a hexagonal room where each walls has a different color. During the task, the correct answer (i.e., the correctly colored wall) must be selected using raycasting via the HMD controller. The simulation includes as distractors shrinking walls and moving colored surfaces. Experimental results show that the combination of an immersive environment and environmental distractors significantly increases subjects' psychophysiological stress compared to previous approaches (both 2D and VR, including the aforementioned [11], [30], [33]).

It is worth noting that while the literature on VRSTs has made significant progress in terms of task design, implementation, and stress induction, there is a notable gap in the assessment of usability, user experience, and sense of immersion/presence. We believe that these aspects play a crucial role in helping researchers gain a more comprehensive understanding of the strengths and limitations of these systems, and inform the development of future VRSTs to improve their effectiveness and acceptance.

B. SOCIAL PRESENCE: COMPETITION AND COOPERATION

Several studies have explored the effects of competitive and collaborative social presence influencers on task performance and induced stress in Stroop tasks. However, while there is a considerable body of research on the effects of competition, cooperation has been realtively less investigated. Mackinnon et al. [16] conducted one of the first studies in which competition was introduced into a traditional Stroop task. Participants performed the task alone and with a competitor, although the competition was asynchronous (i.e., participants completed the test individually and the winner was announced at the end of the session). The study found that the competition condition reduced the Stroop effect and response times by up to 25%, indicating improved performance. However, participants also reported higher mental effort in the competition condition.

In [17], researchers examined the effects of face-to-face competition in a computerized Stroop task. They found that the presence of a physical competitor in the same room resulted in faster response times and a greater reduction in the Stroop effect compared to [16]. The presence of a slower competitor led to a reduction in the Stroop effect of up to 65%, with an average reduction of 52% across all conditions.

This study was further developed in [18] to examine the effects of social comparison on the Stroop effect. In two separate studies, the researchers found that the presence of social comparison led to a remarkable reduction in the Stroop effect by over 70%. Participants also showed faster response times when competing against better opponents, with no significant effect on error rates. These results highlight the significant influence of social factors in modulating information processing.

The study presented in [19] extended the conventional assessment of performance in the Stroop task by analyzing biometric data (heart rate values, heart rate variability, and saliva samples) to examine the impact of competition on stress responses. Participants competed against consistently higher-performing opponents while receiving feedback on their peers' performance. Results showed that the introduction of the social element in this study created a more challenging and stressful environment than the control group, which performed a standard Stroop test.

While previous research has focused primarily on competition as a social factor, with limited examination of the effects of collaboration, [14] presented a study in which participants worked together in one room on a collaborative version of the Stroop task. Results showed minimal influence of social presence and task sharing on Stroop interference and response times. In contrast, [15] examined collaborative

FIGURE 1. An overview of the Virtual Environment where the Stroop task is administered.

FIGURE 2. A screen capture from the IB-VRST as seen from the user's point of view, in Experiment 1. In the image we can see several environmental distractors, such as open windows and several frames on the wall and the plant which fell on the floor.

Stroop-like tasks in which participants worked with a simulated teammate in a different location. In this scenario, social interaction and task sharing significantly reduced Stroop interference compared to a control group. The comparison of the latter two experimental results suggests that the effects of social collaboration on Stroop interference and performance may vary depending on the specific conditions and task requirements. In particular, the physical presence and degree of realism in the collaborative environment appear to influence the impact of the social influencers.

Previous studies examining the effects of social presence on Stroop-like tasks have mainly used traditional paper-and-pencil versions [16] or desktop-based implementations of the Stroop test [14], [15], [17], [18], [19]. Notably, no study used immersive VR and virtual agents. Immersive VR offers a more realistic and engaging environment that can influence participants' sense of presence and cognitive processing while using virtual agents as avatars can ensure consistent and controlled behavior. Thererfore, investigating cooperation and competition under these settings can provide valuable insights into how social factors influence Stroop interference and performance in a more ecologically valid setting.It is also worth noting that the literature lacks a direct comparison between the effects of competition and cooperation. Addressing these gaps would contribute to a more comprehensive understanding of the role of social influencers in Stroop performance and offer insights for future research and practical applications.

III. METHODS

This section provides a detailed description of the virtual environment, simulation system, and technical aspects of the proposed IB-VRST.

A. VIRTUAL ENVIRONMENT

The VE used in IB-VRST presented a room with various furniture, two windows overlooking a courtyard (Fig. 1), and an armchair in front of a widescreen TV, on which instructions and Stroop stimuli are displayed during the session (Fig. 2). Participants wore an HMD and interacted with the VE using the HMD controllers. The Stroop test of IB-VRST is based on the Color-Word Interference Test defined in [27]. This included the Color Naming condition, in which participants had to name the *ink* color of the written word, and the Word Reading condition, in which they had to name the word *meaning*.

To complete the Stroop task, participants had to press the correct button on the console in front of them. The console presented six buttons that corresponded to the color-word stimuli displayed on the TV screen. The stimuli were displayed in a large font, accompanied by a label indicating whether the user should name the *meaning* or the *ink* of the word (Fig. 3). An on-screen timer indicated the time remaining to respond, and a ticking sound marked the passing of time.

There were three possible outcomes for each trial. A *correct* answer was indicated by a green smiley face and a ''success'' sound. A *wrong* answer resulted in a red frowny face and a buzzer sound. If the timer expired without a response, the result was marked as *timer expired* with an appropriate message on the screen and a timer alert sound.

To enhance the sense of presence in the VE, participants were represented by an avatar whose gender corresponded to that of the user. The upper body movements of the virtual avatar were animated using inverse kinematics, computed from the user's head and hand movements obtained from the HMD and controller tracking system. The use of only these two tracking devices ensured maximum freedom of movement for the users but could not provide an exact match of the user's upper body posture (i.e., the position and orientation of the elbows was only an estimate resulting from the inverse kinematics calculations). However, since the participants in the experiment remained seated throughout the session and the range of motion required to interact with the virtual console was limited, positional errors in the virtual representation of the user's body were minimal.

As a final note, using hand tracking instead of HMD controllers could have potentially provided a more natural interaction experience. However, our initial tests using the HMD's cameras for tracking users' hands revealed issues with accuracy due to the frequent head movements required to constantly switch attention between the TV screen and the console, thereby negatively impacting usability. The option of using smart gloves or external bare hand tracking devices was considered in the design phase but ultimately discarded

FIGURE 3. A screen capture from the IB-VRST as seen from the user's point of view in Experiment 2. In the image we can see the button console, the user's virtual hand about to press a button, as well as the relative placement of the teammate in the room. In this example, the screen displays the PURPLE word in green color. Given the INK COLOR prompt under the word, the user would need to press the green button on the console.

to obtain a simpler hardware setup. Given the tradeoffs between natural interaction and practical implementation, we decided to prioritize the user experience and minimize potential technical issues by relying on HMD controllers for interaction.

As for the technical details, IB-VRST was implemented using the Unity game engine.[2] The Oculus Rift S was selected as HMD due to its user-friendly setup (as it features inside-out optical tracking that does not require external sensors), its ergonomic design, comfortable fit, and high-resolution graphics. Nevertheless, the implementation was based on the OpenVR SDK, which allows the applications to be used on most commercially available devices without code changes.

B. INFLUENCERS

As explained in the Introduction, this study investigated the effects of two different types of *influencers* on performance and stress levels in the Stroop task: *distractors* and *social presence*, each of which has several subcategories. There are two types of *distractors*, *task-related non-task-related*, while social presence was represented by avatars that acted as the user's teammates or opponents. The details are as follows.

Task-related distractors, i.e., a strong colored lights behind the TV screen and incongruent verbal cues spoken by a voice coming from the tv screen, were designed to directly influence Stroop interference. These distractors aimed to create cognitive challenges and increase demands on attention and inhibition of irrelevant information or responses during the task.

On the other hand, non-task-related distractors encompassed a range of sensory stimuli, including *auditory* distractors (the shattering of a light bulb, ominous booming music, the sound of a car braking and crashing, loud shouting from outside, and a piece of furniture hitting the floor behind the user), *visual* distractors (a mug and a notebook being thrown into the air from behind the user and the light in the room suddenly dimming), and *audiovisual* distractors (a slamming door, pictures falling off the wall, a flowerpot falling on the floor, the TV screen suddenly cracking, a coat hanger tipping over, a lamp holder falling off the desk, and the back wall of the room collapsing dramatically). These distractors were not directly related to the Stroop task but were intended to simulate real events in the environment and to induce additional cognitive load and distraction. Visual examples of these distractors are depicted in Fig. 2.

Social presence influencers aimed to investigate the effects of social interaction on the Stroop task. Social interactions were simulated by virtual avatars that acted as either the user's opponent or teammate when completing the Stroop test. During the virtual experience, only one avatar type was present in the room and positioned in the user's field of view (Fig. 3) so that its actions were clearly visible. The avatars, hereafter referred to as Non-Playable Characters (NPCs), had different skill levels controlled by the application, and their gender corresponded to that of the user.

