Towards a Methodological Framework for Multimodal Input in Social Interaction in Virtual Reality

Damla Kuleli*
Bournemouth University

Xun He†
Bournemouth University

Charlie Lloyd-Buckingham
Bournemouth University

Liucheng Guo
Tangi0 LTD (TG0)

Andrew James Hanson
Bournemouth University

Laura Vuillier
Bournemouth University

Nicola J. Gregory
Bournemouth University

Chang Hong Liu
Bournemouth University

Ciel Liu
Bournemouth University

Fred Charles‡
Bournemouth University

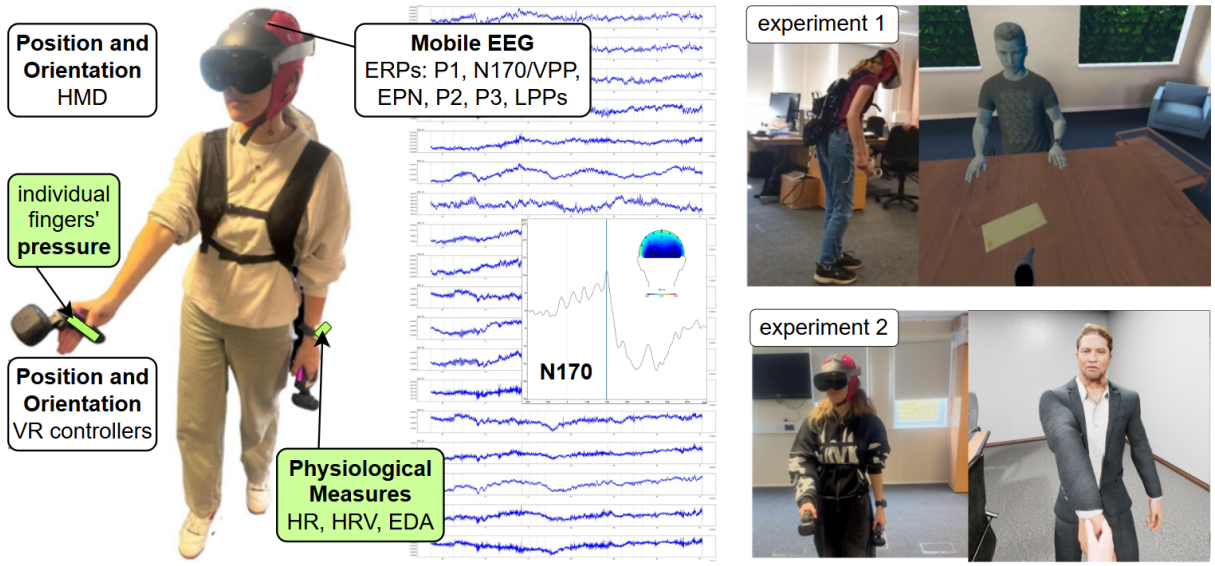

Figure 1: Overview of the experimental setups for the evaluation of human participants' interactions with virtual characters using various multimodal input. Experiment 1 and Experiment 2 show the participant's first-person view of the virtual environment. Experiment 2 includes additional data streams: fingers pressure and physiological measures (highlighted).

Abstract

Studying human behaviour and mind in realistic environments has been a challenge in various research fields. We formulate the foundational elements of a methodological framework for the multimodal research of social interaction in virtual reality (VR) based on two experiments involving human participants interacting with virtual humans. The framework comprises commonly used VR headsets and neurophysiological sensors to measure and evaluate participants' behaviours and brain activities. We provide guidelines related to methodological solutions, data collection and analysis pipelines.

Index Terms: Social Interaction, Face Perception, Social Anxiety, EEG, Virtual Reality.

1 Introduction

Social interaction within virtual environments (VEs) has been the focus of experimental studies for some time. Recent neuroscience research is also using socially interactive virtual characters (VCs) and multimodal neurophysiological recording to study human behaviour. This paper studies how human processes social interactions with believable human-like VCs within a controlled VE, focussing on facial expressions that have a significant role during social interactions. We aim to reveal how social interaction is planned upon viewing the interaction partner's face, by studying event-related potentials (ERPs) related to face and emotion perception (especially early face perception) such as P1, N170, vertex positive peak (VPP), early posterior negativity (EPN), P2, P3, and late positive potential (LPP). N170 and VPP in particular are widely accepted as a biomarker of early face perception [21]. Through a clear definition of the use of VEs for neuroscience research on human interaction with VCs, we hope to offer a framework to assess early human perception of social interactions within VR.