IV. EXPERIMENTAL PROTOCOLS

The experimental protocol aims to validate several research hypotheses related to the immersive design and usability of IB-VRST, its effectiveness in inducing cognitive stress, and the differential effects of distractors and social presence on user performance and stress levels. In particular, our experiments seek to answer the following research questions:

- **RQ1:** How do immersive VR affect user engagement, immersion, and presence in VRSTs?
- **RQ2:** What is the impact of different types of distractors (*task-related* vs. *non-task-related*) on user performance and induced stress in VRSTs?
- **RQ3:** What is the impact of social presence (competing with an opponent vs. collaborating with a teammate) on user performance and induced stress in VRSTs?
- **RQ4:** Which is the differential impact of various types of influencers (i.e., distractors and social presence), in terms of user performance and induced stress?

To answer these RQs, we conducted two experiments whose combined results guided the answers to RQ1 and RQ4, while RQ2 and RQ3 were answered thanks to the results of Experiments 1 and 2, respectively.

Experiment 1 examined the effects of different types of distractors (*task-related* vs. *non-task-related*) while users performed the Stroop task in two modes: the *Normal* mode, in which only the Color Naming condition of the Stroop test was used, and the *Hard* mode, in which participants switched from Color Naming to Word Reading every ten trials. The Stroop tasks were then divided into three rounds, one for each distractor type and one without distractors. All three rounds were completed in both *Normal* and *Hard* modes.

The goal of Experiment 2 was to examine the effects of social presence on users' cognitive engagement, task performance, and stress levels during Stroop tasks. To achieve

[2]The executable is available at the following public repository: https://github.com/CGVGroup/IB-VRST

this, NPCs were introduced as opponents or teammates in the tasks with predefined skill levels.

To gather and validate data for our study, we adopted a mixed method approach. Qualitative data were collected through user feedback, including self-reported stress levels on a scale of 0 to 100 and ratings of usability, engagement, and enjoyment with questionnaires such as the System Usability Scale Questionnaire (SUS) [34], [35] the Simulator Sickness Questionnaire (SSQ) [36], the Flow State Scale Questionnaire (FSS) [37], and an adapted version of the VRUSE questionnaire focusing the following usability factors: User Input, Simulation Fidelity, and Immersion and Presence [38]. Prior to each experiment, participants completed a pre-test questionnaire to collect demographic information and previous familiarity with VR, video games, and the Stroop test.

Quantitative data included participants' physiological measures and their interactions within the VR simulation. Physiological data were collected using an MP160 biosignal amplifier from Biopac Systems, Inc. ECG recordings were obtained by applying pre-gelled shielded electrodes in a Lead II montage with standard limb electrode placement [39]. The signal was sampled at 500 Hz, with a gain parameter set at 1000 ($\pm$ 10 mV), and filtered using a 150 Hz low pass and a 0.05 Hz high pass filter. In addition, a detailed log of the VR simulation state was recorded throughout the experiments. This included user interactions, correct and incorrect responses with reaction times, environmental events related to activated influencers, and the progression of the different phases of the experiment.

In both experiments, participants experienced the IB-VRST while seated and wearing an HMD, using only their dominant hand for interactions, while the non-dominat hand rested on their knee. This setup allowed for reliable measurements of physiological data, as described below.

A. EXPERIMENTS INTRODUCTION

The experimental procedure of both experiments shared the same initial steps (upper grey boxes in Fig. 4 and 5). Subjects first signed an informed consent form and completed the pre-test questionnaire. An experimenter was present to assist with questions, explain the objectives of the study, and provide instructions for using the simulation. The biomonitoring device and sensors were then calibrated for accurate physiological measurements. Then users wore the HMD and adjusted the size of the virtual avatar to their own body size to enhance the sense of embodiment and presence. To collect baseline physiological measurements, users remained in the VE without interacting for 3 minutes.

Finally, the experimenter verbally described to users how to use the HMD controllers and virtual buttons, and then asked them to freely try out the button console (Figure 3). Once users confirmed that they were sufficiently familiar with the user interface, the experimenter described how users should respond to the stimuli presented on the screen. They

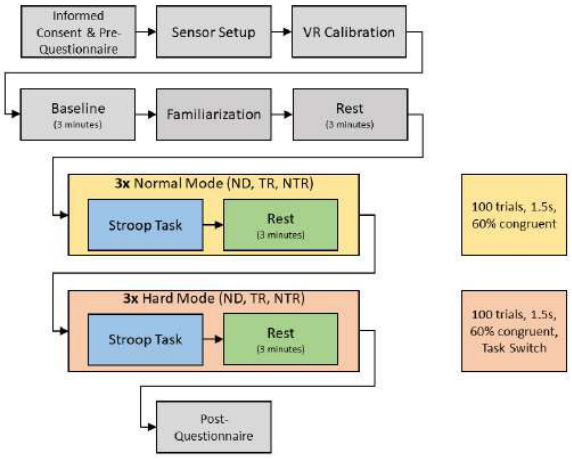


FIGURE 4. The experimental protocol for Experiment 1.

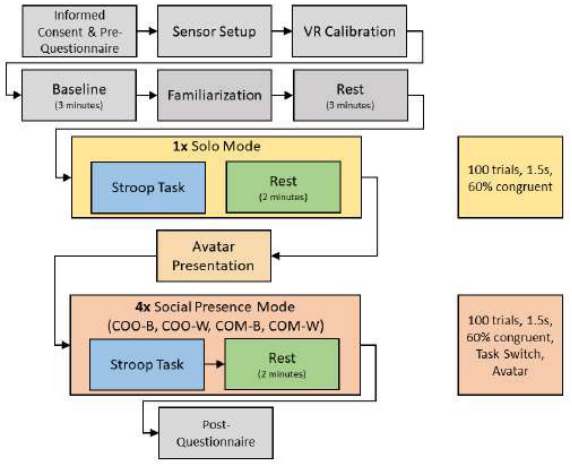


FIGURE 5. The experimental protocol for Experiment 2.

then performed a preliminary *Familiarization* mode in which they were asked to respond to a sequence of stimuli with consistent colors and inks. These initial steps conclude with a 3 minutes rest before starting the different conditions for each experiment, as detailed in the following. Both experimental procedures were approved by the Bioethical Committee of the University of Turin and conducted in accordance with the ethical standards of the 2013 Declaration of Helsinki.

B. EXPERIMENT 1

Fig. 4 illustrates the experimental procedure for Experiment 1. After completing the previously described initial steps (Section IV-A), participants engaged in the VRST in two modalities: *Normal* and *Hard*. The *Normal* mode consisted of 100 trials with a maximum response time of 1.5 seconds per stimulus and 60% congruent trials. Three different configurations of distractors were used: no distractors (ND), task-related distractors only (TR), and non-task-related distractors only (NTR). The order of these configurations was randomized to minimize sequence effects. The *Hard* mode had the same settings as the *Normal* mode, but introduced a task-switching (i.e., switching between Color Naming

and Word Reading conditions) every twelve trials, which increased the difficulty.

After each phase, participants rated their perceived stress level. This was followed by a three-minute rest period during which participants kept the HMD on and remained seated. Finally, participants completed the post-experience questionnaire to share their feedback and impressions. These steps were also followed in Experiment 2.

C. EXPERIMENT 2

In this experiment (Fig. 5), we followed the same initial steps as Experiment 1 (described in Section IV-A). After the familiarization phase, each user enters a *Solo* mode in which they perform the task without the presence of an NPC in the room. The completion of the *Solo* mode marks the beginning of the collaborative or competitive phase.

The social presence mode consists of four different configurations randomly assigned to subjects with equal probability: cooperation with a better-performing NPC (COO-B), cooperation with a worse-performing NPC (COO-W), competition against a better-performing NPC (COM-B), and competition against a worse-performing NPC (COM-W). In the competition phases, users try to beat their opponent's score, while in the cooperation phases, they work with their NPC teammates to beat the highest score of another team (simulated by the application).

The NPCs interact realistically with their virtual console using procedural animations, and their performance is adjusted according to the test condition, either performing better or worse than the user. The algorithm that controls the behavior of the NPCs ensures that their performance ultimately outperforms or lags behind that of the user, with responses appearing randomly distributed. This approach aims to create the impression of unpredictability in the performance of the NPCs and minimize the detectability of the differential effect by the participants.

Before starting each social presence mode, the NPC in the room (i.e., the worse or the better performing one according to the experimental condition) is briefly introduced to the user by displaying its picture and its *Solo* mode results on the TV screen. Then, it is physically introduced in the VE.

In each iteration of the social presence mode, a single stimulus is presented on the screen and each participant (the subject and the NPC) has equal time to respond (1.5 seconds). In the competitive mode, the individual participant score is a combination of the correct response and the response time. In collaborative mode, each teammate has an individual score, and the best of the two is compared to the opponents' best score to update the team score. In both configurations, the system proceeds to the next stimulus once both participants have entered their responses or the timer has expired. Throughout the phase, the subject's and NPC's scores are displayed at the bottom of the TV screen, which (in cooperative modes) also displays the opposing team's score. Since the NPC and the subject are in the same room, users can observe the NPC pressing buttons and hear the audio feedback from its console. The audio in the simulation is spatialized to ensure that the subject can distinguish between their own feedback and that of the NPC.

Upon completion of testing in all four configurations, subjects were given post-questionnaires to gather additional feedback.

D. EXPERIMENTAL SAMPLES

The two experiments were performed 6 months apart. For Experiment 1, we recruited volunteer students from Polytechnic of Turin and University of Turin, with a total of 52 participants. Of them, 32 were male, 19 female, and 1 nonbinary. Their average age was 24.9 years. Regarding technology awareness, only 12 participants had experience with immersive VR simulations, with 5 using it rarely, 3 occasionally, and 4 regularly. In addition, 9 participants had never played video games, 2 rarely, 7 occasionally, and 24 frequently. In terms of familiarity with the Stroop task, 32 participants had never taken the test before.

In Experiment 2, we recruited 50 participants from the same two universities. Of them, 27 were male, 22 were female, and 1 identified as nonbinary. The average age of the participants was 24.57 years. Similar to the first experiment, most participants had no prior experience with immersive VR technologies. Only 11 participants had never played video games, and 14 participants had never taken a Stroop test.

The experimental sessions for both experiments lasted approximately one hour, with 15-20 minutes for calibration and questionnaires and 40-45 minutes for the VR experience.

E. DATA ANALYSIS

A pairwise comparison was conducted for all questionnaire scales between the two experiments. Additionally, the internal consistency of the questionnaire scales was assessed using Cronbach alphas.

For the remaining collected metrics, the following analyses were performed for each experiment. In Experiment 1, a 2×2 ANOVA was conducted with distractor type (ND, TR, and NTR) and difficulty (*Normal* and *Hard* modes) as within-subject factors. In Experiment 2, One-Way repeated measures ANOVAs were initially performed with social condition (Solo, COO, and COM) as the within-subject factor. Subsequently, to examine the differences between various social influencers, 2×2 ANOVAs were conducted with social condition (COO and COM) and other's ability (B and W) as within-subject factors. Post-hoc tests were employed in all cases, with significance thresholds adjusted using Bonferroni correction. In the specific case of self-reported stress levels and heart rate, to present both metrics on the same unit of measure, we performed per-subject z-scoring, and then subtracted baseline values from these scores (we used *Familiarization* mode for Experiment 1, and *Solo* mode for Experiment 2 as baseline values). This correction was necessary to compensate inter-subject variability of both perceived stress and baseline heart rate. To compare the results between the two experiments, for each metric,

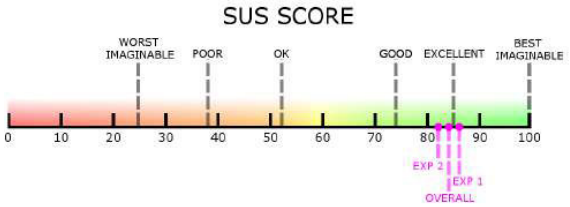


FIGURE 6. SUS scores for the IB-VRST. The subjects rated the usability with a score of 86.44 in Experiment 1 and 82.07 in Experiment 2 and an overall combined score of 84.56.

TABLE 1. VRUSE usability factors for both experiments. For each factor, we report the average score (Mean), the standard deviation (SD), and cronbach's alpha (α).

Usability Factor	Exp 1			Exp 2		
	Mean	SD	α	Mean	SD	α
Input	4.14	0.54	0.82	3.78	0.67	0.85
Fidelity	4.07	0.39	0.62	3.87	0.52	0.76
Immersion/Presence	4.03	0.59	0.80	3.71	0.66	0.80

we created two independent groups, one for each Experiment. In Experiment 1, the TR and NTR conditions of the Hard mode were combined, while in Experiment 2, all conditions except the Solo mode were included. Pairwise tests on unpaired data were then performed between these two groups.

All pairwise comparisons were conducted using appropriate statistical tests depending on the data characteristics. Specifically, Student's t-test was used for normal data with equal variance, Welch's t-test for normal data with unequal variance, and Mann-Whitney U rank test when the normality assumption was not met. The normality and variance equality assessments were performed using the Shapiro-Wilk test and F-test, respectively.

V. RESULTS

A. QUESTIONNAIRES

In the following, we will analyze usability and user experience questionnaires. We will highlight differences between the experiments when present.

1) SUS

The calculated Cronbach's alpha coefficient for the SUS questionnaire in Experiments 1 and 2 was respectively 0.7 and 0.84, indicating acceptable and good internal consistency. The IB-VRST received positive ratings for usability, as assessed by the SUS, with an average score after Experiment 1 of 86.44, which is above the threshold for "excellent" usability [34], [35]. In Experiment 2, the average score was 82.07, and the combined average score for all participants was 84.56, indicating a level between "good" and "excellent".

2) VRUSE

The results of the VRUSE Usability Factors are shown in Table 1. The calculated Cronbach's alpha coefficients in both experiments indicate a good level of internal consistency

TABLE 2. VRUSE diagnostic factors for both experiments. For each factor, we report the average score (Mean), and the standard deviation (SD). Factors with a * symbol indicate significant differences between the two experiments ($p < 0.01$).

Diagnostic Factor	Exp 1		Exp 2	
	Mean	SD	Mean	SD
Ease of Use*	4.18	0.49	3.86	0.63
Appropriateness*	3.94	0.65	3.51	0.79
System Performance	3.90	0.53	3.64	0.62
Input Sensitivity	1.80	0.98	1.63	0.86
Functionality*	4.33	0.99	3.66	1.20
Intuitiveness	3.84	0.72	3.76	0.87
Disorientation	1.45	0.65	1.59	1.02
Immersion	3.89	0.87	3.70	0.97
Presence*	4.14	0.60	3.80	0.58

for the *Input* and *Immersion/Presence* Usability Factors (all $\alpha >= 0.8$) in both experiments. However, the *Fidelity* factor shows questionable internal consistency in Experiment 1 and acceptable internal consistency in Experiment 2.

Examining the scores, we found that the reported values were generally higher in Experiment 1, where all factors were above 4 on a 5-point scale, than in Experiment 2, where all factors were above 3.5 on the same scale. However, no statistical difference was found between the two experiments. When we analyzed the combined scores (the average of the two experiments), we found a significant positive correlation between certain factors. In particular, there was a high positive correlation between the factors *Input* and *Fidelity* (Persons' r = 0.64, $p < 0.001$), as well as between the factors *Fidelity* and *Immersion/Presence* (Persons' r = 0.63, $p < 0.001$). These results suggest that interacting with the system felt natural and likely contributed to the realistic perception of the VE, thereby enhancing the sense of immersion and presence.

Further insight into the differences between the two experiments can be obtained by examining the Diagnostic Factors derived from the VRUSE questionnaire (Table 2). In this analysis, we found similar results, with comparable scores between the two experiments and Experiment 1 consistently scoring higher on all factors. Significant statistical differences ($p < 0.01$) between the two experiments were found in four categories: Ease of Use, Appropriateness, Functionality, and Presence. These differences suggest that the introduction of social elements in Experiment 2 may have influenced participants' perceptions of the VR system in these specific areas. The largest difference was observed in the Functionality category, which measures the user's sense of control over the system during the task [38]. One possible explanation for this result is the combined performance of the subject and the NPC (i.e., how well they worked together to achieve the task objectives, or how the NPC's actions and strategies affected the participant's performance in a competitive scenario), which may have affected the participant's perceived control over the VR system, possibly affecting their sense of functionality and effectiveness in

TABLE 3. Results of the simulator sickness questionnaire.

Category	Exp 1		Exp 2	
	Mean	SD	Mean	SD
General Discomfort	1.22	0.42	1.41	0.67
Fatigue	1.92	0.81	2.07	0.88
Headache	1.37	0.64	1.41	0.74
Eye Strain	1.98	0.78	1.78	0.91
Difficulty Focusing	1.18	0.53	1.27	0.67
Fullness of the Head	1.59	0.86	1.22	0.42
Blurred Vision	1.20	0.50	1.15	0.42
Dizziness with Eyes Closed	1.22	0.55	1.17	0.44
Vertigo	1.08	0.34	1.05	0.22

using the system. The lower value for Presence in Experiment 2 can be justified by the fact that, in the VE design, we used avatars with non-realistic graphical quality due to challenges in real-time management of the VR simulation.