Another important consideration in using VCs is their believability ratio and their potential to enhance social interaction [20], which are further explored using current AI solutions to integrate interactions with believable VCs [22] towards the creation of compelling multimodal frameworks [16]. Our contribution is demonstrating the potential in creating realistic social interactions within VEs via the use of believable VCs animated with real-time immersive game engine technologies, and the presentation of a common analysis pipeline for multimodal neurophysiological recordings (Figure 1).

We present two experiments to establish our new framework. Experiment 1 reveals the effect of planned social interactions on

*e-mail: dkuleli@bournemouth.ac.uk
†e-mail: xhe@bournemouth.ac.uk
‡e-mail: fcharles@bournemouth.ac.uk

early face perception (i.e. embodied perception of social interactions [8]). Neural responses to human-like and statue VCs (VC types) during approach and withdraw behaviours (interaction types) are analysed. This paradigm works succesfully (shown in significant differences between interaction types and VC types) yet does not confirm the social nature of these effects (no interaction between these factors), demanding a better social interaction paradigm. The improved design of Experiment 2 has a more direct social interaction (handshake) to elicit stronger emotional responses in socially anxious participants during affective social interactions [14]. VR is widely used in the treatment of social anxiety disorder (SAD) by exposing individuals to anxiety-eliciting social situations. There are significant number of studies supporting the effectiveness of Virtual Reality Exposure Therapy (VRET) in reducing social anxiety symptoms [1, 7], and a meta-analysis states that there is no significant difference between VRET and in-vivo exposure therapy (IVET) compared to the control group in social anxiety [15]. Although there are a profound number of studies for the treatment of SAD in VR, there are a limited number of studies that assess SAD in VR using objective and direct measurements.

2 Experiment 1: Embodied social interaction in early face perception

Experiment 1 studies whether social interaction planning is embodied in early face perception by studying participants' neural responses to the VCs visualised as either human-like or statue under approach and withdraw behaviours. The study employs a 2 Character (human-like vs. statue) × 2 Interaction (approach vs. withdraw) within-subjects design and therefore repeated-measures ANOVA for analysis. We expect to find differences in early perceptual ERP components (e.g., N170) between the interaction types to the human-like VCs, and confirm its social nature by showing the reduction of such differences to the statue VCs.

2.1 Stimuli and System Architecture

The VE, developed using Unity (v.2021.3.4f1), is an office setting including two identical office rooms facing each other equipped with desk, chair, cabinet and sofas. 20 unique VCs (10 females and 10 males) representing various ethnic backgrounds are selected from Microsoft Rocketbox avatar library [13]. The human-like VCs have normal life-like skin tones. For the statue VCs, grey stone-like texture material is added. Animations and interactions, including head movements, eye gaze, blinks and verbal interactions are applied to the human-like VCs. Participants' proximity, head orientation, and object interactions are used to trigger the human-like VCs' animations to simulate social interaction. Human-like VCs will maintain eye contact with participants, show natural blinking behaviour, and respond by saying "thank you" when participants complete the task (putting an envelope on the desk in front of the VC). The "thank you" message is recorded by various actors to allow voice diversity. Statue VCs have no animation and do not interact with participants.

The experiment was run on desktop PC with Windows 10 and displayed on Meta Quest 2 HMD via Air Link and Steam VR using a local WiFi network . Lab Streaming Layer (LSL) is used to send event markers and synchronise data recording. The virtual display's onset delay is measured (70 ms with little variance) and corrected during EEG pre-processing.

2.2 Experimental Protocol

37 volunteers meeting the eligibility criteria (aged 18-50, no neurological disorder, normal or corrected-to-normal vision, no movement impediments) participated. 5 participants were excluded from analysis due to excessive artifacts.

There are ten 40-trial blocks. The approach vs. withdraw interactions are employed in the two halves of the experiment, with the order counterbalanced across participants. Character types (human vs. statue) and character models are presented randomly in each block.