3) SSQ

Internal consistency of the SSQ was good for Experiment 1 ($\alpha = 0.82$) and acceptable for Experiment 2 ($\alpha = 0.75$). The results (Table 3) showed that participants reported no significant problems with the VR simulation in terms of discomfort or sickness. All parameters were rated less than 2 on a 4-point Likert scale, indicating a generally positive user experience, with the exception of the *fatigue* parameter, which had a value slightly greater than 2 (2.07) in Experiment 2 and an equally high value (1.98) in Experiment 1. It is important to note that these values are still in the lower range of the Likert scale, indicating that the reported fatigue was relatively low and was not a major concern. Possible explanations for these results include the long exposure to the VR environment (lasting approximately 45 minutes) and engagement in the cognitive tasks that required participants to focus their attention on color-word congruence while interacting with the VE. We also report the (relatively) higher values for the *Eye Strain* parameter (1.98 in Experiment 1 and 1.78 in Experiment 2), which in turn can be explained by the duration of the VR experience, in which participants were continuously presented with different visual stimuli that required them to switch visual focus between different points during the tests. Participants had to look buttons' console to enter responses, direct their gaze to the monitor to see the task instructions and stimuli, and also focus their attention on the NPC's activities. This increased visual workload and constant adjustment of focus may have resulted in higher perceived eye strain.

4) FSS

This questionnaire measures the subjective experience of flow, which refers to a state of optimal engagement and immersion in an activity. The FSS is based on 36 items rated on a five-point Likert scale and assigned to nine different categories, including (1) Challenge-Skill Balance, (2) Action-Awareness Merging, (3) Clear Goals, (4) Unambiguous Feedback, (5) Concentration on Task at Hand, (6) Sense of Control, (7) Loss of Self-Consciousness, (8) Transformation of Time, and (9) Autotelic Experience [37].

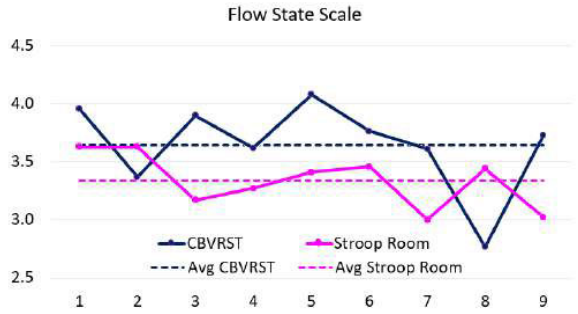


FIGURE 7. Flow State Scale. On the x-axis are the 9 categories of flow, while on the y-axis is the 5-point Lickert scale average score. The blue solid line is the average score of IB-VRST, while the the blue dotted line is the average score across all categories, the purple solid line represents the scores of Stroop Room [32], and the purple dotted line is their average. Better viewed in color.

Internal consistency of the FSS was excellent in both experiments, with Cronbach's $\alpha = 0.90$ in Experiment 1 and $\alpha = 0.92$ in Experiment 2. In Fig. 7, we present the average score for each category, along with the Flow scale, i.e., the average of all subscales. In the same graph, for assessment purposes, we give the FSS scores obtained in the evaluation of the Stroop Room [32]. Since there are no significant differences between Experiment 1 and 2 for any category, we present the average IB-VRST results.

In general, we can say that IB-VRST achieved a good Flow score of 3.56, with particularly high scores in categories 3 and 5 (Clear Goals and Concentration on Task at Hand, suggesting high levels of engagement). However, we also observed a significantly lower score in category 8 (Transformation of Time), suggesting that participants did not perceive a significant alteration in their perception of time during the VR experience. This lower score may be attributed to the presence of a timer on the TV screen, which effectively synchronizes participants' subjective experience of time with the actual flow of time. The clear visibility of the timer may have prevented participants from perceiving time as distorted or stretched, resulting in a lower score in this particular category.

When comparing with the Stroop Room [32], we see that IB-VRST achieved a higher Flow score and higher scores in most categories, suggesting that our system is more effective in immersing and engaging subjects in the activities proposed in the VE.

B. PERFORMANCE METRICS

To gain further insight into the effects of influencers, we examined performance metrics, including response times, Stroop interference, and task failures.

1) RESPONSE TIMES

The average response times for both experiments are reported in Fig. 8. In Experiment 1, the average response time across

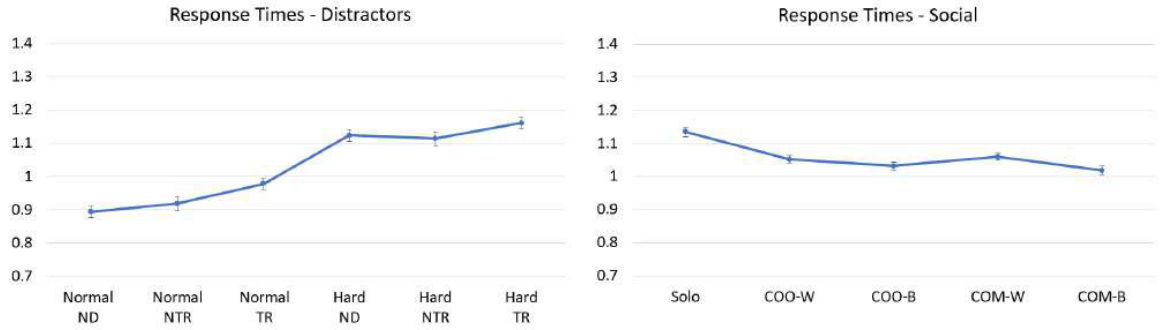


FIGURE 8. The graphs for response times for both experiments. Error bars represent standard errors.

all conditions (ND, TR, and NTR) in the *Normal* mode was 0.93 seconds and increased significantly in *Hard* mode (1.13 seconds, $F_{(1,51)} = 217.12, p < 0.001, \eta_p^2 = 0.80$). The introduction of task-switching in *Hard* mode primarily contributed to the increased response times, as users had to shift their attention between different stimulus-response mappings, leading to longer response times and higher cognitive load.

To investigate the effects of distractor types on response times, we performed pairwise post-hoc tests. Surprisingly, in the *Normal* mode, there was no statistical difference between the ND and NTR conditions, suggesting that the presence of non-task-related distractors does not significantly affect response times. However, response times were significantly higher in the TR condition than in the ND and NTR conditions (both $p < 0.001$). The presence of task-related distractors likely captured users' attention and increased response times because they had to process both the color-word conflict and the task-related distractor effects simultaneously.

In contrast, there is no significant difference between the three conditions in *hard* mode. The reason for this is probably the introduction of task-switching, which increased the complexity of the Stroop task and decreased the relevance of the different classes of distractors.

In Experiment 2, social presence conditions revealed a significant contribution to response times ($F_{(2,98)} = 57.61, p < 0.001, \eta_p^2 = 0.54$). Specifically, the *Solo* mode showed significantly higher response times compared to all social presence conditions (all $p < 0.001$). These results support the findings in [16] and [17], suggesting that the presence of social partners, regardless of their actual role, likely increases user engagement and motivation, leading to faster response times. However, no individual effect was found for the cooperative (COO) and competitive (COM) conditions alone, but the other's ability revealed a significant effect ($F_{(1,49)} = 21.58, p < 0.001, \eta_p^2 = 0.30$). Analyzing each combination in detail, we observed significantly slower response times for COM-W compared to COM-B ($p < 0.001$), suggesting that competition with a better opponent motivated subjects to improve their performance, at least in terms of response time.

In addition, when comparing response times between the two experiments, we found significant differences ($p < 0.001$) with users in Experiment 2 averagely responding faster. These differences can be attributed to the different degrees of task complexity resulting from the combination of Stroop task settings and influencers. Overall, social conditions and the NPCs' abilities acted as facilitators and helped subjects focus on the tasks.

2) STROOP INTERFERENCE

To assess the impact of influencers on participants' cognitive processing, we examined the Stroop interference effect (summarized in Fig. 9, which compares response times between congruent, blue bars, and incongruent stimuli, orange bars).

In Experiment 1, anomalies were observed in the *Normal* mode. In particular, participants in the ND condition showed an interference effect, whereas, surprisingly, no significant interference was found in the TR and NTR conditions. The difference lies in the response times for congruent stimuli. Participants exposed to distractors showed a significant increase in response times for congruent stimuli compared to the ND condition ($p < 0.01$), effectively reducing the gap between congruent and incongruent response times.

One possible explanation for this observation is that the introduction of never-seen-before influencers in the TR and NTR modes may have captured participants' attention, creating an additional source of cognitive conflict. As a result, participants might have shifted their focus from color-word dissonance to processing and integrating the new influencers into their decision-making process. This shift in attentional focus may have led to comparable response times for congruent and incongruent stimuli, reducing the interference effect.

A significant difference in Stroop interference was observed between the different modes ($F_{(1,51)} = 250.17, p < 0.001, \eta_p^2 = 0.80$). In all conditions of the *Hard* mode, participants had to switch between two different tasks, which likely increased the basic cognitive load and attentional demands. As a result, the interference effect became more pronounced, suggesting greater difficulty processing the incongruent stimuli compared to the congruent stimuli.

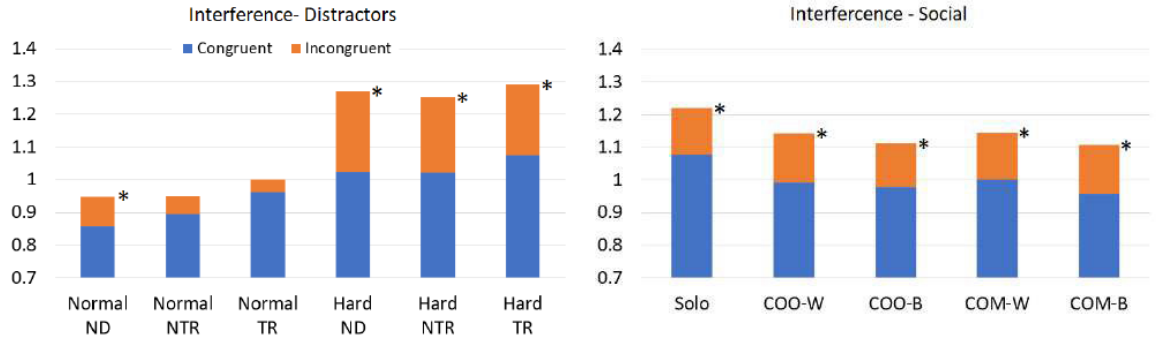


FIGURE 9. Graphs showing the Stroop interference effect. The blue bars show mean response times for congruent stimuli (60%), while the orange bars show response times for incongruent stimuli (40%) in seconds. The gap between the two bars (i.e., the visible portion of the orange bar) represents the amount of interference. The asterisk symbols indicate configurations where the interference was significant.

Interestingly, in the *Hard* mode, there was no significant difference in the interference effect between the different distractor conditions (ND, TR, and NTR). This suggests that the task-switching component was the main driver of the interference effect in *Hard* mode, overshadowing the specific effects of the different types of distractors. These results also highlight the role of task complexity and cognitive demands in modulating the interference effect.

In Experiment 2, the Stroop interference effect was consistently observed in all conditions. The two social presence influencers (i.e., social condition and other's ability) do not show an individual significant effect which, however, is observed from their interaction ($F_{(1,49)} = 3.89, p = 0.054, \eta_p^2 = 0.07$).

Overall, Experiment 2 showed that color-word dissonance consistently affected participants' response times. The type of social interaction modulated the interference effect to some extent, with certain conditions leading to slightly higher cognitive interference.

When comparing the interference effect between the two experiments, we found significant differences ($p < 0.001$). We recall that these two conditions only differ in the presence of distractors or social influencers. Therefore, distractors significantly increase interference, while social presence has no significant bearing on the basic cognitive mechanisms involved in processing conflicting stimuli.

3) FAILURES

Analysis of the failures (Fig. 10) provides additional insight into the subjects' performance. In Experiment 1, the pattern of failures is consistent with observed trends in response time, with *Normal* vs. *Hard* mode being the main influencing factor. Introducing task-switching in *Hard* mode significantly increases the total number of failures compared to *Normal* mode ($F_{(1,51)} = 134.57, p < 0.001, \eta_p^2 = 0.72$). Despite being smaller, also distractors have a significant impact on participants' performance ($F_{(1,102)} = 5.16, p < 0.05, \eta_p^2 = 0.09$). In *Normal* mode, their presence leads to more errors and timeouts although the only significant difference in total failure was found between ND and TR ($p < 0.05$), indicating the influence of task-related distractors.

In the *Hard* mode, the inclusion of task-switching exacerbated the effects of distractors on participants' ability to complete the task accurately and on time. The trend of total failures among conditions was the same of the *Normal* Mode (i.e., first TR, then NTR, and ND) but with consistently higher levels of errors, elapsed timer, and total failures. Surprisingly, however, there is no statistical difference between the *Hard* mode conditions, suggesting that the introduction of task-switching most significantly impacted performance. The presence of task-switching alone causes significant difficulties and makes the additional influence of distractors less notable than in *Normal* mode.

In Experiment 2, failure analysis shows a pattern consistent with response times. Compared to the *Solo* mode, the presence of an NPC has a significant effect on total failures ($F_{(2,98)} = 28.79, p < 0.001, \eta_p^2 = 0.37$). The highest number of failures is observed in *Solo* mode, suggesting that social presence favors performance enhancement as differences with both collaboration and competition are significant (all $p < 0.001$). Conversely, very small and not significant effects were found for each social influencer individually (COO or COM and W or B), but their interaction revealed a significant difference ($F_{(1,49)} = 4.13, p < 0.05, \eta_p^2 = 0.07$). Overall these results suggest that the social facilitation effect might depend only on performing the task with another person and not on these variables. Another interesting result is the higher proportion of unanswered questions in the total number of failures in all conditions in which subjects solve the task alone (i.e., in Experiment 1 and in the Solo condition of Experiment 2). This observation suggests that social presence increases user attention and motivation, resulting in a lower number of unanswered questions.

These observations are also supported by the comparison of total errors between the two experiments. In Experiment 2, users significantly ($p < 0.001$) end up committing fewer errors, thus, suggesting that the introduction of social

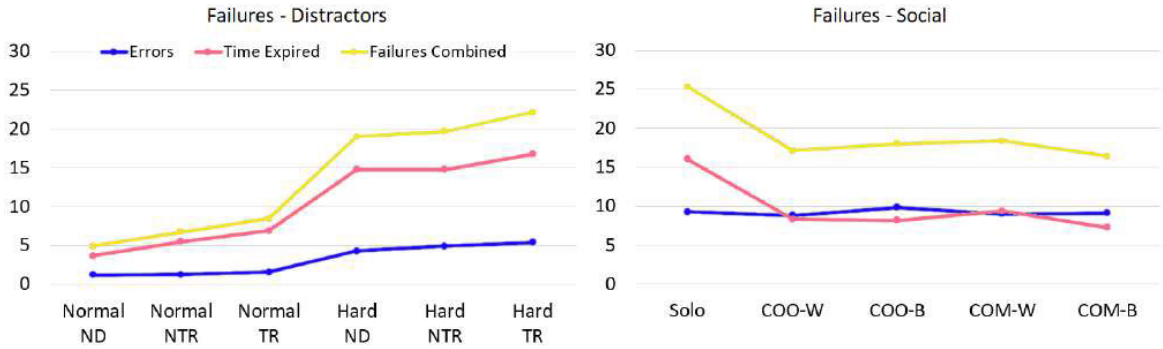


FIGURE 10. Graphs showing the number of errors for both experimental sessions. The graphs report both *errors* and *time expired*, meaning wrong responses and failures to respond in time.

influencers can act as facilitators compared to the detrimental effect of distractors.

C. STRESS ASSESSMENT

To evaluate the stress induced by IB-VRST in both experiments, we will now discuss the subjective self-reported stress levels and heart rate (HR). We chose this physiological measure as previous studies show HR to be a reliable measure of induced stress both during a competitive Stroop test [19] and in VEs [40].

1) SELF REPORTED STRESS LEVELS

Analysis of self-reported stress levels (whose z-scores are summarized in Fig. 11) revealed several notable results. In *Hard* mode, users reported higher stress levels than in *Normal* mode ($F_{(1,49)} = 73.73, p < 0.001, \eta_p^2 = 0.60$), confirming the effect of task-switching on stress (and performance, as mentioned earlier). Also, the distractor type is a source of induced stress ($F_{(2,98)} = 23.79, p < 0.001, \eta_p^2 = 0.33$). In the *Normal* mode, the NTR condition elicited the highest stress compared to ND ($p < 0.001$) and TR ($p < 0.01$), although the TR condition resulted in worse performance. One possible explanation is that the NTR distractors are characterized by sudden and unexpected events in the virtual environment that triggered stress due to the expectation of the unknown. In contrast, the TR distractors were easier to anticipate due to their repetitive nature, but had a negative effect on the cognitive aspect of the task and led to more errors. These results suggest that self-perceived stress and cognitive performance may not always align. Similarly, in the *Hard* mode, the NTR condition was the most stressful compared to ND and TR (both $p < 0.01$).