Each participant is informed about the interaction type (approach vs. withdraw) when a block starts. Each experimental trial starts in darkness in VR. The participant sees two symbols, "X" (representing participant's fixation) and "O" (denoting the between-eyes point of the virtual character). The participant moves the head to align "X" and "O" (ensuring eye contact with the VC later). The lights turn on after a random interval of 500-1500 ms after the X-O alignment, showing a room and a VC (human or statue) sitting behind a desk (3 meters away) (Figure 1). Human-like characters maintain eye contact with participants, and show natural blinking behaviour; statue characters stay still. The participant holds an envelope in the right (VR) hand. On hearing a "go" sound (1000 ms after lights-on), the participant will either go forward to the VC (approach blocks) or turn around and approach another VC in the room behind (withdraw blocks), and put the envelope on the desk. Then, the participant returns to the starting position and facing direction. The trial ends by turning the lights off.

After the main experiment, to assess the participant's perception of VCs, perceived anthropomorphism and animacy ratings for the human-like and statue VCs are measured [3] outside of the VE.

2.3 EEG Recording and Pre-processing

EEG is recorded from 30 scalp locations with a mobile EEG system (ANT Neuro EegoSports) at a 500-Hz sampling rate, and pre-processed using EEGLAB (2022.1) and ERPLAB (2023.1). First, the 70 ms display delay is corrected. Then, non-experimental data, gross artifacts, and DC offset are removed. A 1-30 Hz bandpass digital filter is applied, along with a 50 Hz noise correction using cleanline 2.0. Bad channels are identified and interpolated with the automated subspace removal (ASR) plugin. Then, the data are segmented into epochs from -100 to 500 ms relative to the lights-on markers. EEG activities over $\pm$500 μV are removed. Then, EEG is re-referenced to the average and cleaned with independent component analysis (ICA) before baseline corrected to the 100 ms baseline. Finally, epochs are re-referenced to the infinity for mininal reference dependence [27] and averaged to generate ERPs for each experimental condition. Participants having less than 30 accepted epochs per condition are rejected from analysis.

2.4 Results and Discussion

Paired *t*-tests showed higher ratings for the human-like than statue VCs in both anthropomorphism (3.5 vs 1.8) and animacy (3.7 vs 1.2), *t*s >4.77, *p*s <.003, confirming that the human-like VCs are perceived more "social" than the statues.

Mean ERP amplitudes are quantified in time windows and scalp locations showing the maximal amplitudes (determined with the collapsed localiser method [18]), and analysed with 2 Character (human vs statue) × 2 Interaction (approach vs. withdraw) ANOVAs (JASP 0.17.3.0). An early Interaction effect in a frontal negativity (Fp1/Fp2, 82-201 ms) is significant, $F(1,31) = 4.84$, $p < .05$). N170 (P7/P8, 150-190 ms) and VPP (Cz, 152-172 ms) show no any effects (Fs <1.60, ps >.2) despite a significant Character effect in VPP, $F(1,30) = 9.71$, $p < .01$.

The Interaction effect in early ERPs validated our design and replicated previous findings in VR that early ERPs are modulated by action planning [6] and image realism [12]. However, we did not find the interaction between the factors. It is possible that social interaction is not embodied in early perception in highly realistic VEs (cf. the notion of embodied cognition [8]). Alternatively, the behaviour in Experiment 1 may not be social enough to engage attention on the VCs. To tell these explanations apart, we need a better paradigm to require more direct social interactions, thus demanding participants' focal attention on VCs (such as their attitudes