In Experiment 2, a main effect on self-reported stress was detected for both social condition ($F_{(1,47)} = 19.54, p < 0.001, \eta_p^2 = 0.29$) and other's ability ($F_{(1,47)} = 12.82, p = 0.001, \eta_p^2 = 0.21$). The COM-B social condition elicited the highest stress levels compared to all the other social conditions (all $p < 0.001$), whereas the other social conditions had comparable and lower stress levels. These results highlight that cooperation had a minimal effect on stress arousal but contributed to better performance, similar to what happens when competing against a weaker opponent. However, when competing against a more skilled opponent, performance improvements are also associated with the highest stress levels among all conditions, which may be attributed to the combination of higher challenge, pressure to perform, and fear of failure.

By comparing stress levels between the two experiments, we can observe a significant difference in reported stress levels ($p < 0.001$). Specifically, we are comparing *Hard* modes conditions of Experiment 1 with the social conditions of Experiment 2 as they both share the same task complexity introduced by task-switching. In general, the stress levels reported in Experiment 1 exhibit higher values, which nearly double when we compare the most stressful condition, *Hard* NTR, with the least stressful COO-B. An exception is made for condition COM-B which is comparable to the *Hard* ND one. Overall, the results suggest the greater impact on stress originated from task complexity (i.e., given the difference between *Normal* vs. *Hard* modes), which subsequently affects performance. In Experiment 2, on the other hand, stress arises primarily from the competitive nature of the task and the opponent's skill, which in turn urges users to improve their performance.

2) HEART RATE

Looking at the HR data in Experiment 1 (Fig. 12, left), we observe a significant increase (all $p < 0.01$) in all phases compared to *Familiarization* baseline. We can also observe that HR remains essentially stable within each mode as distractors type do not have a significant effect on HR ($F_{(2,100)} = 0.55, p = 0.57, \eta_p^2 = 0.01$). Conversely, switching to the *Hard* mode results in a slightly decreasing average level of HR ($F_{(1,50)} = 28.93, p < 0.001, \eta_p^2 = 0.36$). Given these results, it is interesting to note that the lower scores in the *Hard* mode contradict the patterns observed in other metrics (e.g., increased Stroop interference, lower performance, and higher self-reported stress levels). This discrepancy could be due to users physiologically stabilizing as they become more familiar with the environment, task, and distractors. However, further research is needed to understand the underlying factors contributing to these observations.

In contrast, Experiment 2 shows a clearer trend in the HR data compared to Experiment 1 (Fig. 12, right). With

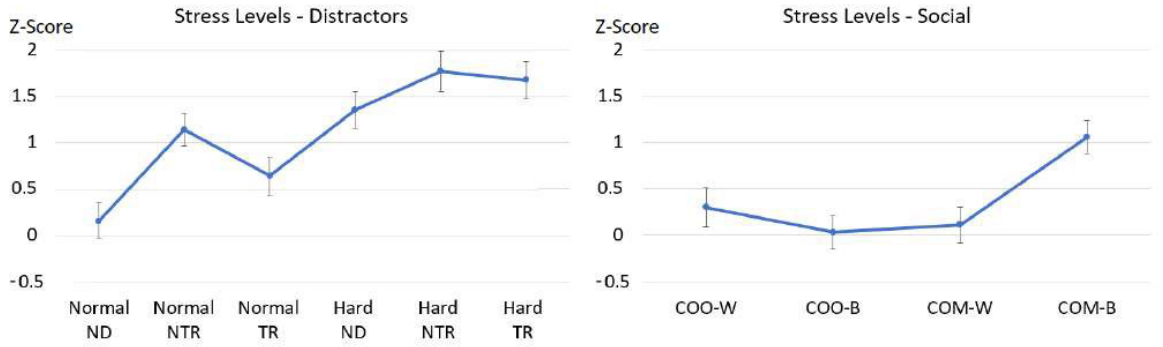


FIGURE 11. Z-scores for Self-reported stress levels in both experiments. All values have been adjusted according to baseline values: Familiarization mode for Experiment 1, and Solo mode for Experiment 2. Error bars represent standard errors.

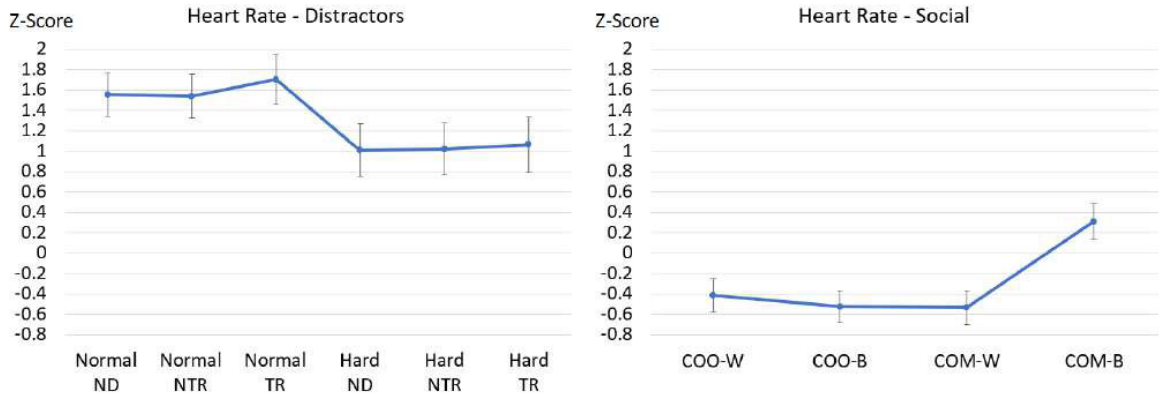


FIGURE 12. Z-scores for Heart Rate in both experiments. All values have been adjusted according to baseline values: Familiarization mode for Experiment 1, and Solo mode for Experiment 2. Error bars represent standard errors.

the exception of COM-B, all conditions are below the baseline (*Solo* mode), with significant differences observed for COO-B and COM-W (both $p < 0.01$). A main effect on HR was detected for both social condition ($F_{(1,49)} = 5.97, p = 0.018, \eta_p^2 = 0.10$) and other's ability ($F_{(1,49)} = 13.41, p < 0.001, \eta_p^2 = 0.21$). However, COM-B is the only condition where the HR z-scores are above the baseline, with significantly different scores from the other social conditions (all $p < 0.001$). The trend in HR is consistent with that of self-reported stress scores, confirming that competition with a better opponent is the most stressful social condition in both self-perception and physiological measures, primarily due to the competitive nature of the task rather than its complexity.

Comparison of HR between the two experiments reveals differential effects of distractors and social influencers ($p < 0.001$), with the former leading to an increase in subjects' stress levels and the latter primarily reducing stress. A notable exception is competition with a better opponent, which can induce a sense of pressure and challenge that triggers the subjects' psychological and physiological stress response.

VI. DISCUSSION

In this section, we try to answer the RQs proposed in this study by broadly discussing the results presented in Section V.

RQ1: *How do immersive VR affect user engagement, immersion, and presence in VRSTs?*

Analysis of the questionnaires provided valuable insights into the subjective experiences and perceptions of the participants. Overall, the results were encouraging, reflecting positive responses from users and suggesting the effectiveness of the interventions implemented. The results from SUS indicate a high level of user satisfaction with the VR system, which effectively assisted them in accomplishing their tasks. The results from VRUSE and FSF indicate that users felt engaged and immersed in the virtual environment and were not or minimally uncomfortable with the proposed immersive VR system, according to the SSQ.

These overall positive results suggest that the proposed VR system provides an immersive, engaging, and user-friendly experience. This is a positive indication for the future adoption of immersive VR technologies in similar contexts, as they have the potential to enhance user engagement and achieve the desired outcomes. It is worth noting that in previous VRST literature, no comparable study has been conducted to investigate multiple dimensions of user experience using standardized questionnaires. Therefore, this study contributes to gain a better understanding of user experience in VRSTs.

In addition, the efficacy of IB-VRST in inducing the Stroop effect contributes to the significance of this study. The results

confirmed that participants took more time to complete incongruent trials compared to congruent trials, which is consistent with previous Stroop studies. This confirms that the virtual version of the Stroop task is an effective tool for eliciting the Stroop effect.

Considering these results, we believe that the design choices adopted for IB-VRST, which address the VE characteristics and the interaction system, can serve as valuable reference guidelines for researchers and practitioners.

RQ2: *What is the impact of different types of distractors (task-related vs. non-task-related) on user performance and induced stress in VRSTs?*

The results of Experiment 1 revealed several significant patterns. With respect to performance, we observed that the introduction of task-switching in *Hard* mode had a detrimental effect. Participants showed higher response times, and made more errors than in *Normal* mode.

As for the Stroop effect, the results highlight the influence of both task-related and non-task-related distractors, with non-task-related distractors causing more interference. However, the impact of distractors on the Stroop effect was modulated by the complexity of the task (i.e., *Normal* vs. *Hard* mode), with task-switching playing a dominant role in the *Hard* mode.

Examining the effects of different distractors in the *Normal* mode, we found that the NTR condition elicited the highest sense of stress in participants. This was in contrast to the TR condition, which had worse performance results (more errors and higher response times) but did not significantly increase stress levels compared to the ND condition. These results suggest that self-perceived stress and performance on cognitive tasks may not always be interdependent. The NTR distractors, which presented sudden and unexpected events in the virtual environment, elicited stress due to the anticipation of the unknown, whereas the TR distractors, which were more predictable, primarily affected the cognitive aspect of the task.

Interestingly, in terms of self-reported stress level, we did not observe significant differences between the different distractor conditions (ND, TR, and NTR) within the *Hard* mode. This suggests that the stress induced by task-switching overridden the effects of the specific types of distractors.

Regarding heart rate (HR), our data showed an overall increase in HR in all phases of Experiment 1 compared with baseline (*Familiarization* phase). This increase in HR suggests that distractors generally induced stress in users, which is consistent with their self-reported stress levels. However, we also found a decrease in HR in the *Hard* mode compared to the *Normal* mode. This is inconsistent with the patterns observed in performance measures and self-reported stress levels. One possible explanation for this discrepancy is a physiological stabilization effect as participants become more familiar with the environment, task, and distractors encountered. However, further research is sorely needed to provide a sound support to these observations.

RQ3: *What is the impact of social presence (competing with an opponent vs. collaborating with a teammate) on user performance and induced stress in VRSTs?*

Comparing the different social conditions in Experiment 2, leads to several interesting observations. First, all social conditions facilitate performance improvement (i.e., faster response times and fewer errors) compared to the Solo condition. This suggests that both collaboration and competition improve task performance. Second, performance improvement in competition with a more skilled opponent was associated with increased stress levels. Indeed, the COM-B condition was rated as the most stressful, significantly differing from the other social conditions, which had lower stress levels and were not significantly different from each other. Third, in contrast to the results of Experiment 1, in Experiment 2, the overall trend of HR was consistent with participants' perceived stress levels, supporting the relationship between subjective stress and physiological responses. In particular, the self-reported most stressful condition (COM-B) was paired with a higher HR than the other social conditions.

In summary, the analysis of Experiment 2 suggests that social influencers elicit a social facilitation effect, i.e., an improvement in performance when a task is performed with other individuals compared to when it is performed alone [41]. Competing against a better opponent also improves performance but with a significant increase in stress levels compared to collaborating with any partner or competing with a poorer performer. These results highlight the complex interplay between social presence, task performance, and induced stress in VE. Of particular note, these results suggest a possible extension to tasks other than the Stroop task, where collaborative social presence could be used to improve user performance, without negatively affecting subjects' psychophisiological stress.

RQ4: *Which is the differential impact of various types of influencers (i.e., distractors and social presence)?*

The differential effects of distractors and social presence on performance and stress are evident from the experimental results. Distractors primarily impair performance and induce stress, whereas collaboration in social presence improves performance and even acts as a stress alleviator when compared to a *Solo* condition. Competition, however, has a more nuanced effect and, as noted, leads to better performance, but possibly also to higher stress levels, depending on how skilled the opponent is. These results highlight the importance of considering the design and implementation of influencers in VE to induce interference effects, optimize performance and effectively manage stress.

VII. LIMITATIONS AND FUTURE WORK

This work was subject to some limitations. First, our results highlighted the complex interplay between social factors, task demands, and individual differences. This interplay could have potentially introduced co-occurring variables that made it more difficult to isolate and examine the specific

effects of each factor independently. Future studies with more controlled experimental designs could help shed further light on the individual effects of each influencer and provide a more comprehensive understanding of their contributions.

Second, interactions with the VE were mediated by the use of controllers bundled with the HMD. This approach may have imposed constraints on the naturalness of users' interactions with the Stroop interface and their embodied avatar. In future studies, we could investigate the effects of smart gloves or external bare-hand tracking devices on the naturalness and immersion of the interaction.

Third, the quality of the VR experience itself may have had limitations. While efforts have been made to create visually appealing and realistic avatars, there is still room to increase the level of realism in terms of avatar appearance.

Finally, our study focused on social presence with virtual avatars, but we did not validate the results by comparing them to interactions with real human competitors or teammates. The use of virtual avatars may lead to differences in behavior and social cues compared to real human interactions, which could influence the observed effects. This research will be also part of future work.

VIII. CONCLUSION

In this paper, we presented the Influencer-Based Virtual Reality Stroop test (IB-VRST), a novel implementation of the Stroop task in immersive VR that can support different types of influencers such as distractors and social presence. This work makes the following contributions to the research. First, we have conducted a comprehensive investigation of usability, user experience, and sense of immersion and presence in a VRST. Second, the comparative analysis of different influencers provides novel insights into the specific effects of these elements on cognitive performance and stress responses. Third, our work is the first immersive VRST to use virtual avatars to explore the effects of social presence. Finally, we conducted a unique comparison of the psychophysiological and performance effects of collaborative and competitive social conditions on Stroop task performance.

REFERENCES

[1] J. R. Stroop, "Studies of interference in serial verbal reactions," *J. Experim. Psychol.*, vol. 18, no. 6, pp. 643–662, Dec. 1935.

[2] B. M. Ben-David, L. L. T. Nguyen, and P. H. H. M. van Lieshout, "Stroop effects in persons with traumatic brain injury: Selective attention, speed of processing, or color-naming? A meta-analysis," *J. Int. Neuropsychological Soc.*, vol. 17, no. 2, pp. 354–363, Feb. 2011.

[3] T. D. Parsons, C. G. Courtney, B. Arizmendi, and M. Dawson, "Virtual reality Stroop task for neurocognitive assessment," in *Medicine Meets Virtual Reality*, vol. 18. Amsterdam, The Netherlands: IOS Press, 2011, pp. 433–439.

[4] J. A. King, M. Colla, M. Brass, I. Heuser, and D. Y. von Cramon, "Inefficient cognitive control in adult ADHD: Evidence from trial-by-trial Stroop test and cued task switching performance," *Behav. Brain Functions*, vol. 3, no. 1, p. 42, 2007.

[5] T. D. Parsons and A. R. Carlew, "Bimodal virtual reality Stroop for assessing distractor inhibition in autism spectrum disorders," *J. Autism Develop. Disorders*, vol. 46, no. 4, pp. 1255–1267, Apr. 2016.

[6] P. Renaud and J.-P. Blondin, "The stress of Stroop performance: Physiological and emotional responses to color–word interference, task pacing, and pacing speed," *Int. J. Psychophysiol.*, vol. 27, no. 2, pp. 87–97, Sep. 1997.

[7] N. Cowan and A. Barron, "Cross-modal, auditory-visual Stroop interference and possible implications for speech memory," *Perception Psychophysics*, vol. 41, no. 5, pp. 393–401, Sep. 1987.

[8] E. M. Elliott, C. C. Morey, R. D. Morey, S. D. Eaves, J. T. Shelton, and D. A. Lutfi-Proctor, "The role of modality: Auditory and visual distractors in Stroop interference," *J. Cognit. Psychol.*, vol. 26, no. 1, pp. 15–26, Jan. 2014.

[9] A. A. Rizzo, J. G. Buckwalter, T. Bowerly, C. Van Der Zaag, L. Humphrey, U. Neumann, C. Chua, C. Kyriakakis, A. Van Rooyen, and D. Sisemore, "The virtual classroom: A virtual reality environment for the assessment and rehabilitation of attention deficits," *CyberPsychology Behav.*, vol. 3, no. 3, pp. 483–499, Jun. 2000.

[10] A. A. Rizzo, T. Bowerly, J. G. Buckwalter, D. Klimchuk, R. Mitura, and T. D. Parsons, "A virtual reality scenario for all seasons: The virtual classroom," *CNS Spectrums*, vol. 11, no. 1, pp. 35–44, Oct. 2009.

[11] T. D. Parsons and M. D. Barnett, "Virtual apartment Stroop task: Comparison with computerized and traditional Stroop tasks," *J. Neurosci. Methods*, vol. 309, pp. 35–40, Nov. 2018.

[12] A. D. Eastvold, H. G. Belanger, and R. D. Vanderploeg, "Does a third party observer affect neuropsychological test performance? It depends," *Clin. Neuropsychologist*, vol. 26, no. 3, pp. 520–541, Apr. 2012.

[13] M. Barnett, J. Sawyer, and J. Moore, "An experimental investigation of the impact of rapport on Stroop test performance," *Appl. Neuropsychol., Adult*, vol. 29, no. 5, pp. 941–945, Sep. 2022.

[14] D. R. Saunders, D. Melcher, and W. van Zoest, "No evidence of task co-representation in a joint Stroop task," *Psychol. Res.*, vol. 83, no. 5, pp. 852–862, Jul. 2019.