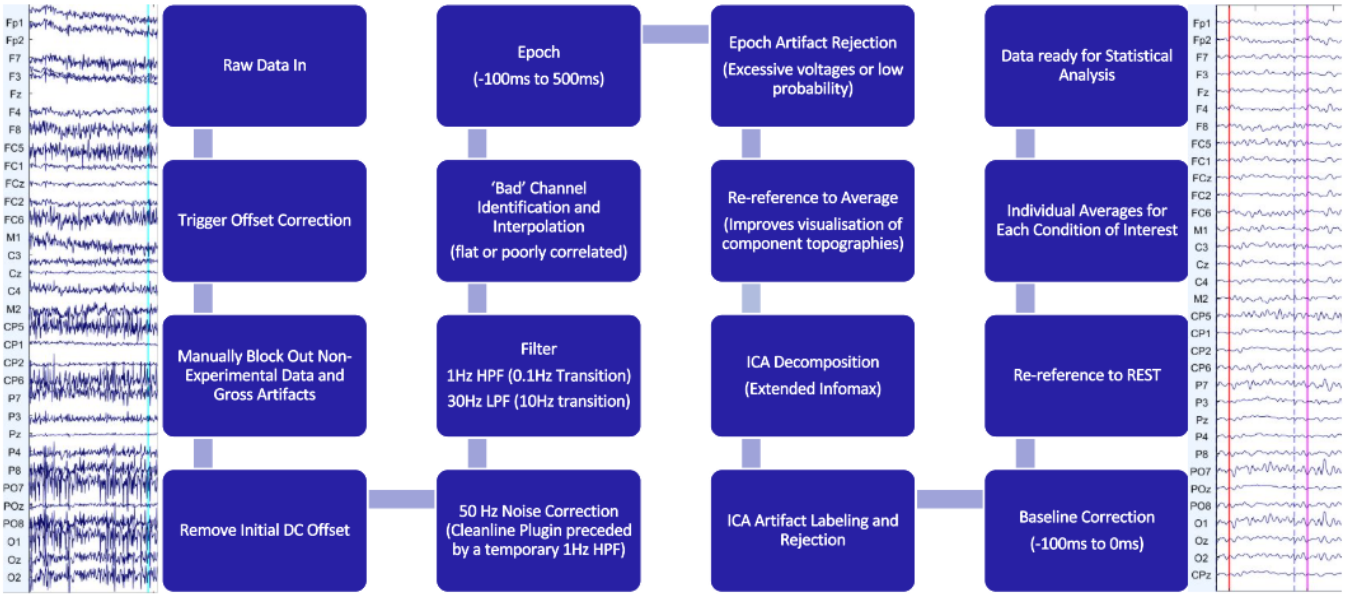

Figure 2: The custom-made EEG data pre-processing pipeline. The side panels show an example of raw (left) and processed (right) EEG data.

towards participants). This led to the introduction of bodily touch in the subsequent experiment (described next) as an effective method to elicit emotional responses [11].

3 Experiment 2: Embodied Affective Social Interaction in Early Face Perception in Social Anxiety

Experiment 2 uses a more direct and scalable social interaction with VCs, i.e., having a handshake. This requires attentional focus on VCs and tends to elicit stronger emotional responses, especially anxious feelings in people with Socia Anxiety (SA) [24]. Hence, we employ an anxiety-provoking scenario of meeting authority figures [26] and requires handshake with a male boss showing positive, negative or neutral facial expressions (Valence factor). We expect to see differences among the three valence conditions in ERP components related to face and emotion perception, such as P1, N170/VPP, EPN, P2, P3 and LPP. Moreover, correlations between these effects and SA scores will help confirming biomarkers of social interaction perception. Given that physiological responses are intensified in SA [9], HR, HRV, and EDA will also be recorded.

3.1 Stimuli and System Architecture

The experiment will be conducted using a Meta Quest Pro HMD using Unreal Engine 5.3. It simulates participants, as employees, meeting with their male bosses. The VE is a typical individual office environment (3m × 4m) with office furniture, with floor and ceiling having similar colour and texture to the experimental lab to create smooth transition from the lab to VE. VCs are created using Metahuman Creator (Epic Games) showing various ethnic backgrounds. We validated the VCs' face and body images on perceived trustworthiness, attractiveness, approachability, humanlikeness, and dominance. 24 VCs rated highest in dominance, humanlikeness and attractiveness and lowest in approachability and trustworthiness are selected for the experiment . To enhance dominance [5], VCs are set to be taller than the average height of males in the UK. For better presence and embodiment, participants' own VCs are partially visible (torso, arms, hands and legs in first-person view, skin tone adjusted). Inverse kinematics is used to replicate the participants' hand and body movements.

The VCs show positive, negative and neutral facial expressions generated with a face motion capture app (Live Link Face) based on actors' real-time facial expressions. Each VC has unique facial expressions for increased believability. VCs maintain eye contact with participants. The participants will initiate handshakes with VCs who will respond.

EmotiBit (physiological measures) and Etee VR controllers (TG0) are additional sensors to the previously presented Experiment 1 setup (Figure 1). Etee VR controllers (replacing traditional VR controllers) are clipped onto the participants' hands without the need to grip, allowing an open-hand position for handshake. Their vibration feedback also enhances the handshake's believability. The built-in sensors track individual fingers' pressure exerted onto the controller. All data streams (EEG, physio, pressure, HMD/Etee position and orientation) are recorded and synchronised in real-time using LSL.

3.2 Experimental Protocol and Analysis

Participants will be recruited to the inclusion criteria: 1) aged 18-50, 2) identifying themselves as women, 3) normal or corrected-to-normal vision, 4) no diagnosis of SA disorder, 5) no neurological condition, 6) no movement impediments. Only participants who identify themselves as women will be recruited because women typically experience a higher level of SA [23] and exhibit more pronounced SA symptoms, particularly when facing men [2]. Each participant will choose one of six skin tone options for their own VC, and complete the Liebowitz Social Anxiety Scale [17] and Social Interaction Anxiety Scale [19] (assessing their fear of and avoidance towards social interactions).

There are three experimental blocks (positive, negative, and neutral), the order of which is counterbalanced across participants. Each block has 80 trials (8 VCs repeating 10 times each). As in Experiment 1, each trial starts in darkness with the X-O alignment to trigger the lights-on moment to reveal the VE and a standing VC 1.2 m away, displaying the corresponding facial expression of the block. After 1000 ms, a doorbell sound signals the participant to approach and initiate handshake (using the preferred hand) with the VC. The participant then returns to their initial starting point and facing direction to trigger the lights-off (trial end). After each block, the participants will rate the overall emotion of the VCs in that block with Self-Assessment Manikin (SAM) [4] in the VE, then sit down and watch a 2 min neutral video to return to their emotional

baseline before the next block.

The EEG recording and analysis pipelines remain the same as those in Experiment 1. Amplitudes of various ERP components will be analysed with one-way ANOVA (factor: Valence) and ANCOVA (controlling for physiological responses including HR, HRV, and EDA, which might confound ERP results). We will also correlate the valence effects in ERPs against SA scores to help identifying the biomarkers of social interaction perception.

4 General Discussion

The paper introduces a flexible methodological framework for investigating early face perception during social interactions in VEs. It has heightened ecological validity, especially in the highly believable handshake, improving the generalisability of the findings to real-life scenarios. The design can be easily adjusted to meet different needs with different social interaction styles and intensity. For instance, the handshake can be initiated by participants or VCs. The VCs' valence and arousal levels can also be scaled. This makes the design highly adaptable.

The framework integrates multiple synchronised data streams via LSL with a high precision in temporal alignment. This is crucial for EEG analysis which is of high temporal resolution. The results have demonstrated that the system can separate different social behaviours in VE, and are moving towards identifying biomarkers of social perception in SA. This advances beyond traditional methods by incorporating embodied and believable social interactions in VR, and fills the methodology gap of social perception research.

One possible limitation of the proposed methodology is the use of the less realistic "lights-on" moment. This approach is a compromise for better ERP data quality, because visual onsets produce the strongest ERPs [25]. Recently, fixation-related ERP has shown good data [10]. It is worth adapting these methods to VR-EEG research despite the challenge of the much lower signal-to-noise ratio of EEG in VR. The inclusion of only women participants in Experiment 2 may appear as a limiting factor; however, previous findings that women show stronger SA symptoms [23] support our experimental setup towards identifying biomarkers of SA. After this initial step, we will include both women and men to enhance the generalisability.

References

[1] P. L. Anderson, M. Price, S. M. Edwards, M. A. Obasaju, S. K. Schmertz, E. Zimand, and M. R. Calamaras. Virtual reality exposure therapy for social anxiety disorder: a randomized controlled trial. *Journal of consulting and clinical psychology*, 81(5):751, 2013. 2

[2] M. Asher, A. Asnaani, and I. M. Aderka. Gender differences in social anxiety disorder: A review. *Clinical psychology review*, 56:1–12, 2017. 3

[3] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1:71–81, 2009. 2

[4] A. Betella and P. F. Verschure. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one*, 11(2):e0148037, 2016. 3

[5] N. M. Blaker and M. van Vugt. The status-size hypothesis: How cues of physical size and social status influence each other. *The psychology of social status*, pp. 119–137, 2014. 3

[6] M. Bortoletto, J. B. Mattingley, and R. Cunnington. Action intentions modulate visual processing during action perception. *Neuropsychologia*, 49(7):2097–2104, 2011. 2

[7] S. Bouchard, S. Dumoulin, G. Robillard, T. Guitard, E. Klinger, H. Forget, C. Loranger, and F. X. Roucaut. Virtual reality compared with in vivo exposure in the treatment of social anxiety disorder: a three-arm randomised controlled trial. *The British Journal of Psychiatry*, 210(4):276–283, 2017. 2

[8] E. W. Carr, A. Kever, and P. Winkielman. Embodiment of emotion and its situated nature. In *The Oxford Handbook of 4E Cognition*. Oxford University Press, 09 2018. 2

[9] D. M. Clark and A. Wells. A cognitive model of social phobia. In R. G. Heimberg, M. R. Liebowitz, D. A. Hope, and F. R. Schneier, eds., *Social phobia: Diagnosis, assessment, and treatment*, pp. 69–93. The Guilford Press, 1995. 3

[10] F. Degno and S. P. Liversedge. Eye movements and fixation-related potentials in reading: a review. *Vision*, 4(1):11, 2020. 4

[11] A. Gallace and C. Spence. The science of interpersonal touch: an overview. *Neuroscience & Biobehavioral Reviews*, 34(2):246–259, 2010. 3

[12] A. R. Geiger and B. Balas. Not quite human, not quite machine: Electrophysiological responses to robot faces. *bioRxiv*, pp. 2020–06, 2020. 2

[13] M. Gonzalez-Franco, E. Ofek, Y. Pan, A. Antley, A. Steed, B. Spanlang, A. Maselli, D. Banakou, N. Pelechano, S. Orts-Escolano, et al. The rocketbox library and the utility of freely available rigged avatars. *Frontiers in virtual reality*, 1:561558, 2020. 2

[14] D. Kuleli, F. Charles, L. Guo, L. Vuillier, C. H. Liu, N. Gregory, and X. He. Exploring influence of social anxiety on embodied face perception during affective social interactions in vr. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, pp. 1–5, 2024. 2

[15] D. Kuleli, P. Tyson, N. H. Davies, and B. Zeng. Examining the comparative effectiveness of virtual reality and in-vivo exposure therapy on social anxiety and specific phobia: A systematic review & meta-analysis. *Journal of Behavioral and Cognitive Therapy*, 35(2):100524, 2025. 2

[16] M. Lataifeh, I. Afyouni, Z. A. S. Shaduly, A. Abdulkarim, and N. Ahmed. An adaptive multimodal framework for designing intelligent virtual agents in mixed reality. IUI '25 Companion, p. 133–136. Association for Computing Machinery, New York, NY, USA, 2025. 1

[17] M. R. Liebowitz. Liebowitz social anxiety scale. *Journal of Anxiety Disorders*, 1987. 3

[18] S. J. Luck and N. Gaspelin. How to get statistically significant effects in any erp experiment (and why you shouldn't). *Psychophysiology*, 54(1):146–157, 2017. 2

[19] R. P. Mattick and J. C. Clarke. Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behaviour research and therapy*, 36(4):455–470, 1998. 3

[20] D. Roth, J.-L. Lugrin, D. Galakhov, A. Hofmann, G. Bente, M. E. Latoschik, and A. Fuhrmann. Avatar realism and social interaction quality in virtual reality. In *2016 IEEE virtual reality (VR)*, pp. 277–278. IEEE, 2016. 1

[21] G. A. Rousselet, M. J. Macé, and M. Fabre-Thorpe. Spatiotemporal analyses of the n170 for human faces, animal faces and objects in natural scenes. *Neuroreport*, 15(17):2607–2611, 2004. 1

[22] A. Shoa, R. Oliva, M. Slater, and D. Friedman. Sushi with einstein: Enhancing hybrid live events with llm-based virtual humans. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, pp. 1–6, 2023. 1

[23] C. L. Turk, R. G. Heimberg, S. M. Orsillo, C. S. Holt, A. Gitow, L. L. Street, F. R. Schneier, and M. R. Liebowitz. An investigation of gender differences in social phobia. *Journal of anxiety disorders*, 12(3):209–223, 1998. 3, 4

[24] F. H. Wilhelm, A. S. Kochar, W. T. Roth, and J. J. Gross. Social anxiety and response to touch: incongruence between self-evaluative and physiological reactions. *Biological psychology*, 58(3):181–202, 2001. 3

[25] G. F. Woodman. A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, & Psychophysics*, 72:2031–2046, 2010. 4

[26] Y. Xu, F. Schneier, R. G. Heimberg, K. Princisvalle, M. R. Liebowitz, S. Wang, and C. Blanco. Gender differences in social anxiety disorder: Results from the national epidemiologic sample on alcohol and related conditions. *Journal of anxiety disorders*, 26(1):12–19, 2012. 3

[27] D. Yao. A method to standardize a reference of scalp eeg recordings to a point at infinity. *Physiological measurement*, 22(4):693, 2001. 2