[15] R. Sellaro, B. Treccani, and R. Cubelli, "When task sharing reduces interference: Evidence for division-of-labour in Stroop-like tasks," *Psychol. Res.*, vol. 84, no. 2, pp. 327–342, Mar. 2020.

[16] D. P. MacKinnon, R. E. Geiselman, and J. A. Woodward, "The effects of effort on Stroop interference," *Acta Psychologica*, vol. 58, no. 3, pp. 225–235, Mar. 1985.

[17] P. Huguet, F. Dumas, and J.-M. Monteil, "Competing for a desired reward in the Stroop task: When attentional control is unconscious but effective versus conscious but ineffective," *Can. J. Experim. Psychol. Revue canadienne de Psychologie Expérimentale*, vol. 58, no. 3, pp. 153–167, Sep. 2004.

[18] F. Dumas, P. Huguet, and E. Ayme, "Social context effects in the Stroop task: When knowledge of One's relative standing makes a difference," *Current Psychol. Lett.*, vol. 2, no. 16, Jun. 2005.

[19] V. Mueller, R. Richer, L. Henrich, L. Berger, A. Gelardi, K. M. Jaeger, B. M. Eskofier, and N. Rohleder, "The Stroop competition: A social-evaluative Stroop test for acute stress induction," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Sep. 2022, pp. 1–4.

[20] K. Kreijns, K. Xu, and J. Weidlich, "Social presence: Conceptualization and measurement," *Educ. Psychol. Rev.*, vol. 34, no. 1, pp. 139–170, Mar. 2022.

[21] S. Coelli, G. Tacchino, E. Rossetti, M. Veniero, L. Pugnetti, F. Baglio, and A. M. Bianchi, "Assessment of the usability of a computerized Stroop test for clinical application," in *Proc. IEEE 2nd Int. Forum Res. Technol. Soc. Ind. Leveraging Better Tomorrow (RTSI)*, 2016, pp. 1–5.

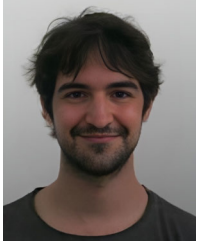
[22] G. G. Robertson, S. K. Card, and J. D. Mackinlay, "Three views of virtual reality: Nonimmersive virtual reality," *Computer*, vol. 26, no. 2, p. 81, Feb. 1993.

[23] J. Blascovich, J. Loomis, A. C. Beall, K. R. Swinth, C. L. Hoyt, and J. N. Bailenson, "Immersive virtual environment technology as a methodological tool for social psychology," *Psychol. Inquiry*, vol. 13, no. 2, pp. 103–124, 2002.

[24] M. Henry, C. C. Joyal, and P. Nolin, "Development and initial assessment of a new paradigm for assessing cognitive and motor inhibition: The bimodal virtual-reality Stroop," *J. Neurosci. Methods*, vol. 210, no. 2, pp. 125–131, Sep. 2012.

[25] C. Belletier, A. Normand, and P. Huguet, "Social-facilitation-and-impairment effects: From motivation to cognition and the social brain," *Current Directions Psychol. Sci.*, vol. 28, no. 3, pp. 260–265, Jun. 2019.

[26] P. Huguet, M. P. Galvaing, J. M. Monteil, and F. Dumas, "Social presence effects in the Stroop task: Further evidence for an attentional view of social facilitation," *J. Personality Social Psychol.*, vol. 77, no. 5, pp. 1011–1025, 1999.
[27] D. C. Delis, E. Kaplan, and J. H. Kramer, "Assessment," in *Delis-Kaplan Executive Function System*. USA: SAGE, 2001.
[28] D. Wu, C. G. Courtney, B. J. Lance, S. S. Narayanan, M. E. Dawson, K. S. Oie, and T. D. Parsons, "Optimal arousal identification and classification for affective computing using physiological signals: Virtual reality Stroop task," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 109–118, Jul. 2010.
[29] C. M. Armstrong, G. M. Reger, J. Edwards, A. A. Rizzo, C. G. Courtney, and T. D. Parsons, "Validity of the virtual reality Stroop task (VRST) in active duty military," *J. Clin. Experim. Neuropsychol.*, vol. 35, no. 2, pp. 113–123, Feb. 2013.
[30] T. D. Parsons, C. G. Courtney, and M. E. Dawson, "Virtual reality Stroop task for assessment of supervisory attentional processing," *J. Clin. Experim. Neuropsychol.*, vol. 35, no. 8, pp. 812–826, Oct. 2013.
[31] T. McMahan, T. Duffield, and T. D. Parsons, "Feasibility study to identify machine learning predictors for a virtual school environment: Virtual reality Stroop task," *Frontiers Virtual Reality*, vol. 2, Aug. 2021, Art. no. 673191.
[32] S. Gradl, M. Wirth, N. Mächtlinger, R. Poguntke, A. Wonner, N. Rohleder, and B. M. Eskofier, "The Stroop room: A virtual reality-enhanced Stroop test," in *Proc. 25th ACM Symp. Virtual Reality Softw. Technol.*, Nov. 2019, pp. 1–12.
[33] R. Poguntke, M. Wirth, and S. Gradl, "Same same but different: Exploring the effects of the Stroop color word test in virtual reality," in *Proc. IFIP Conf. Hum.-Comput. Interact.* Cham, Switzerland: Springer, 2019, pp. 699–708.
[34] J. Brooke, "SUS-A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, no. 194, pp. 4–7, 1996.
[35] J. Brooke, "SUS: A retrospective," *J. Usability Stud.*, vol. 8, no. 2, pp. 29–40, Feb. 2013.
[36] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *Int. J. Aviation Psychol.*, vol. 3, no. 3, pp. 203–220, Jul. 1993.
[37] S. A. Jackson and H. W. Marsh, "Development and validation of a scale to measure optimal experience: The flow state scale," *J. Sport Exercise Psychol.*, vol. 18, no. 1, pp. 17–35, Mar. 1996.
[38] R. S. Kalawsky, "VRUSE—A computerised diagnostic tool: For usability evaluation of virtual/synthetic environment systems," *Appl. Ergonom.*, vol. 30, no. 1, pp. 11–25, Feb. 1999.
[39] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," *Psychiatry Invest.*, vol. 15, no. 3, pp. 235–245, Mar. 2018.
[40] V. Aspiotis, A. Miltiadous, K. Kalafatakis, K. D. Tzimourta, N. Giannakeas, M. G. Tsipouras, D. Peschos, E. Glavas, and A. T. Tzallas, "Assessing electroencephalography as a stress indicator: A VR high-altitude scenario monitored through EEG and ECG," *Sensors*, vol. 22, no. 15, p. 5792, Aug. 2022.
[41] R. B. Zajonc, "Social facilitation: A solution is suggested for an old unresolved social psychological problem," *Science*, vol. 149, no. 3681, pp. 269–274, Jul. 1965.



EDOARDO BATTEGAZZORRE is currently a Research Fellow with the Department of Control and Computer Engineering, Politecnico di Torino, a part of the Computer Graphics and Vision Group and the VR@Polito Group. His research interests include virtual and augmented reality applications for public health, safety and professional training, and agent-based models applied to emergency and natural disaster situations.

FRANCESCO STRADA is currently an Assistant Professor with the Department of Control and Computer Engineering, Politecnico di Torino, where he conducts research under the Computer Graphics and Vision Research Group. His current research interests include virtual and augmented reality for learning and training, serious games, human–computer interaction, and the use of immersive technologies in collaborative scenarios.

LUCIA DE FRANCESCO is currently pursuing the Ph.D. degree with the Department of Psychology, University of Turin, where she is a part of the Social Interaction Laboratory. Her research interests include the influence of social factors on behavior and physiological activation, through real and virtual settings.

ALESSANDRO MAZZA is currently pursuing the Ph.D. degree with the Department of Psychology, University of Turin, where he is a part of the Social Interaction Laboratory. His research interests include the behavioral aspects of social cognition and underlying neurophysiological mechanisms.

OLGA DAL MONTE is currently an Associate Professor and the Principal Investigator of the Social Interaction Laboratory, Department of Psychology, University of Turin. Her research interest includes understanding the neurophysiological mechanisms that underlie social behavior. She has been at the forefront of using naturalistic and ecological social interaction paradigms for studying several complex social behaviors. Some of the areas, she has studied include prosocial and antisocial behavior, cooperation and competition, interpersonal distance, affective touch, eye contact, and social attention.

ANDREA BOTTINO (Member, IEEE) is currently an Associate Professor with the Department of Control and Computer Engineering, Politecnico di Torino, where he heads the Computer Graphics and Vision Research Group. His current research interests include computer vision, machine learning, human–computer interaction, serious games, and virtual and augmented reality.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement