DEVELOPMENTS IN IMPLICIT MEASUREMENTS

EDITED BY: Alexander Toet, Anne-Marie Brouwer, Jaroslaw Jerzy Barski, Andrew Spink, Gernot Riedel and Raphaëlle N. Roy
PUBLISHED IN: Frontiers in Psychology, Frontiers in Human Neuroscience, Frontiers in Neuroergonomics and Frontiers in Neuroscience

frontiers Research Topics

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

DEVELOPMENTS IN IMPLICIT MEASUREMENTS

Topic Editors:
Alexander Toet, Netherlands Organisation for Applied Scientific Research, Netherlands
Anne-Marie Brouwer, Netherlands Organisation for Applied Scientific Research, Netherlands
Jaroslaw Jerzy Barski, Medical University of Silesia, Poland
Andrew Spink, Noldus Information Technology, Netherlands
Gernot Riedel, University of Aberdeen, United Kingdom
Raphaëlle N. Roy, Institut Supérieur de l'Aéronautique et de l'Espace (ISAE-SUPAERO), France

Table of Contents

The Shape of Water Stream Induces Differences in P300 and Alpha Oscillation

Noriaki Kanayama[1,2]*, Shumpei Mio[3], Ryohei Yaita[3], Takahiro Ohashi[3] and Shigeto Yamawaki[2]

[1]Human Informatics Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan, [2]Center for Brain, Mind and KANSEI Sciences Research, Hiroshima University, Hiroshima, Japan, [3]TOTO Limited, Research Institute, Chigasaki, Japan

Touching is a fundamental human behavior used to evaluate objects in the external world. Many previous studies have used tactile stimulation to conduct psychological and psychophysiological experiments. However, most of these studies used solid material, not water stream, as an experimental stimulus. To investigate water perception, or to easily control the temperature of an experimental stimulus, it is important to be able to control the water stimulus. In this study, we investigated the usability of water as an experimental stimulus for electroencephalography (EEG) experiments and report the basic EEG response to water stimulus. We developed a tactile stimulation device using a water stream to study EEG responses, with the ability to control the stimulus onset timing. As stimuli, we selected two types of water stream, normal and soft, based on a psychological experiment to confirm a difference of subjective feeling induced by these water streams. We conducted a typical oddball task using the two different water streams and recorded EEG waveforms from 64 electrodes while participants touched the water streams. We calculated P300 at the Pz electrode, alpha asymmetry at the frontal electrodes, and alpha suppression at the parietal area. As a result, we observed typical P300 differentiation based on the stimulus proportion (target 20% and standard 80%). We observed a weaker alpha suppression when participants touched the soft water stream compared to the normal shower. These results demonstrate the usability of water stream in psychophysiological studies and suggested that alpha suppression could be a candidate to evaluate comfort of water stream.

Keywords: water, EEG, P300-event related potential, alpha oscillations, touch

INTRODUCTION

Recent technologies for modulation of tactile experiences demonstrated that, in the very near future, tactile perception might be freely created to induce a specific affective response, in the same way that is currently possible with visual and auditory perception. For example, the softness experience could be modulated with vibrotactile stimuli (Hayward, 2008; Visell and Okamoto, 2014). These stimuli could not only be induced with typical solid materials but possibly with ultrasonic mid-air stimulation (Hoshi et al., 2009; Long et al., 2014).

A water stream is an ethologically important stimulus for inducing tactile perception on the surface of the human body. We often touch water in ordinary life, such as when washing one's face, hand washing, dishwashing, toothbrushing, bathing, and drinking. Some uses of water include cleaning one's body; in this regard, touching water can induce a comforting feeling. To elucidate the neural mechanisms underscoring the pleasure of touching, water may be a useful stimulation.

Experimental stimuli should be shared across laboratories to enhance replicability of results obtained from different cognitive neuroscience studies. For this reason, many researchers developed homemade devices for precise control of the experience of touching solid materials (McGlone et al., 2012; Muñoz et al., 2014; Kanayama et al., 2019). However, this is not always possible, as materials to be touched in an experiment vary greatly (Sakamoto and Watanabe, 2017).

The shape of a water stream could be altered by changing the shape of a faucet. Water stimulation can easily be reproduced in any location using an identical faucet shape. The different shapes of water stream could have different impacts on the subjective feeling of affective evaluation, including tactile comfort and richness of the water stream. Comfortable touching experiences induced by water stream could be modeled by affective/Kansei engineering (Nagamachi, 1995) using subjective sense of comfort about the water stream. Based on subjective reports of comfort about various water streams, the shape of a water stream can be optimized. However, subjective reports could be varied and modified by spontaneous appraisal, which makes stable modeling based on subjective reports challenging. Psychophysiological measurements, as an objective index, can support the model of subjective feelings of comfort (Balters and Steinert, 2017). To this end, we explored electroencephalography (EEG) components as a reflection of tactile comfort during perception of touching a water stream. However, EEG components are also not always more reliable than subjective reports, as they can vary depending on mental health, age, and current arousal state. Measurement of implicit and automatically activated evaluation of affective experience could potentially provide further insight not obtained from explicit reflective and subjective verbal reports.

To relate subjective comfort from touching a water stream, we used two typical EEG components of alpha band oscillation. Alpha oscillations are related to emotional responses, as the generation of alpha oscillations stems from regional cerebral blood flow in various cortical areas including emotion-related areas, such as the amygdala, basal prefrontal cortex, and insula (Sadato et al., 1998).

The first component we focused on in this study was alpha suppression, which is typically observed after stimulus presentation. Previous EEG studies using tactile stimuli have repeatedly showed alpha/beta band suppression after stimulation (van Ede et al., 2011; Singh et al., 2014). Singh et al. (2014) reported that beta band mu-suppression and beta-band oscillation showed a relationship with tactile caressing and subjective ratings of pleasantness. Bauer et al. (2006) reported that parieto-occipital alpha/beta suppression was modulated by spatial attention. This component over the parieto-occipital distribution has been observed in response to audiovisual stimulation, suggesting that this component is generated by cognitive processes independent of sensory modality (Schelenz et al., 2013). Some studies demonstrated that this component was more strongly suppressed when participants perceived a negative emotional stimulus compared to when perceiving a neutral stimulus (Jessen and Kotz, 2011; Swingle, 2013). Kostyunina and Kulikov (1996) reported that a decrease in alpha power was related to negative emotional states of fear and sorrow. These findings spurred us to measure participants' emotional states using alpha suppression. We hypothesized that a softer stream would elicit weaker alpha suppression compared to a normal shower stream.

The second component we focused on was alpha asymmetry. This EEG component has repeatedly been observed when viewing emotional movie clips (Killeen and Teti, 2012; Lopez-Duran et al., 2012; Meyer et al., 2014; Zhao et al., 2018). Evidence that frontal asymmetry is modulated by tactile comfort is scarce; furthermore, physical distress during sleep can have an impact on this component (Flo et al., 2011). Thus, we aimed to investigate whether touching a soft water stream would induce alpha asymmetry. We hypothesized that a softer stream would elicit greater alpha asymmetry compared to a normal shower stream.

The difficulty of controlling a water stream has typically precluded the use of water as a tactile stimulus, especially in neuroscience and psychophysiology research. To our knowledge, no psychophysiological study using a water stream has been undertaken to date. Here, we examined using a water stream as stimulus in a typical oddball paradigm and confirmed that a water stream can induce a typical P300 component. By this investigation, we demonstrate the usability of water in psychophysiological studies, alongside solid and visual stimuli.

MATERIALS AND METHODS

Pilot Study for Stimulus Selection

For this study, it was necessary to be able to compare whether or not pleasant touch was elicited in response to different shapes of water streams. To select stimuli, we conducted a pilot experiment on subjective evaluation of water streams. Fifteen people (five females, 10 males) participated in the pilot experiment. The mean age of participants was 31.67 years ($SD = 5.95$; range = 25–46). Twelve of these people also participated in the main experiment with EEG measurements. The pilot experiment was performed more than a year before the EEG experiment.

Participants evaluated five types of water streams (**Table 1**) in terms of "richness" and "comfort" after touching the water streams freely. For "richness," participants rated whether they have "touched a high-quality thing:" 1 for very low quality, 9 for very high quality, and 5 for average quality. For "comfort," participants rated whether they "felt comfort" when touching the water stream: 1 for high discomfort, 9 for high comfort, and 5 for neither. The difference between "comfort" and "richness" was the evaluation target. Comfort was the evaluation of the participant's state, which is induced by the touch of water. Richness was the evaluation of the water stream itself. We could

TABLE 1 | Physical properties of water streams used in the pilot experiment.

Shape of flow	Comfort mean (SD)	Richness mean (SD)	Amount of water flowing (L/min)	Hole diameter (mm)	Hole numbers	Aperture area (mm^2)
Normal shower	1.73 (0.88)	1.60 (0.91)	3.0	0.6	18	5.09
Soft flow	7.47 (1.36)	7.67 (0.82)	3.0	1.5	19	33.56
Laminar	7.00 (1.07)	6.80 (1.15)	3.0	12.5	1	122.66
Modified Soft flow	7.13 (1.49)	7.60 (1.18)	3.0	1.5	18	31.81
Numerous hole shower	5.87 (1.46)	1.73 (0.88)	3.0	-	82	43.36
Flow with aerator	5.00 (0.00)	5.00 (0.00)	3.0	11.9	1	111.16

assume, for example, a case in which a participant feels richness on touching the water stream but does not feel comfort. To make the evaluation easier, participants were instructed to touch a flow with an aerator before the experiment started, an item not included in the stimulus list. Participants were instructed to consider this as the baseline and evaluate all other stimuli relative to the baseline score of 5. During the experiment, all water streams were hidden by a wall to exclude visual effects from the stimuli. The amount of water flowing was identical for each water stream, which was monitored by a flowmeter (sensor: FD-MH10A, display: FD-MA1A, KEYENCE, Osaka, Japan) and controlled within a range of ± 0.2 L/min. The temperature of the water was kept within 23–25°C. Other physical properties of the two stimuli are listed in **Table 1**.

To confirm pairs of water streams with different subjective feelings, we statistically analyzed comfort and richness scores for all pairs of stimuli. We observed significant differences in comfort scores between the pairs of normal shower vs. soft flow ($t_{(14)} = -14.94$, $p < 0.001$, Cohen's $d = -3.86$), normal shower vs. laminar ($t_{(14)} = -14.19$, $p < 0.001$, Cohen's $d = -3.66$), normal shower vs. modified soft flow ($t_{(14)} = -11.34$, $p < 0.001$, Cohen's $d = -2.93$), Normal shower vs. numerous hole shower ($t_{(14)} = -9.75$, $p < 0.001$, Cohen's $d = -2.52$) and soft flow vs. numerous hole shower ($t_{(14)} = -3.36$, $p < 0.05$, Cohen's $d = -0.87$). We also observed significant differences in richness scores between pairs of normal shower vs. soft flow ($t_{(14)} = -25.25$, $p < 0.001$, Cohen's $d = -6.52$), normal shower vs. laminar ($t_{(14)} = -14.38$, $p < 0.001$, Cohen's $d = -3.71$), normal shower vs. modified soft flow ($t_{(14)} = -18.33$, $p < 0.001$, Cohen's $d = -4.73$), normal shower vs. numerous hole shower ($t_{(14)} = -10.81$, $p < 0.001$, Cohen's $d = -2.79$), soft flow vs. numerous hole shower ($t_{(14)} = -7.74$, $p < 0.001$, Cohen's $d = -2.00$), and modified soft flow vs. numerous hole shower ($t_{(14)} = -6.17$, $p < 0.01$, Cohen's $d = -1.59$). All p-values were corrected by Bonferroni method. Based on these results, we selected normal shower and soft flow as the paired stimuli for the EEG experiment.

Participants

Thirty healthy individuals (15 females, 15 males) participated in the EEG experiment. Mean age of participants was 31.28 years ($SD = 5.60$; range = 24–46). All participants were employees of TOTO Limited. None of the participants were informed of the experimental aims before participation. Based on the ethical guideline of TOTO Limited, all participants provided oral informed consent before participation.

Materials

Stimulation Device

A stimulation device using a water stream was developed for this experiment (**Figure 1A**). Two faucets with different shapes were attached to an actuator (MISUMI, RS102, Tokyo, Japan). One faucet produced a normal shower water stream, whereas the other produced soft flow of water. The water stream from the latter faucet was straighter and induced softer touch sense compared to that of the normal shower (details in subsequent section). The device could randomly switch between the two types of faucets for water stimulation during experimentation for each trial, which enabled us to conduct a discrimination task, typically termed the "oddball task," using water stream as a stimulus. Water streams were continuously delivered to the faucets in parallel. Solenoid valves (TOTO Limited, THE13, Fukuoka, Japan) attached to the faucet controlled whether the water stream was released or stopped. A controller (MISUMI, EXRS-C1, Tokyo, Japan) of the actuator monitored the position of the faucet. When the actuator moved the faucet to the programmed position, transistor–transistor logic signal was sent to the EEG amplifier as a trigger signal of touch onset using a programmable logic controller (KEYENCE, CPU: KV-7300, KV-B8XTD, Osaka, Japan). The aforementioned system was attached to a stage made of aluminum frames. A basin with a drain hose was attached to the stage and covered by a thin plastic wall preventing water splashes and visualization of the water stream.

Water Stream Stimuli

Two types of water stream stimuli (normal shower and soft flow shower) were used in the EEG experiment. Before the stimulation, the water stream was released by controlling a solenoid valve. The actuator position was placed at a distance from the finger such that the water stream or any water splash would not contact the finger. The actuator was controlled to move from the right side to the participant's left hand. When the actuator reached the intended position where the water stream first touched the right end of the left index finger, the transistor–transistor logic signal as stimulus onset trigger was sent to the EEG amplifier. This onset trigger timing was confirmed by a pilot experiment using an acrylic rod instead of a finger. The position of the actuator, where the water stream first touched the right end of the acrylic rod, was searched in 0.1-mm steps using a high-speed camera (Photron, FASTCAM Mini UX100, Tokyo, Japan; **Figure 1B**). The discrepancy between timing of the

FIGURE 1 | Stimulation device and water stream used in the electroencephalography (EEG) experiment. **(A)** Schematic representation of the stimulation device. **(B)** Pictures of touching a water stream with an acrylic rod (substitute for a finger) captured by a high-speed camera. **(C)** Schematic illustrations of arm and finger fixtures.
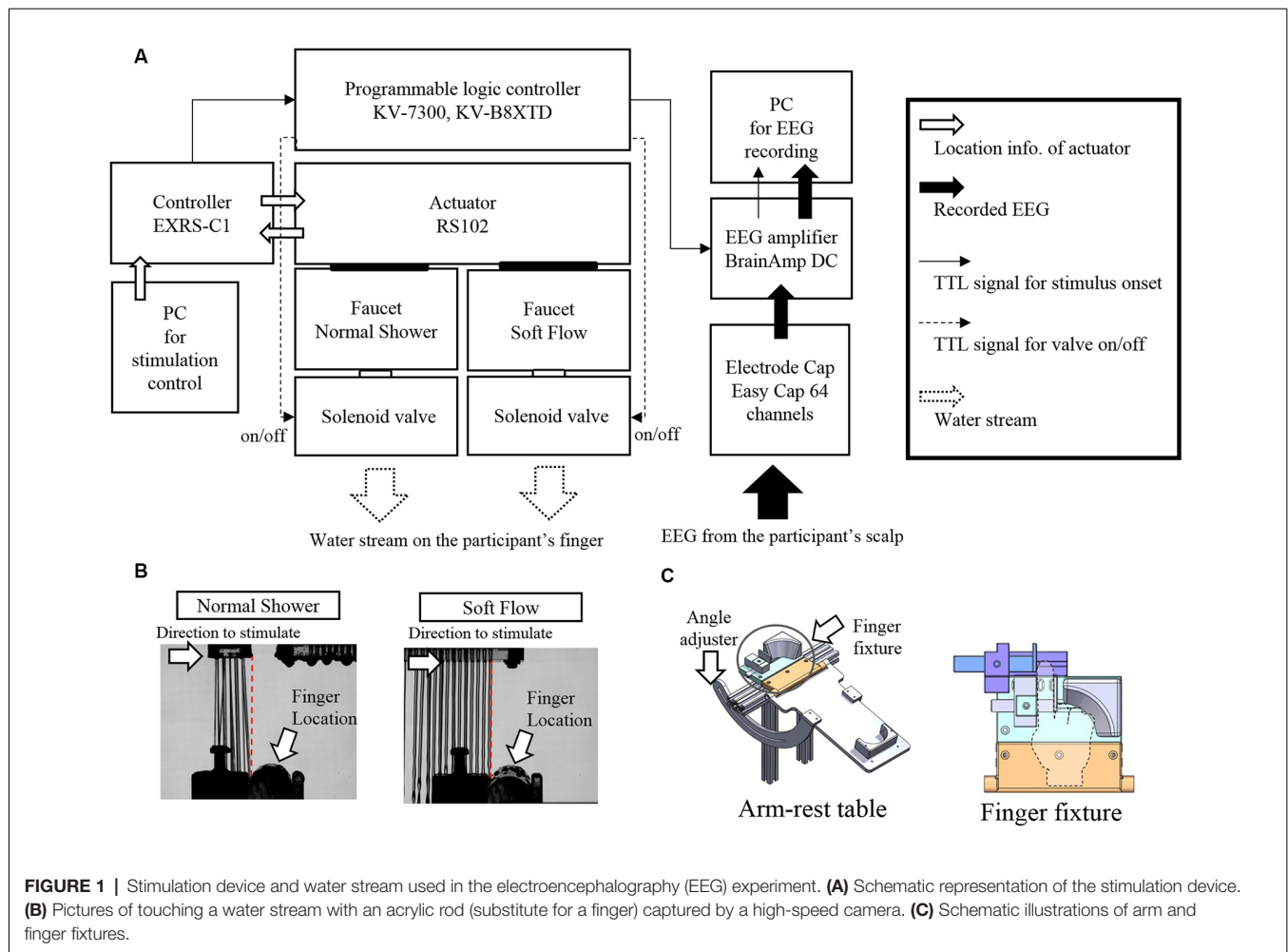
trigger and that of water touching was kept below 1 ms (0.1 mm).

Experimental Procedures

Participants were fitted with an EEG electrode cap and received instructions about the task. In the experimental room, participants sat on a comfortable chair and placed their arm and hand on the arm-rest table (**Figure 1C**) of the stimulation device in a palm-up posture. The angle of the table could be adjusted seamlessly for each participant to feel comfortable. To constrain incidental movement of the participants' finger, participants were required to place the index finger of their left hand into a finger fixture. For adjustment of the program to control the actuator based on individual finger shape, we simulated the finger touch position using a scale. Trigger timing was adjusted by the finger touch position.

Before the beginning of experimentation, participants wore canal-type earphones (ZERO AUDIO, DX211-WB, Kyoto, Japan) to hear white noise as a masking sound. Participants heard the sound induced by movement of the actuator without touching the water stream. Volume level of the white noise was adjusted such that participants could not hear device-induced noise. The maximum volume was 70 dB for adjustments.

Participants were instructed to view the center of a box which occluded sight of the water stream in front of their face and to maintain their posture. Participants were encouraged to interfere with movement of the index finger of the left hand. Participants were required to count the touches of the water streams during the experimental block as target stimuli. The order of targets was counterbalanced across all participants. At the beginning of the experimental block, after a 10-s blank period, one of the water streams touched the index finger's inner surface of the left hand.

Touching continued for 2,000 ms. After the stimulation offset, a resting period was inserted. The interstimulus interval was varied from 5,450 to 5,700 ms by 50-ms steps. The average interstimulus interval was 5,575 ms. In one block, participants touched the water stream 60 times. The standard stimulus was presented 48 times, whereas the target stimulus was presented 12 times. After all trials in one block, participants were required to report the number of target stimuli based on their internal count. The next block started with a short break. Each session comprised five blocks without changing the target stimulus. Breaks were provided as per participants' request. The second

session started with the altered target stimulus. Participants perceived the water stream 600 times in total, divided into 10 blocks.

EEG Recordings

EEG waveforms were recorded using the BrainAmp DC amplifier (Brainproducts, GmbH, Munich, Germany). The sampling rate was 1,000 Hz. The reference electrode was placed on the nose tip, and the grand electrode for the scalp EEG was placed on the back of the neck. The 64-channel active electrodes were distributed over the whole scalp according to the 10-10 international standard position using an EEG recording cap (EasyCap, GmbH, Herrsching, Germany). Horizontal electrooculograms (EOGs) were recorded by bipolar surface electrodes placed on the left and right outer canthus. The vertical EOG was recorded by bipolar surface electrodes placed above and under the left eye (eyebrow). The hardware filter settings were as follows: the low frequency cutoff was 0.016 Hz (time constant, 10 s), whereas the high frequency cutoff was 1,000 Hz. The impedance at each electrode was maintained at least below 50 kΩ and typically below 10 kΩ. The grand average of impedance across all participants and electrodes at the beginning of recording was 6.24 kΩ ($SD = 1.85$).

EEG Data Analysis

EEG waveforms were analyzed using EEGLAB eeglab14_1_1b (Delorme and Makeig, 2004) under the Matlab R2018a (MathWorks, Natick, MA, USA). First, the recorded waveforms were digitally filtered by 1 Hz high-pass and 40 Hz low-pass filters. Continuous EEG waveforms were segmented from −600 to 1,200 ms after stimulus touch timing. Independent component analysis (ICA) decomposition with the Infomax method was applied to the segmented data. Based on the component waveforms, artifact-contaminated trials were discarded according to maximum and minimum amplitude, mean trial probability, kurtosis value, and spectrum power. Details are described in the **Supplementary Material**. Another ICA was applied to trial-rejected datasets, and the component waveforms were obtained anew. Dipole locations for all components were estimated using dipfit2 (EEGLAB plug-in using FieldTrip toolbox functions; Oostenveld et al., 2011).

For event-related potential (ERP) analysis to differentiate P300 deflection by standard and target stimulus presentation, we conducted cluster-based IC rejection to discard the eye-movement related EEG deflection. After this rejection, we selected the Pz electrode and calculated ERP waveforms. For statistical analysis, the max value during 300–800 ms across all trials was used for P300 amplitude for each participant. Wilcoxon-signed rank test was conducted, and the p values were corrected using the Bonferroni method.

Regardless of target/standard stimulus, we visualized brain activation related to the difference between normal and soft water streams, and calculated alpha power. For further analyses, we modified the previously mentioned data set by merging all trials into one condition of water stream shape regardless of presentation probability (target/standard). For alpha asymmetry, we calculated the values for the stimulated period (2 s) using the traditional formula [ln(F4) − ln(F3) (Allen et al., 2004)]

and statistically tested the difference between the two streams using Wilcoxon-signed rank test with Bonferroni correction of the p-value. For alpha suppression, we calculated event-related spectrum power (ERSP). Using the merged dataset, we conducted ICA clustering analysis and derived 10 clusters. Based on the hypothesis of alpha wave distribution, we focused on the parieto-occipital cluster. ERSPs at the cluster were calculated for −100 to 800 ms and 3–30 Hz. The baseline was a period between −300 and −50 ms. The difference of ERSPs between two water streams was statistically tested using Monte Carlo permutation statistics with cluster correction (channel neighbor parameters: triangulation, clustering method: max-sum) implemented in FieldTrip toolbox (Oostenveld et al., 2011).

Follow-Up Water Flow Evaluation Task

All participants were recruited for a follow-up evaluation task more than 6 months after participation in the EEG experiment. Participants sat on a comfortable chair in a light room and adopted an identical posture to that in the EEG experiment. Participants received identical water streams on the index finger of their left hand. One of the two types of water streams was alternately delivered to the participant's finger. The order of stimuli was counterbalanced across participants. Participants were required to rate the water streams after each stimulation in terms of "richness" and "comfort," as conducted in the EEG experiment. In total, 10 stimulations and ratings were conducted. The averaged score across all ratings was used as an evaluation of each participant.

RESULTS

P300 to the Water Stream Stimulus

The grand averaged ERP waveforms at Pz electrode for all conditions are plotted in **Figure 2**. Based on visual inspection of the waveforms, ERP waveforms to the target stimulus were greater than those to standard stimuli regardless of the shape of the water stream. The time period in which the waveforms to target stimuli were greater than those to standard stimuli was from about 300–800 ms. This time period is typical for P300 to low-frequency target stimuli. Statistical analysis of the maximum values during that time period revealed significant differences between target and standard stimuli for both blocks (soft target/normal standard condition, $Z = 2.62$, $p < 0.05$, $r = 0.34$, normal target/soft standard, $Z = 3.92$, $p < 0.01$, $r = 0.51$).

Subjective Rating for Each Shower Shape

Participants evaluated the richness and comfort of perceived water streams by normal and silky shower. The averaged values of subjective ratings in the follow-up evaluation of water streams are illustrated in **Figure 3**. Both subjective rating scores were higher when participants received the soft water stream than when they received the normal shower. The difference in scores between two types of water streams were statistically analyzed using a t-test with corrected p-values (Bonferroni method p-value correction). Significant differences in both ratings were observed ($t_{(29)} = 6.01$, $p < 0.01$, Cohen's $d = 1.09$ for comfort; $t_{(29)} = 11.05$, $p < 0.01$, Cohen's $d = 2.02$ for richness).
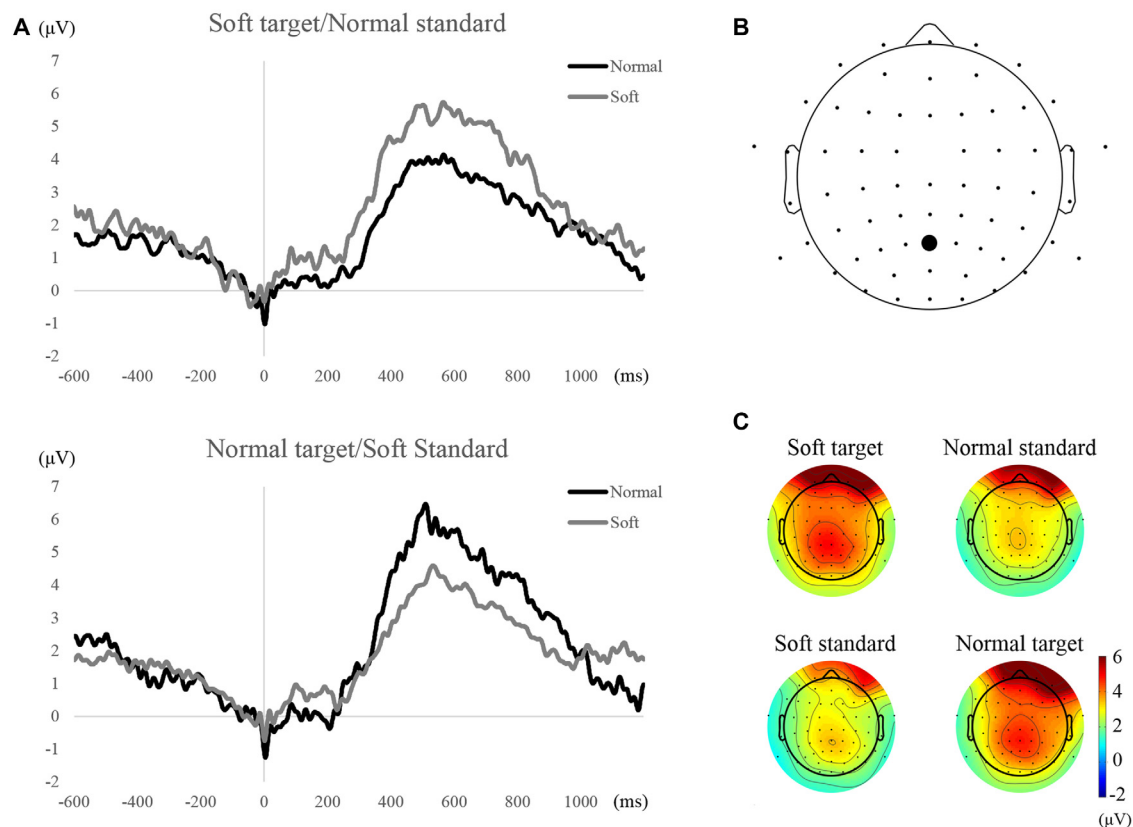
FIGURE 2 | Waveforms and topographical maps of P300 components. **(A)** Grand averaged event-related potential (ERP) waveforms at the Pz electrode for each condition and stimulation. **(B)** The location of Pz electrode. **(C)** Scalp distribution at amplitudes averaged across 300–800 ms time period.

FIGURE 3 | Subjective rating scores for water stream perception averaged across 23 participants. Error bars indicate the standard errors for each condition.
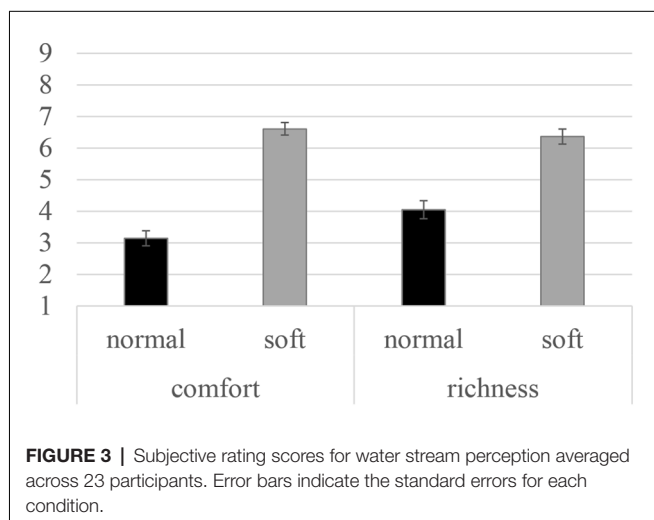
Alpha Asymmetry for Each Shower Shape

No significant difference between different water streams was observed in alpha asymmetry index. The alpha asymmetry for each water stream was almost identical (-0.05 for normal

shower vs. -0.06 for soft flow, $Z = 0.61$, n.s.; **Figure 4**, right panel). Almost one-third of the participants showed the opposite (positive) value to the trend (negative value), which suggests great interindividual variability in alpha asymmetries (**Figure 4**, left panel).

Alpha Suppression for Each Shower Shape

We observed alpha suppression at the parieto-occipital cluster after touching a water stream (**Figure 5**). The centroid of the parietal cluster was located at X, 26; Y, -41; Z, 49. More than half of the ICs' dipole positions (41/76) are in the parietal area, including postcentral, superior parietal, precuneus, and angular gyrus (**Table 2**). Statistical analysis revealed that alpha suppression to the normal shower was significantly greater than that to the silky shower stimulus. The significant time frequency window was 8–17 Hz and 200–600 ms.

Correlation Analyses of Subjective Ratings and EEG Component

We calculated the Spearman rho correlation coefficient to analyze the relationship between EEG activity (alpha asymmetry and alpha suppression) and subjective rating scores (comfort and richness) obtained by follow-up experiments. No significant

FIGURE 4 | Alpha asymmetries in individuals for each water stream and averaged value for each stimulus. In the left panel, alpha asymmetry in each individual is illustrated. The y-axis denotes participant numbers, whereas the x-axis denotes the value of alpha asymmetries (microV^2). In the right panel, the alpha asymmetries are averaged across all participants. The y-axis denotes the value of alpha asymmetries (microV^2).

FIGURE 5 | (A) Event-related spectrum power (ERSP) for each condition, dipole locations, and scalp topography of the target cluster. **(B)** Schematic representation of differentiated alpha suppression by touch with normal and silky water streams. The illustrated waveforms are typical waveforms at the occipital electrodes before and after stimulation with normal and silky water streams.

correlations were observed. Scatter plots are illustrated in **Figure 6**.

DISCUSSION

In this study, we aimed to test two hypotheses: first, whether a water stream as a tactile stimulus could induce P300 and whether amplitudes were different between target and standard stimulation in the typical oddball task; second, whether tactile comfort induced by a water stream could be indexed by alpha oscillation. Significant modulation of P300 was observed in this experiment, suggesting that the methodology used in this experiment was effective for electrophysiological studies using water streams. Alpha suppression was significantly modulated by the shape of the water stream, which differentiated subjective reports of comfort and richness, implying that alpha oscillation could be involved in affective processing when touching water streams.

TABLE 2 | Dipole characteristics for each independent component involved in target cluster 3.

AAL name of the dipole position	N of ICs	Average RV	SD of RV
Postcentral_R	14	10.12	4.19
Precuneus_R	12	7.63	3.03
Parietal_Sup_R	9	10.15	4.55
Angular_R	6	8.97	4.18
Precentral_R	5	13.03	2.23
Parietal_Inf_R	5	5.41	2.41
Cingulum_Mid_R	4	7.82	3.99
Cingulum_Post_R	3	6.64	4.36
Precuneus_L	3	9.43	2.15
Cingulum_Mid_L	2	7.02	0.24
Cingulum_Post_L	2	3.69	0.87
Paracentral_Lobule_R	2	4.93	1.17
Occipital_Sup_R	2	5.54	0.49
Frontal_Sup_R	2	13.81	2.47
Insula_R	1	9.71	0
Temporal_Mid_R	1	6.92	0
Paracentral_Lobule_L	1	3.22	0
Frontal_Mid_R	1	10.85	0
Occipital_Mid_R	1	2.92	0

For example, at the top of the rows, the dipole positions of 14 independent components (ICs) fall in the area named "Postcentral_R" (the right postcentral gyrus) based on the automated anatomical labeling (AAL) atlas 3. The average residual variance of dipoles across all ICs of "Postcentral_R" was 10.12, and the standard deviation was 4.19. AAL, automated anatomical labeling; IC, independent component; RV, residual variance.

P300 amplitude was greater for the target stimulus regardless of the shape of the water stream. This suggested that P300 could be elicited by the shape of the water stream in a similar manner as that of typical visual and/or tactile oddballs using surface toughness. Scalp topography was also very typical of P300 component during the oddball task. In the topographical map of P300, positive deflections were also observed over the frontal area. This is possibly related to eye movement-related noise deflections, as we did not instruct participants to refrain from eye blinking when touching the water stream. We tried to discard eye movement-related EEG waveforms using ICA. We successfully detected and excluded a cluster distributed over

the frontal pole before ERP calculation. However, the frontal distributed positive deflection remained in the resulting EEG data, suggesting that there were difficulties in ICA-based removal of EOG-related activity from EEG waveforms recorded in free eye movement conditions. In these results, we observed clear boundaries between frontal and parietal positive deflections, which supported the conclusion that the P300 difference at Pz was not contaminated by EOG-related positive deflections.

Furthermore, we calculated the alpha asymmetry index for each EEG waveform to different water streams; however, there was no significant difference between them. Subjective reports have revealed that the comfort and richness evaluation for soft flow was significantly higher than that for normal shower, which led to the assumption that alpha asymmetry during perception of soft flow is greater than that during normal shower because negative emotional states induce greater alpha asymmetry than do normal and positive emotional stimuli. However, many studies using alpha asymmetry have adopted longer time periods for analysis. For example, the waveforms when viewing video clips ranging from 6 s to several tens of minutes were used for calculating spectrum powers (Davidson et al., 1999; Allen et al., 2001; Meyer et al., 2014). In this study, water stream stimulation continued for only 2 s, which may have been insufficient for alpha asymmetry analysis. In the future, experiments with longer stimulation periods are required to clarify whether alpha asymmetry can be an emotional index of water stream perception.

Finally, we assumed that alpha suppression at the parietal area could be modulated by different stimuli. Compared to normal shower, soft flow showed significantly lower alpha suppression of a cluster at the parietal area. The relationship between emotional response and alpha suppression has previously been reported (Kostyunina and Kulikov, 1996; Swingle, 2013), which suggests that alpha suppression could be related to emotional responses to water stream stimulation. The cluster involved many ICs with dipole locations in the postcentral, superior parietal, precuneus, and angular gyrus, which suggested that the primary

FIGURE 6 | Scatter plots of subjective rating scores and EEG responses to water streams. **(A)** Alpha asymmetry. **(B)** Alpha suppression.
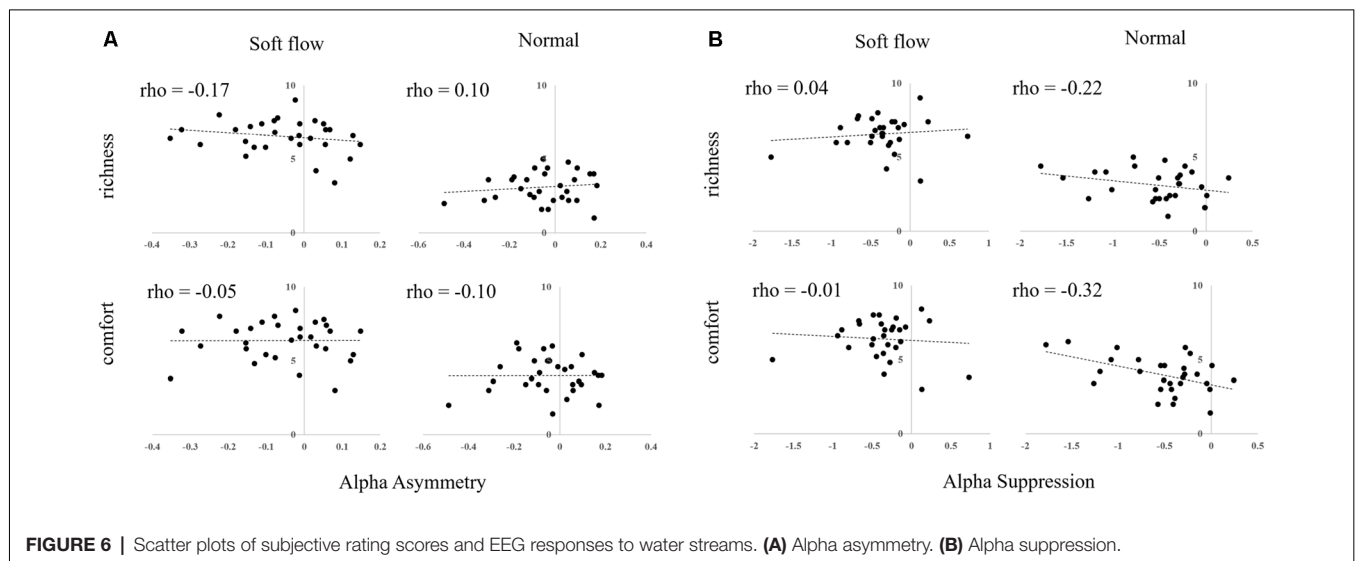
and secondary somatosensory area and the sensory association area could be related to the water pleasantness perception. However, because the correlations between alpha suppression and subjective evaluations were not significant, we were unable to conclude that alpha suppression was directly related to the feeling of comfort or richness when perceiving a water stream. In addition, we have to note that nonparametric statistical tests were used as we could not presume a normal distribution; therefore, we could only determine that the rank (high or low amplitude) was consistent across participants. Previous studies have demonstrated that alpha suppression was greater when participants paid attention to the stimulus (Siegel et al., 2008; Wyart and Tallon-Baudry, 2008), especially using visuospatial attention. In this study, exogenous covert attention is a candidate for modulation of alpha suppression after touching a water stream. Future research is required to clarify the relationships among the amount of induced attention, alpha suppression, and tactile comfort.

To conclude, we have demonstrated that water stream could be used as tactile stimulation for electrophysiological experiments. During a typical oddball task, water stream-induced P300 components and a difference between target and standard stimuli were observed. Furthermore, we demonstrated that less alpha suppression was related to comfort/richness of a water stream, although no direct correlation between them was observed. Future research is required to clarify the relationships among the amount of induced attention, alpha suppression, and tactile comfort. Our findings may be used as an electrophysiological index of affective evaluation for engineering.

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available. As a rule of the Ethical Committee,

we cannot disclose the raw biological data acquired from human participants.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by TOTO Limited, Research Institute, Ethical Committee. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

NK, SM, RY, TO and SY have read, discussed and approved the manuscript for submission. NK, SM, RY, TO and SY discussed to design the behavioral and EEG experiment. NK, SM, RY and TO designed and developed the device. NK, SM and RY collected the data. NK wrote the article.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2019.00460/full#supplementary-material.

REFERENCES

Allen, J. J. B., Coan, J. A., and Nazarian, M. (2004). Issues and assumptions on the road from raw signals to metrics of frontal EEG asymmetry in emotion. *Biol. Psychol.* 67, 183–218. doi: 10.1016/j.biopsycho.2004.03.007

Allen, J. J. B., Harmon-Jones, E., and Cavender, J. H. (2001). Manipulation of frontal EEG asymmetry through biofeedback alters self-reported emotional responses and facial EMG. *Psychophysiology* 38, 685–693. doi: 10.1017/s0048577201991255

Balters, S., and Steinert, M. (2017). Capturing emotion reactivity through physiology measurement as a foundation for affective engineering in engineering design science and engineering practices. *J. Intell. Manuf.* 28, 1585–1607. doi: 10.1007/s10845-015-1145-2

Bauer, M., Oostenveld, R., Peeters, M., and Fries, P. (2006). Tactile spatial attention enhances gamma-band activity in somatosensory cortex and reduces low-frequency activity in parieto-occipital areas. *J. Neurosci.* 26, 490–501. doi: 10.1523/JNEUROSCI.5228-04.2006

Davidson, R. J., Coe, C. C., Dolski, I., and Donzella, B. (1999). Individual differences in prefrontal activation asymmetry predict natural killer cell activity at rest and in response to challenge. *Brain, Behav. Immun.* 13, 93–108. doi: 10.1006/brbi.1999.0557

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Flo, E., Steine, I., Blågstad, T., Grønli, J., Pallesen, S., and Portas, C. M. (2011). Transient changes in frontal alpha asymmetry as a measure of emotional and physical distress during sleep. *Brain Res.* 1367, 234–249. doi: 10.1016/j.brainres.2010.09.090

Hayward, V. (2008). A brief taxonomy of tactile illusions and demonstrations that can be done in a hardware store. *Brain Res. Bull.* 75, 742–752. doi: 10.1016/j.brainresbull.2008.01.008

Hoshi, T., Iwamoto, T., and Shinoda, H. (2009). "Non-contact tactile sensation synthesized by ultrasound transducers", in *Proceedings-3rd Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems* (Washington, DC: IEEE Computer Society).

Jessen, S., and Kotz, S. A. (2011). The temporal dynamics of processing emotions from vocal, facial and bodily expressions. *Neuroimage* 58, 665–674. doi: 10.1016/j.neuroimage.2011.06.035

Kanayama, N., Hara, M., Watanabe, J., Kitada, R., Sakamoto, M., and Yamawaki, S. (2019). Controlled emotional tactile stimulation during functional magnetic resonance imaging and electroencephalography. *J. Neurosci. Methods* 327:108393. doi: 10.1016/j.jneumeth.2019.108393

Killeen, L. A., and Teti, D. M. (2012). Mothers' frontal EEG asymmetry in response to infant emotion states and mother-infant emotional availability, emotional experience and internalizing symptoms. *Dev. Psychopathol.* 24, 9–21. doi: 10.1017/S0954579411000629

Kostyunina, M. B., and Kulikov, M. A. (1996). Frequency characteristics of EEG spectra in the emotions. *Neurosci. Behav. Physiol.* 26, 340–343. doi: 10.1007/bf02359037

Long, B., Seah, S. A., Carter, T., and Subramanian, S. (2014). Rendering volumetric haptic shapes in mid-air using ultrasound. *ACM Trans. Graph.* 33, 1–10. doi: 10.1145/2661229.2661257

Lopez-Duran, N. L., Nusslock, R., George, C., and Kovacs, M. (2012). Frontal EEG asymmetry moderates the effects of stressful life events on internalizing symptoms in children at familial risk for depression. *Psychophysiology* 49, 510–521. doi: 10.1111/j.1469-8986.2011.01332.x

McGlone, F., Olausson, H., Boyle, J. A., Jones-Gotman, M., Dancer, C., Guest, S., et al. (2012). Touching and feeling: differences in pleasant touch processing between glabrous and hairy skin in humans. *Eur. J. Neurosci.* 35, 1782–1788. doi: 10.1111/j.1460-9568.2012.08092.x

Meyer, T., Quaedflieg, C. W. E. M., Giesbrecht, T., Meijer, E. H., Abiad, S., and Smeets, T. (2014). Frontal EEG asymmetry as predictor of physiological responses to aversive memories. *Psychophysiology* 51, 853–865. doi: 10.1111/psyp.12230

Muñoz, F., Reales, J. M., Sebastián, M. Á., and Ballesteros, S. (2014). An electrophysiological study of haptic roughness: effects of levels of texture and stimulus uncertainty in the P300. *Brain Res.* 1562, 59–68. doi: 10.1016/j.brainres.2014.03.013

Nagamachi, M. (1995). Kansei engineering: a new ergonomic consumer-oriented technology for product development. *Int. J. Ind. Ergonomics* 15, 3–11. doi: 10.1016/0169-8141(94)00052-5

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011:156869. doi: 10.1155/2011/156869

Sadato, N., Nakamura, S., Oohashi, T., Nishina, E., Fuwamoto, Y., Waki, A., et al. (1998). Neural networks for generation and suppression of alpha rhythm: a PET study. *Neuroreport* 9, 893–897. doi: 10.1097/00001756-199803300-00024

Sakamoto, M., and Watanabe, J. (2017). Exploring tactile perceptual dimensions using materials associated with sensory vocabulary. *Front. Psychol.* 8:569. doi: 10.3389/fpsyg.2017.00569

Schelenz, P. D., Klasen, M., Reese, B., Regenbogen, C., Wolf, D., Kato, Y., et al. (2013). Multisensory integration of dynamic emotional faces and voices: method for simultaneous EEG-fMRI measurements. *Front. Hum. Neurosci.* 7:729. doi: 10.3389/fnhum.2013.00729

Siegel, M., Donner, T. H., Oostenveld, R., Fries, P., and Engel, A. K. (2008). Neuronal synchronization along the dorsal visual pathway reflects the focus of spatial attention. *Neuron* 60, 709–719. doi: 10.1016/j.neuron.2008.09.010

Singh, H., Bauer, M., Chowanski, W., Sui, Y., Atkinson, D., Baurley, S., et al. (2014). The brain's response to pleasant touch: an EEG investigation of tactile caressing. *Front. Hum. Neurosci.* 8:893. doi: 10.3389/fnhum.2014.00893

Swingle, P. G. (2013). The effects of negative emotional stimuli on alpha blunting. *J. Neurother.* 17, 133–137. doi: 10.1080/10874208.2013.785797

van Ede, F., de Lange, F., Jensen, O., and Maris, E. (2011). Orienting attention to an upcoming tactile event involves a spatially and temporally specific modulation of sensorimotor alpha- and beta-band oscillations. *J. Neurosci.* 31, 2016–2024. doi: 10.1523/jneurosci.5630-10.2011

Visell, Y., and Okamoto, S. (2014). "Vibrotactile sensation and softness perceptionpitch," in *Multisensory Softness: Springer Series on Touch and Haptic Systems*, ed. M. Di Luca (London: Springer), 31–47.

Wyart, V., and Tallon-Baudry, C. (2008). Neural dissociation between visual awareness and spatial attention. *J. Neurosci.* 28, 2667–2679. doi: 10.1523/JNEUROSCI.4748-07.2008

Zhao, G., Zhang, Y., and Ge, Y. (2018). Frontal EEG asymmetry and middle line power difference in discrete emotions. *Front. Behav. Neurosci.* 12:225. doi: 10.3389/fnbeh.2018.00225

Corrigendum: The Shape of Water Stream Induces Differences in P300 and Alpha Oscillation

Noriaki Kanayama[1,2]*, Shumpei Mio[3], Ryohei Yaita[3], Takahiro Ohashi[3] and Shigeto Yamawaki[2]

[1] Human Informatics Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan, [2] Center for Brain, Mind and KANSEI Sciences Research, Hiroshima University, Hiroshima, Japan, [3] TOTO Limited, Research Institute, Chigasaki, Japan

A Corrigendum on

The Shape of Water Stream Induces Differences in P300 and Alpha Oscillation
by Kanayama, N., Mio, S., Yaita, R., Ohashi, T., and Yamawaki, S. (2020). Front. Hum. Neurosci. 13:460. doi: 10.3389/fnhum.2019.00460

In the original article, there was a mistake in **Figure 6** as published. **In the figure, the locations of the scatter plot were wrong. In original Figure 6B, a plot for Softflow x comfort (left bottom) was a plot for Normal x richness (right up).** The corrected **Figure 6** appears below.

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

FIGURE 6 | Scatter plots of subjective rating scores and EEG responses to water streams. **(A)** Alpha asymmetry. **(B)** Alpha suppression.
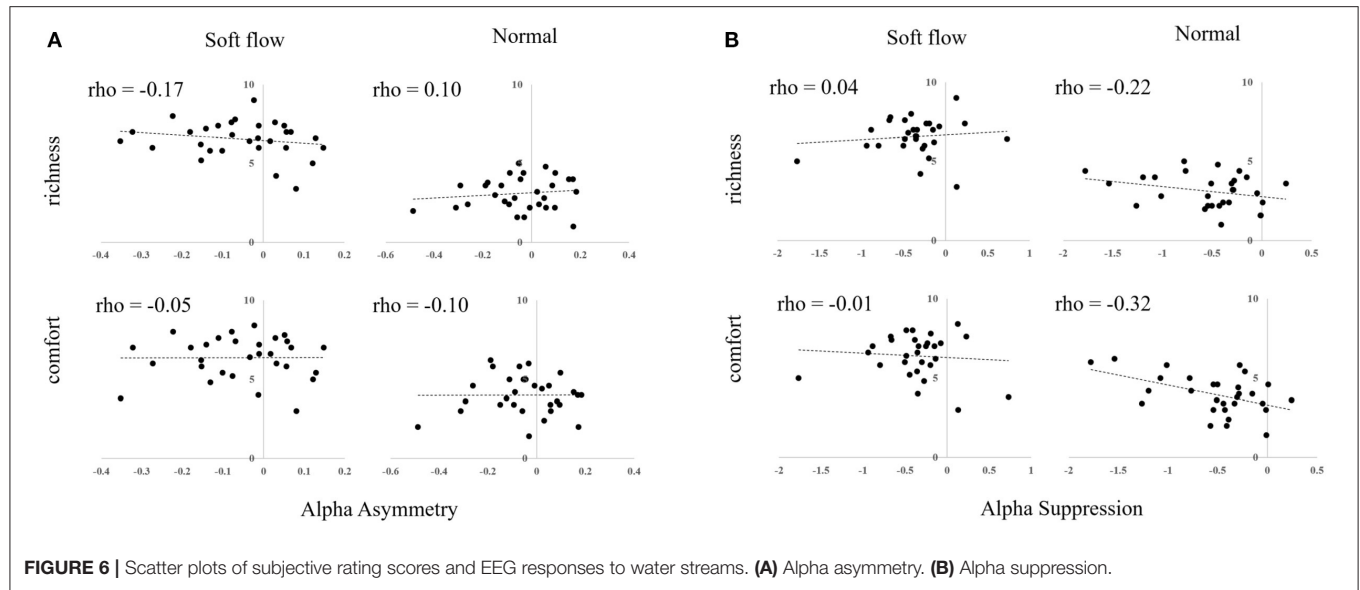
Gaze Behavior in Social Fear Conditioning: An Eye-Tracking Study in Virtual Reality

Jonas Reichenberger*, Michael Pfaller and Andreas Mühlberger

Department of Clinical Psychology and Psychotherapy, Institute of Psychology, University of Regensburg, Regensburg, Germany

The vigilance-avoidance hypothesis of selective attention assumes that socially anxious persons initially direct their attention toward fear-related stimuli and subsequently avoid these social stimuli to reduce emotional distress. New technical developments provide tools to implicit measure overt attention on fear-related stimuli via eye-tracking in ecological valid virtual environments presented via a head-mounted display. We examined in 27 low (LSA) and 26 high socially anxious (HSA) individuals fear ratings, physical behavior (duration of approach), hypervigilance (time to first fixation), and attentional avoidance (count of fixations) toward virtual female and male agents (CS) during social fear conditioning (SFC) and extinction in virtual reality (VR). As hypothesized, generally SFC was successfully induced and extinguished concerning the fear ratings. Our findings partly support the vigilance-avoidance hypothesis as HSA directed especially at the first half of the fear acquisition their initial attention more at CS+ than CS− agents, and avoided subsequently the CS+ more than the CS− agents during the fear acquisition. In contrast, in LSA participants initial and sustained attention did not differ between CS+ and CS− agents during fear acquisition. We conclude that HSA individuals guide their initial attention to emotionally threatening stimuli and subsequently avoid the threatening stimuli to possibly reduce their emotional distress, whereas LSA individuals regulate themselves less in their (fear) responses during SFC. Measuring implicit gaze behavior within a well-controlled virtual environment is an interesting innovative tool to in deeply investigate the impact of attention on emotional learning processes.

Keywords: social anxiety disorder, social fear conditioning, vigilance-avoidance hypothesis, eye-tracking, virtual reality

INTRODUCTION

A lot of research has investigated selective attention that describes social anxiety to conceive the architecture of information processing in social anxiety disorder (SAD). The most popular theoretical models of SAD are the cognitive model from Clark and Wells (1995) and the cognitive-behavioral model from Rapee and Heimberg (1997). Both models comprise biases in information

processing that develop and/or maintain SAD and implicate the importance of selective attention (e.g., vigilance, avoidance) in social anxiety (Schofield et al., 2012).

Clark and Wells (1995) suggest that central factors of information processing in SAD include increased self-focused attention, safety behavior, and problematic anticipatory and subsequent processing. For example, avoiding attention toward feared stimuli during social interactions serves as safety behavior to regulate the internal distress (e.g., reducing eye contact), to avert feared negative situations (e.g., assessments), and/or to avoid real or perceived feared social appraisal of others as an attempt of self-regulation. Accordingly, conscious attentional avoidance of salient feared stimuli is active during social evaluations and is assumed to maintain SAD.

In comparison, Rapee and Heimberg (1997) focus more on vigilance as selective attention toward emotionally negative information in SAD. The authors emphasize that socially anxious persons utilize an exaggerated attentional allocation toward any sign of impending negative feedback of others, which can lead to biased estimations of a more threatening social environment and that a wisp of negative evaluation will be perceived rapidly and reinforce the threatening negative self-evaluation (Heimberg et al., 2014).

Chen and Clarke (2017) report in their review that there are different results in empirical research among socially anxious humans with attentional bias to relevant emotionally threat stimuli (Gilboa-Schechtman et al., 1999; Perowne and Mansell, 2002; Lange et al., 2011; Çek et al., 2016; Lazarov et al., 2016) and who also exhibited attentional avoidance toward socially feared information (Chen et al., 2002; Wieser et al., 2009b; Singh et al., 2015; Shechner et al., 2017). Mogg and Bradley (2002) suggest that vigilant patterns of attention may be located in initially attentional processing. Furthermore, Chen and Clarke (2017) conclude that there is an association between social anxiety and vigilance (e.g., hyperscanning the environment) as well as attentional avoidance (e.g., reducing eye contact) according to the information processing of emotionally social stimuli.

These mixed empirical results caused the formulation of a vigilance-avoidance hypothesis of selective attention (Bögels and Mansell, 2004; Wieser et al., 2009a,b; Chen and Clarke, 2017). It assumes that socially anxious humans guide their initial attention to emotionally threatening information (hypervigilance) and avoid the negative information subsequently (attentional avoidance) to reduce emotional distress (Bögels and Mansell, 2004; e Claudino et al., 2019).

In recent years the rising use of eye-tracking technology offered novel insights into diverse aspects of selective attention that are responsible in the etiology and maintenance of SAD (Lange et al., 2011; Schofield et al., 2012; Lazarov et al., 2016; Chen and Clarke, 2017). Furthermore, patterns of gaze behavior could be used as sensitive index of fear learning (Hopkins et al., 2015). Using an eye-tracking device allows us to directly and continuously measure visual selective attention in real time. This methodology registers the precise location of eye movement and gaze behavior over the course of time as a relatively naturalistic assessment of attention without an explicit response of the participant. Bögels and Mansell (2004) described in their review

a lot of paradigms like the emotional Stroop task (Maidenberg et al., 1996; Becker et al., 2001), visual search task (Rinck et al., 2003), modified dot-probe task (Mansell et al., 2002; Mogg et al., 2004), and eye-tracking tasks (Garner et al., 2006) to investigate hypervigilance and/or avoidance in the information processing of selective attention in social anxiety.

Relating to the nature of SAD, one major deficiency of these experimental paradigms is that participants observe stimuli with a low ecological validity (e.g., images of faces or words on a screen). Consequently, more research is necessary to validate how socially anxious persons respond to more ecological valid socially relevant stimuli. Empirical studies on gaze behavior in social interactions (e.g., interviews or having a speech causing fear of negative evaluation) are rare and have methodical shortcomings, e.g., less valid and reliable measurement of dependent variables and reduced control of independent variables in real interactions (Mühlberger et al., 2008; Wieser et al., 2009a).

Several empirical studies with an innovative technology such as virtual reality (VR) investigated social interaction processes with good experimental control, high ecological validity, low time costs, and highly valid assessments of dependent variables like physical and gaze behavior. Further advantages are the systematic and independent manipulations of eye-gaze directions or gender of the interaction partner as well as the presented social situations, all of which are either very costly or difficult to control in vivo. Most importantly, eye and head movements or interpersonal space between the participant and the counterpart can simply be recorded with the use of eye- and head-tracking within VR (Mühlberger et al., 2008; Wieser et al., 2009a,b; Dechant et al., 2017; Reichenberger et al., 2019). This innovative technology gives us the opportunity to assess directly approach-avoidance behavior in social interactions as well.

Wieser et al. (2010) investigated the impact of sex, gaze, and interpersonal distance (0.5 m vs. 1.5 m) on social anxiety in women in a virtual social interaction scenario. Socially anxious women showed attentional avoidance in response to virtual male counterparts, which were farther away and showed a straight gaze. In addition, they revealed more backward movements of the head (avoidance behavior) toward male agents, regardless of the interpersonal space. Moreover, Dechant et al. (2017) reported that HSA presented more attentional avoidance than LSA participants in a virtual social interaction paradigm. As we know, avoidance behavior is a key element in the maintenance of anxiety and also in fear learning.

In our previous study, we investigated differences between the gender and the influence of female vs. male agents in $N = 60$ high and low socially anxious individuals concerning social fear conditioning (SFC) and extinction in VR (Reichenberger et al., 2019). This SFC paradigm used social interactions in a standardized and experimentally controlled way, and disorder-relevant US (spitting simulated by an aversive air blast and verbal rejection) to examine affective learning in social anxiety. We measured with an enhanced ecological validity the experience, psychophysiology, behavior and cognition in emotional learning processes corresponding on each level of emotional reactions. Besides the successfully induced and extinguished SFC, we could present higher social fear conditionability in women, but

participants reported no higher fear while approached male than female agents. Interestingly, we found enhanced fear responses in the fear-potentiated startle to male than to female agents which might indicate higher social fear conditionability toward male persons. In addition, high socially anxious women revealed more behavioral avoidance to male than to female agents during fear acquisition. In comparison, HSA men did not discriminate between male or female agents. We concluded that the gender of the participants rather affect reflective processes (e.g., reported fear as well as contingency and skin conductance response), while the gender of the virtual agents influence more automatic measures (e.g., fear-potentiated startle and behavioral avoidance). Besides these self-reported and psychophysiological measures, binocular gaze behavior was also continuously recorded.

In the current study, we analyzed the recorded physical and gaze behavior (hypervigilance and attentional avoidance) during the SFC paradigm in a subsample from Reichenberger et al. (2019). Therefore, we aim to investigate the effect of induced and extinguished social fear on physical behavior (duration of approach) as well as hypervigilance (time of the first fixation) and attentional avoidance (count of fixations) toward female and male agents in HSA and LSA students in the SFC paradigm in VR. Based on the aforementioned empirical results, we hypothesize that (1) hypervigilance and attentional avoidance for CS+ would increase compared to CS− during fear acquisition. (2) In addition, we hypothesize that HSA will show enhanced hypervigilance to CS+ compared to CS− and will avoid more the CS+ compared to CS− than LSA participants during fear acquisition. At least, we expected that hypervigilance and attentional avoidance for CS+ would return to baseline levels and won't differ to CS− during fear extinction.

MATERIALS AND METHODS

Participants

One hundred and eighty Psychology and Media Informatics Science students at the University of Regensburg filled in a pre-screening questionnaire consisting of demographic and exclusion criteria as well as social anxiety using the German version of the Social Phobia Inventory (SPIN; Stangier and Steffens, 2002). Exclusion criteria were age <18 or >55 years, an existing or former diagnosed neurological or mental disorder (excepting SAD), a current neurological, psychiatric or psychotherapeutic treatment, a history of psychotropic drug use, pregnancy or lactation, and an attendance in a previous SFC study (Reichenberger et al., 2019).

Based on our previous study investigating SFC and extinction in VR (Reichenberger et al., 2017) a sample size of 54 was estimated with G*Power 3.1.9.2 to detect medium effects ($f = 0.25$) at power = 0.95 and $\alpha = 0.05$. The current study consists of a subsample from Reichenberger et al. (2019). However, on the basis of eye-tracking failure during data acquisition, we had to exclude seven participants from the original sample of $n = 60$. Thus, the current study contained a sample of 53 undergraduate students (27 LSA: 51.85% female, aged between 18 and 43; and 26 HSA: 53.85% female, aged between 18 and

33). The SPIN cut-off value of 19 differentiates patients with SAD and healthy persons with a diagnostic accuracy of 79% (Connor et al., 2000). A successful segmentation into a HSA and LSA group was given, since HSA reveal significantly higher scores than LSA participants with quite large between group effect sizes in the SPIN and in the German version of the Social Interaction Anxiety Scale (SIAS; Stangier et al., 1999). Please see **Table 1** for the questionnaire data. In addition, the groups did not differ in age [$t(51) = 0.832$, $p = 0.409$] and sex ratio [HSA: 12 men; LSA: 13 men; $\chi^2(1, 53) = 0.021$, $p = 0.884$]. The final sample was free of any neurological or mental disorder (self-report) and had unimpaired or corrected hearing and vision. All of the students received course credit as reimbursement for their attendance. The Ethics Committee of the University of Regensburg approved the study.

Apparatus

The VR was displayed with the HTC VIVE head-mounted display (HMD; HTC Corporation, Taoyuan, Taiwan) and was generated via the Steam Source engine (Valve Corporation, Bellevue, WA, United States). The presented virtual environment was controlled by CyberSession Research 5.6 (VTplus GmbH, Würzburg, Germany). Eye-tracking was continuously recorded binocular with a sampling rate of 250 Hz by the SMI (SensoMotoric Instruments, Boston, United States) head gear integration into the HMD. The trackable field of view is 110° with an accuracy of typ. 0.2° for eye-gaze vectors. CyberSession Research submits the processing of eye-tracking data in log files. All of the sounds were presented over headphones (Sennheiser HD-215, Sennheiser electronic GmbH, Germany). The participants could move in the virtual environment using a joystick (Logitech Extreme 3D Pro Joystick, Logitech GmbH, Germany).

During the study participants were immersed into a virtual room, which was modeled after a corridor of the University of Regensburg (see **Figure 1A**). In all three phases (pre-acquisition, acquisition, and extinction), the starting position was at one end of the room and at the opposite end a female or male agent was presented. The agents moved their head and upper body slightly and gazed dynamically at the participant to appear alive (please

TABLE 1 | Questionnaire data.

	LSA (n = 27)		HSA (n = 26)				
	M	**SD**	**M**	**SD**	**t**	**p**	**d**
SPIN	10.8	3.82	27.3	7.06	−10.624	<0.001	2.98
SIAS	16.2	10.1	29.5	11.1	−4.585	<0.001	1.28
PRF-D	9.78	3.97	5.31	3.74	4.219	<0.001	1.18
SBS	22.6	8.18	27.9	6.51	−2.636	0.011	0.74
VIG	10.7	3.63	13.5	2.69	−3.007	0.004	0.84
CAV	10.4	3.63	9.15	2.63	1.392	0.170	0.25

Means (M) and standard deviations (SD) and also t- and p-values and Cohen's d as well as number of participants (n) are given for low (LSA) and high socially anxious (HSA) participants for questionnaire data. SPIN = Social Phobia Inventory; SIAS = Social Interaction Anxiety Scale; PRF-D = German Personality Research Form – Dominance Scale; SBS = Submissive Behavior Scale; VIG = Vigilance of the Mainz Coping Inventory; CAV = Cognitive Avoidance of the Mainz Coping Inventory.
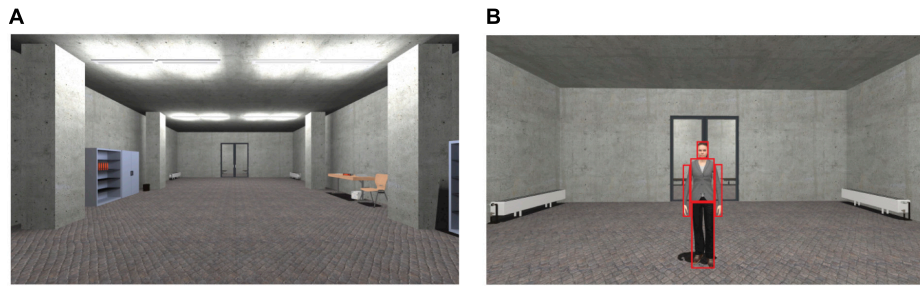
FIGURE 1 | Virtual environment and stimuli. **(A)** Starting point in the room in which the learning phases took place. **(B)** Social stimuli (agents) used for social fear conditioning (SFC) with the dynamic social area of interest (red rectangles). For interpretation of the references to color in this figure, the reader is referred to the web version of this article.

see Reichenberger et al., 2019). The task of the participants was to actively approach in all three phases the female and male agents until a distance of 30 cm where movement stopped. In 75% of the SFC trials female and male agents (CS+) were paired with an aversive unconditioned stimulus (US) existing of a sound of spitting attended by an air blast toward the right neck of the participant (5 bar, 10 ms) followed by the verbal rejection "Get lost!." The agents' facial expression was adapted to the spitting and verbal rejection. The other female and male agent (CS−) was not paired with an US. The pre-acquisition and extinction phase proceeded in exactly the same way as the fear conditioning, except for presenting no US and the appearance of a neutral agent (NS) of both genders in the extinction phase. The order of the agents was pseudo-randomized in each phase (see Shiban et al., 2015; Reichenberger et al., 2017, 2019).

Measures

At the beginning of the experiment, participants completed the following questionnaires: a socio-demographic questionnaire, the SPIN, the SIAS, the Dominance scale of the German Personality Research Form (PRF-D; Stumpf et al., 1985), the Submissive Behavior Scale (SBS; Allan and Gilbert, 1997), and the subtest Ego threat of the German version of the Mainz Coping Inventory (MCI; Krohne and Egloff, 1999) with the two coping strategies Vigilance (VIG) and Cognitive Avoidance (CAV). A detailed description of the questionnaires is given in our previous study (see Reichenberger et al., 2019).

During the VR, we quantified the experienced anxiety in regard of each agent with verbally ratings ("On a scale from 0 to 100, how intense is your anxiety during the presence of this person?") in the rating phases which followed each of the three phases. Participants approached each of the diverse agents until they reached the specific distance of 30 cm to the agents, lights faded out and they verbally rate their anxiety.

Besides these self-reported measures, physical behavior (duration to approach) and binocular gaze behavior was continuously recorded during the SFC paradigm. To ensure that the gaze behavior was measured correctly by the SMI integration eye-tracking device, a 5-point calibration was conducted before the learning phases in VR. The continuous gaze behavior was analyzed as the time to first fixation (hypervigilance) and the

mean percentage of the fixation counts (attentional avoidance) in predefined dynamic social areas of interest (AOI) for each agent (CS+, CS−, and NS) in each phase (pre-acquisition, acquisition, extinction). The social AOI was defined by a rectangle around the face and the body of each agent separately (see **Figure 1B**).

Procedure

The study consisted of briefing the participants in written form, signing the informed consent, filling out questionnaires, and the SFC paradigm in VR (see **Figure 2**). At the beginning of the VR, participants could explore the virtual environment and learn how to navigate with the joystick. Afterward, the recorded instruction "You will now meet several human beings. Please try to move directly toward the persons until they are right in front of you" was presented. Therefore, participants approached female and male agents actively until the specific distance of 30 cm to the agents, lights faded out, the participants were returned to the starting point and the next trial with a new agent at the opposite end of the room was presented. Which agent was presented as CS+ or CS− was balanced across participants.

In the pre-acquisition phase each female and male CS+/CS− agent was presented four times, resulting in 16 total trials and the US was not presented yet. Following this, the first rating phase took place for each female and male CS+/CS− agent.

The acquisition phase consisted of 32 trials in total. Each female and male CS+/CS− agent was presented eight times. Only the conditioned agent (CS+) was paired with 75% contingency of the US during the acquisition phase.

Following the fear conditioning, the second rating phase took place for each agent. After that, participants had a 5 min break where they took off the HMD and had the possibility to sit down and close their eyes.

After the break, the extinction phase took place exactly the same way as the fear acquisition, except for the absence of the US as well as the presentation of one additional female and one additional male NS, resulting in 48 total trials. At least, the third rating phase took place for each female and male CS+, CS−, and NS agent. A more detailed description of the procedure is given by Reichenberger et al. (2019).

Phase	Stimulus	Agents	US	Total Trials
Pre-Acquisition	CS+	4 x ♂ 4 x ♀	-	16
	CS-	4 x ♂ 4 x ♀		
Pre-Acquisition Rating	CS+	1 x ♂ 1 x ♀	-	4
	CS-	1 x ♂ 1 x ♀		
Acquisition	CS+	8 x ♂ 8 x ♀	75 %	32
	CS-	8 x ♂ 8 x ♀	-	
Post-Acquisition Rating	CS+	1 x ♂ 1 x ♀	-	4
	CS-	1 x ♂ 1 x ♀		
Extinction	CS+	8 x ♂ 8 x ♀	-	48
	CS-	8 x ♂ 8 x ♀		
	NS	8 x ♂ 8 x ♀		
Post-Extinction Rating	CS+	1 x ♂ 1 x ♀	-	6
	CS-	1 x ♂ 1 x ♀		
	NS	1 x ♂ 1 x ♀		

FIGURE 2 | Experimental procedure. As unconditioned stimulus (US) a sound of spitting attended by an air blast followed by the verbal rejection was applied. The order of the agents was pseudorandomized in each phase. CS+ = agent paired with aversive US; CS− = agent without aversive US; NS = agent without aversive US and only appearing during the extinction phase.

Data Reduction and Statistical Analyses

We utilized MATLAB 9.5 (MathWorks Inc., Germany) to analyze the behavioral data and SPSS 25.0 (IBM Corp., Armonk, NY, United States) to perform further analyses of the subjective and behavioral variables.

The mean fear rating for each female and male agent (CS+, CS−, NS) measured after each phase (pre-acquisition, acquisition, and extinction) were calculated. Investigating changes in fear ratings concerning SFC, two repeated-measures ANOVAs with the within-subject factors time (pre vs. post-acquisition for acquisition, and post-acquisition vs. post-extinction for extinction), stimulus (CS+ vs. CS−) and agent (female vs. male), and the between-subject factors gender (women vs. men) and social anxiety (LSA vs. HSA) were computed.

For the statistical analyses of the behavioral outcome variables, we defined that one trial consisted of the onset (as soon as the participant starts to approach using the joystick) and the offset (30 cm distance to the agent) in each phase.

For the physical behavior (duration of approach) as well as both gaze behavioral outcome variables, we computed means for the pre-acquisition phase, while the first four and the last four approaches in the acquisition and the extinction phase were calculated as the means of the beginning and the end of the acquisition as well as extinction phase, respectively.

Hypervigilance was measured by the time to first fixation on a social AOI in a trial. Since the duration of approach toward the agents was different between the phases and both groups (see **Figure 4**), attentional avoidance was calculated by the mean percentage of the count of fixations (number of fixations in the social AOI divided by the total number of fixations in the trial). The social AOI contain the face and the body of each agent in each approach trial. Fixations were defined based on a spatial (a diameter of 1° visual angle) and a temporal criterion (a minimum of 150 ms). Gaze behavior data were excluded if for more than 20% of the sampling points of a trial the eye-tracker couldn't identify the gaze direction of the participant.

Checking for potential distinctions in pre-acquisition, ANOVAs with the within-subject factors stimulus (CS+ vs. CS−) and agent (female vs. male) and the between-subject factors gender (women vs. men) and social anxiety (low vs. high) were calculated for each behavioral outcome variable. In order to analyze conditioning effects, repeated-measures ANOVAs with the within-subject factors time (beginning vs. end), stimulus (CS+ vs. CS−) and agent (female vs. male) and the between-subject factors gender (women vs. men) and social anxiety (LSA vs. HSA) were calculated for the acquisition and extinction phase. Testing for possible generalization effects, ANOVAs with the within-subject factors stimulus (CS+ vs. CS− vs. NS) and agent (female vs. male) and the between-subject factors gender (women vs. men) and social anxiety (low vs. high) were calculated at the beginning of the extinction phase as well.

All of the significant interactions were performed by separate follow-up ANOVAs or Student's t-tests. Partial η^2 (η_p^2) scores and Cohen's d served as indices of effect size. All statistical analyses utilized a $p < 0.05$ as level of statistical significance.

RESULTS

As expected, significant group differences (see **Table 1**) were found in social anxiety (SPIN), in anxiety of social interactional situations (SIAS), in dominant (PRF-D) and submissive behavior (SBS), and vigilance (VIG). However, groups did not differ in the total score of cognitive avoidance (CAV). As this is a subsample of the study published by Reichenberger et al. (2019), results in these analyses reflect the earlier results based on the whole sample.

Self-Report

Figure 3 displays that in each phase all fear ratings are higher for HSA than for LSA participants. After pre-acquisition, both stimuli are rated almost equal in both groups. After fear acquisition, the self-reported fear for CS+ agents is clearly higher than for CS− agents in both groups. After fear extinction, the ratings for CS+ agents decrease and the self-reported fear for

FIGURE 3 | Fear ratings (n = 53) for CS+, CS−, and NS in the three rating phases for low (LSA) and high socially anxious (HSA) participants. CS+ = agent paired with aversive unconditioned stimulus (US); CS− = agent without aversive US; NS = agent without aversive US and appearing only in the extinction phase; pre Acquisition = before acquisition phase; post Acquisition = after acquisition phase; post Extinction = after extinction phase. Mean fear ratings (0 = very low fear to 100 = very high fear) were given. Significant differences are indicated with an asterisk. Standard errors are presented by error bars.

CS− and NS agents do not distinguish after fear extinction, whereas the CS+ agents are rated slightly higher in both groups.

For fear acquisition, an ANOVA comparing the self-reported fear pre and post acquisition revealed a significant effect of Time × Stimulus × Gender × Social Anxiety (please see **Table 2** for all significant results of the ANOVA). Follow-up ANOVAs were performed for the HSA and LSA group separately to disentangle this fourfold interaction. For the HSA group, a significant effect of Time × Stimulus × Gender could be approved. Follow-up ANOVAs were conducted for HSA women and men separately. For HSA women, a significant interaction effect of Time × Stimulus could be found and the follow-up t-tests indicate that the self-reported fear increased significantly for CS+, $t(13) = -3.35$, $p = 0.005$, $d = 0.69$, but not for CS− ($p = 0.313$). In comparison, for HSA men a significant main effect of Time $F(1,11) = 4.83$, $p = 0.050$, $\eta_p^2 = 0.31$, Stimulus $F(1,11) = 4.88$, $p = 0.049$, $\eta_p^2 = 0.31$, but not for Time × Stimulus $F(1,11) = 2.83$, $p = 0.121$, $\eta_p^2 = 0.20$, was detected. Exploratory t-tests might give a first hint that also in men the self-reported fear might increase for CS+, $t(11) = -2.29$, $p = 0.043$, $d = 0.5$, but not for CS− ($p = 0.237$). For the LSA group, we found significant interaction effects for Time x Stimulus, and the follow-up t-tests show that the subjective fear increased significantly for CS+, $t(26) = -3.91$, $p < 0.001$, $d = 0.51$, but not for CS− ($p = 0.424$) as expected. In sum, the findings of the self-reported fear reveal that successful SFC took place like in the whole sample by Reichenberger et al. (2019), with the exception of HSA men.

Concerning fear extinction, an ANOVA comparing fear ratings pre and post extinction approved significant effects of Stimulus × Gender, and Time × Stimulus × Agent (please see **Table 2**). Follow-up ANOVAs were performed for female and male agents separately to unravel this threefold interaction. For female agents, a significant interaction effect of Time × Stimulus

TABLE 2 | Significant results of the ANOVA for fear ratings of the acquisition and extinction phase.

Effect	df	F	η_p^2	p
Acquisition				
Total				
Time	1, 49	18.9	0.28	<0.001
Stimulus	1, 49	23.5	0.32	<0.001
Gender	1, 49	5.13	0.10	0.028
Social Anxiety	1, 49	4.30	0.08	0.043
Time × Stimulus	1, 49	25.8	0.35	<0.001
Time × Stimulus × Gender × Social Anxiety	1, 49	4.13	0.08	0.048
LSA				
Time	1, 25	11.8	0.32	0.002
Stimulus	1, 25	12.0	0.32	0.002
Time × Stimulus	1, 25	10.7	0.30	0.003
HSA				
Time	1, 24	8.05	0.25	0.009
Stimulus	1, 24	11.6	0.33	0.002
Gender	1, 24	4.29	0.15	0.049
Time × Stimulus	1, 24	14.9	0.38	<0.001
Time × Stimulus × Gender	1, 24	5.62	0.19	0.026
HSA women				
Stimulus	1, 13	8.64	0.40	0.011
Time × Stimulus	1, 13	13.7	0.51	0.003
HSA men				
Time	1, 11	4.83	0.31	0.050
Stimulus	1, 11	4.88	0.31	0.049
Extinction				
Total				
Time	1, 49	28.0	0.36	<0.001
Stimulus	1, 49	28.6	0.37	<0.001
Agent	1, 49	4.77	0.09	0.034
Gender	1, 49	5.12	0.10	0.028
Stimulus × Gender	1, 49	4.88	0.09	0.032
Time × Stimulus	1, 49	16.6	0.25	<0.001
Time × Stimulus × Agent	1, 49	5.89	0.11	0.019
Female agents				
Time	1, 49	24.8	0.34	<0.001
Stimulus	1, 49	20.9	0.30	<0.001
Time × Stimulus	1, 49	23.3	0.32	<0.001
Male agents				
Time	1, 49	25.9	0.35	<0.001
Stimulus	1, 49	21.1	0.30	<0.001
Time × Stimulus	1, 49	5.73	0.11	0.021

df = degrees of freedom; η_p^2 = effect size; Time = pre vs. post acquisition for acquisition and post acquisition vs. post extinction for extinction; Stimulus = CS+ vs. CS−; Agent = female vs. male agent; Gender = women vs. men; Social Anxiety = LSA vs. HSA.

could be found and follow-up t-tests detect that the self-reported fear significantly decreased for the CS+, $t(52) = 5.57$, $p < 0.001$, $d = 0.53$, but not clearly for the CS−, $t(52) = 1.99$, $p = 0.051$, $d = 0.12$. For male agents, a significant interaction effect of Time × Stimulus was identified and follow-up t-tests highlight that the fear ratings of the CS+, $t(52) = 4.61$, $p < 0.001$, $d = 0.36$, and CS−, $t(52) = 3.23$, $p = 0.002$, $d = 0.31$, significantly

decreased. For the Stimulus × Gender interaction, follow-up *t*-tests reveal that women rated the CS+ significantly higher than men, *t*(45.7) = 2.48, *p* = 0.017, *d* = 0.68, however, there was no difference for the CS− (*p* = 0.096). The results of the fear ratings indicate that social fear extinction was also successful in the subsample as it was in the sample reported by Reichenberger et al. (2019). However, no generalization effect was found.

Physical Behavior

Figure 4 displays that at the beginning of the fear acquisition the LSA approach the CS− faster than the CS+, whereas the HSA don't differ between both agents. Interestingly, during fear acquisition HSA need slightly more time to approach the CS+ and CS− agents than the LSA group. During fear extinction, the duration of approach toward the agents differs not between both groups, besides the duration of approach toward the NS agent was shorter for the LSA than the HSA group at the beginning of the extinction.

Investigating differences in the pre-acquisition phase an ANOVA confirmed only a significant effect of Agent, $F(1,49) = 5.48$, $p = 0.023$, $\eta_p^2 = 0.10$, indicating a longer approach time toward male ($M = 9.08$, $SD = 0.25$) than female ($M = 9.01$, $SD = 0.20$) agents.

Relating to the fear acquisition, we detected no significant effect for Social Anxiety, $F(1,48) = 3.08$, $p = 0.086$, $\eta_p^2 = 0.06$, and Time × Stimulus × Social Anxiety, $F(1,48) = 3.02$, $p = 0.088$, $\eta_p^2 = 0.06$, but a significant effect for Agent × Gender × Social Anxiety, $F(1,48) = 4.38$, $p = 0.042$, $\eta_p^2 = 0.08$. Follow-up ANOVAs were conducted for both HSA and LSA participants separately, but no main or interaction effect reached significance. For LSA, the effect of Agent × Gender, $F(1,24) = 3.57$, $p = 0.071$, $\eta_p^2 = 0.13$, and Time × Stimulus, $F(1,24) = 4.13$, $p = 0.053$, $\eta_p^2 = 0.15$, narrowly missed significance. With regard to fear conditioning, exploratory *t*-tests might give a preliminary hint that at the first part of the fear acquisition the approaching time toward the CS+ might be higher than to the CS−, $t(25) = 2.091$, $p = 0.047$, $d = 0.53$, whereas at the end of the fear acquisition the difference of the approaching time seems vanished ($p = 0.415$).
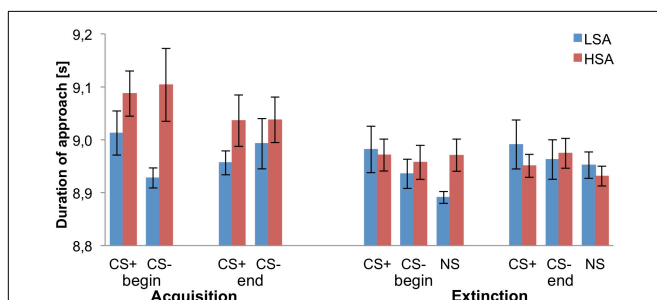
For fear extinction, a significant main effect of Time × Stimulus × Agent × Gender, $F(1,49) = 6.52$, $p = 0.014$, $\eta_p^2 = 0.12$, was found. Follow-up ANOVAs were performed for both genders separately, but no significant main or interaction effect was found. Furthermore, no generalization effect was found.

Hypervigilance

Figure 5 shows that at the beginning of the fear acquisition HSA fixate more early the CS+ than the CS− agent. In comparison, at the end of the fear acquisition HSA show a lower time of the first fixation to the CS− compared to the CS+ agent. The LSA show a small increase of the time of the first fixation toward both agents during fear acquisition. In the extinction phase, we can detect a decrease of the time of the first fixation toward the agents in both groups.

Proving differences before fear conditioning, an ANOVA confirmed no significant differences during pre-acquisition. For fear acquisition, an ANOVA revealed a significant effect of Time × Gender, $F(1,46) = 4.59$, $p = 0.037$, $\eta_p^2 = 0.10$, Time × Stimulus, $F(1,46) = 6.06$, $p = 0.018$, $\eta_p^2 = 0.12$, and Time × Stimulus × Social Anxiety, $F(1,46) = 4.06$, $p = 0.050$, $\eta_p^2 = 0.08$. Follow-up ANOVAs were conducted for the HSA and LSA group separately. For HSA, the interaction effect of Time × Stimulus, $F(1,24) = 12.73$, $p = 0.002$, $\eta_p^2 = 0.35$, reached significance. Follow-up *t*-tests indicate that at the beginning (first half) of the acquisition phase the time till the first fixation toward the CS+ was significantly lower than to the CS−, $t(25) = -2.377$, $p = 0.025$, $d = 0.18$, whereas at the end (second half) of the acquisition phase the time till the first fixation toward the CS+ was significantly higher than to the CS−, $t(25) = 2.687$, $p = 0.013$, $d = 0.21$. For LSA, no significant main or interaction effects were found.

Regarding fear extinction, an ANOVA detected significant effects of Time, $F(1,48) = 7.01$, $p = 0.011$, $\eta_p^2 = 0.13$, and Stimulus × Agent, $F(1,48) = 10.75$, $p = 0.002$, $\eta_p^2 = 0.18$. Follow-up *t*-tests reveal a significant lower time till the first fixation toward the male CS− compared to the female CS−, $t(52) = -3.641$, $p < 0.001$, $d = 0.29$, but no significant difference

FIGURE 4 | Mean duration of approach for low (LSA) and high socially anxious (HSA) participants (*n* = 53). CS+ = agent paired with aversive unconditioned stimulus (US); CS− = agent without aversive US; NS = agent without aversive US and only appearing during the extinction phase. Standard errors are presented by error bars.
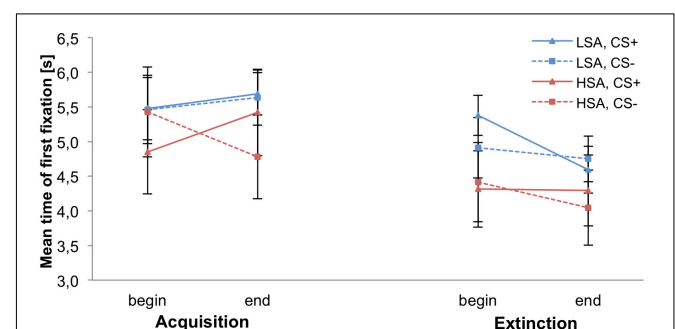
FIGURE 5 | Mean time of the first fixation within the social area of interest (agent) for low (LSA) and high socially anxious (HSA) participants (*n* = 53). CS+ = agent paired with aversive unconditioned stimulus (US); CS− = agent without aversive US. Standard errors are presented by error bars.

between the male and female CS+ ($p = 0.125$). However, no generalization effect was found.

Attentional Avoidance

As we can see, the mean count of fixations toward the face and the body of the agents (social AOI), as well as toward the environment is higher for the HSA compared to the LSA participants (see **Figure 6**). **Figure 7** displays the mean percentage of the count of fixations that contain the given social AOI (agent) for LSA and HSA participants during fear acquisition and extinction. In the fear acquisition phase, HSA show more attentional avoidance to CS+ than to CS−, whereas LSA avoid both CS+ and CS− in the same level. In the extinction phase, we can detect a higher increase of fixations to CS− than to CS+.

Checking differences in the pre-acquisition phase, an ANOVA showed a significant effect of Stimulus × Agent × Gender, $F(1,46) = 7.33$, $p = 0.009$, $\eta_p^2 = 0.14$, but follow-up ANOVAs conducting for women and men separately detected no significant main or interaction effects for both gender.

Concerning fear acquisition, an ANOVA approved significant effects of Time × Gender, $F(1,46) = 4.76$, $p = 0.034$, $\eta_p^2 = 0.09$,

FIGURE 6 | Mean count of fixations within the social area of interest (face and body) as well as environment for low (LSA) and high socially anxious (HSA) participants ($n = 53$). CS+ = agent paired with aversive unconditioned stimulus (US); CS− = agent without aversive US; NS = agent without aversive US and ap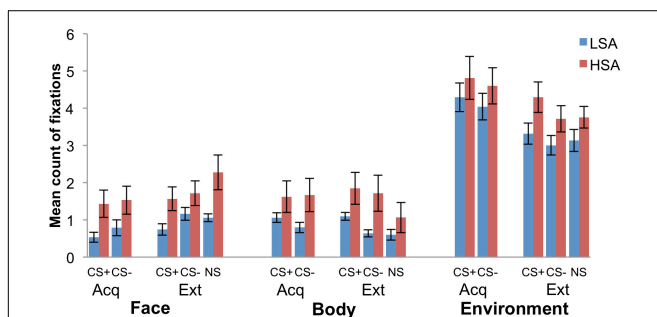pearing only in the extinction phase; Acq = acquisition phase; Ext = extinction phase. Standard errors are presented by error bars.

FIGURE 7 | Mean percentage of the count of fixations within the social area of interest (agent) for low (LSA) and high socially anxious (HSA) participants ($n = 53$). CS+ = agent paired with aversive unconditioned stimulus (US); CS− = agent without aversive US. Standard errors are presented by error bars.

Time × Stimulus × Social Anxiety, $F(1,46) = 8.51$, $p = 0.005$, $\eta_p^2 = 0.16$, and Stimulus × Agent × Social Anxiety, $F(1,46) = 6.20$, $p = 0.016$, $\eta_p^2 = 0.12$. Follow-up ANOVAs were conducted for the HSA and LSA group. For HSA, the interaction effects of Stimulus × Agent, $F(1,24) = 9.84$, $p = 0.004$, $\eta_p^2 = 0.29$, and Time × Stimulus, $F(1,24) = 12.5$, $p = 0.002$, $\eta_p^2 = 0.34$, reached significance. For the Stimulus × Agent interaction, follow-up t-tests show that HSA fixated significantly more frequently the female than the male CS+, $t(25) = −2.51$, $p = 0.019$, $d = 0.22$, whereas between the female and the male CS− was no significant difference ($p = 0.170$). Regarding the Time × Stimulus interaction, follow-up t-tests reveal that HSA looked significantly more frequently toward the CS+ than to the CS− at the beginning of the fear acquisition, $t(25) = 2.39$, $p = 0.025$, $d = 0.22$, whereas at the end of the acquisition they exhibited significantly more fixations counts toward the CS− than to the CS+, $t(25) = −2.98$, $p = 0.006$, $d = 0.22$, indicating an effect of fear conditioning according the attentional avoidance in HSA. For LSA, we detected no significant main or interaction effect.

For fear extinction, an ANOVA showed significant effects of Stimulus x Agent, $F(1,48) = 16.5$, $p < 0.001$, $\eta_p^2 = 0.26$, Time × Agent × Social Anxiety, $F(1,48) = 4.63$, $p = 0.037$, $\eta_p^2 = 0.09$, and Stimulus × Agent × Gender × Social Anxiety, $F(1,48) = 4.96$, $p = 0.031$, $\eta_p^2 = 0.09$. To disentangle this threefold interaction, we conducted follow-up ANOVAs for HSA and LSA separately. For HSA, significant effects of Stimulus × Agent, $F(1,24) = 21.5$, $p < 0.001$, $\eta_p^2 = 0.47$, and Stimulus × Agent × Gender, $F(1,24) = 9.17$, $p = 0.006$, $\eta_p^2 = 0.28$, were identified. Another follow-up ANOVAs were performed for HSA women and men separately. For HSA women, no significant main or interaction effect was detected. For HSA men, we found a significant interaction effect of Stimulus × Agent, $F(1,11) = 17.1$, $p = 0.002$, $\eta_p^2 = 0.61$, and the follow-up t-tests show that HSA men fixated significantly more female than male CS+ agents, $t(11) = −3.17$, $p = 0.009$, $d = 0.9$, whereas no significant difference between the female and male CS− was identified ($p = 0.059$). With regard to LSA, we found significant effects of Gender, $F(1,24) = 8.21$, $p = 0.009$, $\eta_p^2 = 0.26$, Time × Agent, $F(1,24) = 6.76$, $p = 0.016$, $\eta_p^2 = 0.22$, and Time × Stimulus × Agent × Gender, $F(1,24) = 4.77$, $p = 0.039$, $\eta_p^2 = 0.17$. Therefore, another follow-up ANOVAs were conducted for LSA women and men separately. No significant main or interaction effect was found for LSA women. For LSA men, a significant interaction effect of Time × Agent $F(1,12) = 6.94$, $p = 0.022$, $\eta_p^2 = 0.37$, could be detected and follow-up t-tests reveal that the fixations toward female agents significantly increased, $t(12) = −2.24$, $p = 0.045$, $d = 0.94$, whereas the fixations toward the male agents did not change during fear extinction ($p = 0.814$). However, no generalization effect was found.

DISCUSSION

The present study is one of the first to investigate social anxiety related physical behavior as well as hypervigilance and attentional avoidance through the use of implicit eye-tracking in HSA and

LSA participants (1) within the context of SFC, (2) with the use of anthropomorphic stimuli (3) in a social interaction with enhanced ecological validity (4) in VR. Therefore, we investigated the effect of induced and extinguished social fear on physical behavior as well as hypervigilance and attentional avoidance toward virtual female and male agents in HSA and LSA students in the SFC paradigm from Reichenberger et al. (2019), where participants actively approached different agents using a joystick. The outcome variables include fear ratings, physical behavior (duration of approach), hypervigilance (time of the first fixation), and attentional avoidance (count of fixations).

Concerning fear ratings, our results showed that social anxiety was successfully acquired and extinguished, except for a clear SFC in HSA men. It would be interesting to investigate whether this results comes from differences in coping strategies in HSA men, from somehow reduced conditionability, from reduced attention or higher variance between HSA men. The findings on the physical behavior might give a preliminary hint that HSA seemed to need more time to approach all agents during fear acquisition than LSA participants, which could mean that HSA were more carefully in approaching agents in this social interaction indicating enhanced social anxiety. With regard to hypervigilance, the HSA fixated the CS+ agents earlier than the CS− agents at the beginning of the fear acquisition, whereas at the end of the fear acquisition they fixated the CS+ agents later than the CS− agents. On the contrary, in LSA participants initial attention did not distinguish between the CS+ and CS− agents during fear acquisition. Moreover, according to our first hypothesis, we found an increased attentional avoidance to CS+ compared to CS− agents for HSA participants during fear acquisition. In contrast, in LSA participants sustained attention did not differ between the CS+ and CS− agents during acquisition, although they reported higher fear ratings for the CS+ than the CS− agents. Therefore we could conclude that LSA might not distinguish between an aversive or not-aversive human according to measures of attentional avoidance. Regarding fear extinction, the variation in attentional avoidance toward the aversive and non-aversive agents after fear acquisition disappeared for the HSA and LSA group after the extinction.

Relating to the vigilance-avoidance hypothesis, our results indicate that HSA directed especially their initial attention at CS+ than CS− agents at the first half of the fear acquisition, and avoided subsequently the CS+ more than the CS− agents at the second half of the fear acquisition to possibly reduce emotional distress. Our findings are compliant with the assumption that HSA or persons with SAD guide their initial attention to emotionally threatening information and tend to avoid eye contact or threatening stimuli to might reduce anxiety directly (Chen et al., 2002; Wieser et al., 2009b; Singh et al., 2015; Shechner et al., 2017). For clear evidence of a hypervigilance bias, we would have expected that HSA participants should guide their initial attention more toward threatening than non-threatening agents at the second half of the fear acquisition as well. The results of our current study are in line with Mühlberger et al. (2008) as well, who found that HSA participants avoided emotional facial expressions in a virtual fear-relevant situation. However, the authors could not affirm clear results of hypervigilance

to threat-relevant stimuli, like angry faces, in VR. Relating to these results e Claudino et al. (2019) call into question in their review the ecological validity of other studies, which indicated hypervigilance regarding emotions. Most research mainly used visual search and dot-probe tasks or measured gaze behavior (e.g., hypervigilance in time periods of 500–1500 ms) when persons looked at images of different faces on a computer display. In contrast, we utilized a SFC paradigm in VR, in which the participant had to approach different agents during fear learning. Furthermore, our non-clinical sample consisted of low and HSA students without a diagnosed SAD. Most of the studies found clear hypervigilant patterns of attention in persons with SAD.

Social anxiety is related with selective attention to social threatening stimuli and self-focused attention to internal cues, which are assumed as maintaining factors of SAD. For example, self-focused attention on internal cues (e.g., negative thoughts, emotions, and body sensations), could impair the performance and even inhibit perceiving positive social feedback which invalidates the false impressions of how others judge them (Perowne and Mansell, 2002; Spurr and Stopa, 2002).

Perowne and Mansell (2002) examined in LSA and HSA individuals their self-focused and selective external attention of non-verbal behaviors in a social-evaluative stress situation. The authors revealed that the HSA related more self-focused attention and perceived a more negative view of their performance than the LSA group. However, the HSA showed no less selective attention than the LSA group. These findings are in line that HSA persons monitor others and as far as they find a hint for a negative evaluation they turn their attention to internal cues (Clark and Wells, 1995; Perowne and Mansell, 2002).

Thus, the question arises how could self-focused attention influence hypervigilance and attentional avoidance in the current study? With regard to the cognitive model of Clark and Wells (1995), social anxiety leads to limited attentional processing of external cues. Moreover, the remaining reduced attentional processing of the external social situation tends to be biased, as ambiguous behaviors are more likely to be interpreted as negative (Clark and Ehlers, 2002). Therefore, HSA participants should focus more to internal than to external cues. This assumption is partly in line with our findings on the hypervigilance that at the first half of the fear acquisition the HSA fixated more early the CS+ than the CS− agents, whereas at the second half of the fear acquisition they fixated the CS+ later than the CS− agents. With regard to the attentional avoidance, we found for the HSA a reduced amount of fixations toward the CS+ than to the CS− agents at the end of the fear acquisition. Our findings partly support the vigilance-avoidance hypothesis that HSA directed especially their initial attention at threat-relevant agents with their negative hints, and at the end of the fear conditioning they avoided subsequently the threat-relevant more than the non-threatening agents to possibly reduce their emotional distress. For a clearer insight of the impact of self-focused attention to hypervigilance and attentional avoidance, the self- and other-focused attention could be measured with the Focus of Attention Questionnaire (Woody, 1996) in future studies.

There are less empirical findings relating the effects of fear conditioning on gaze behavior. Michalska et al. (2017) showed

that children hold their gaze more often and longer toward the area of the eyes of a CS+ than a CS−. The authors used two neutral faces as conditioned stimuli (CS+ and CS−). However, we know no empirical study measuring eye-gaze during fear conditioning with an enhanced ecological validity in VR.

Regarding ecological validity, Foulsham et al. (2011) examined whether gaze behavior in walking through the real world differs compared to watching videos of the same real world in a laboratory setting. The authors showed that the gaze behavior toward the path nearby and objects in the distance is remarkably different in the real world compared to viewing the videos. Moreover, participants looked more often to persons that were close to them on the video than in the real world. These findings demonstrate that the generalization of laboratory results to attentional processes in natural situations is restricted. Because, our participants were immersed in a virtual environment, where they could actively move and interact with female and male agents, we can assume different responses than toward photographs or videos of humans (McCall et al., 2016). We assume that behavioral and attentional processes measures in VR is more related to real-life situations than traditional laboratory paradigms. Moreover, an potential key element assessing reliable and valid physical or gaze behavior in a virtual environment is social presence, the feeling of the participant to interact with another sentient being which involve cognitive and emotional processes (Blascovich et al., 2002; Oh et al., 2018; Felnhofer et al., 2019). Thus, future research on attention and emotion regulation should include a measure of social presence in VR.

Furthermore, our study measures gaze behavior during approaching different social agents. Therefore, the question arises how physical behavior is related to gaze behavior. McCall et al. (2016) showed in their Affect Gallery paradigm that participants gazed more and came closer to positive than negative images. In contrast, the authors presented in their Crowded Room paradigm that participants gazed more directly at agents with angry or sad than neutral facial expressions, but revealed a larger interpersonal distance toward angry than neutral or sad agents. These results illustrate that gaze behavior does not have to match continuously with physical behavior. For example, in a social situation attention is more directed to a threat-relevant facial expression causing to physically avoid the potential threatening person.

Weeks et al. (2019) recommend the use of eye-tracking techniques in more future studies as these suitable tools could improve social skills training (e.g., providing an objective index of gaze avoidance) or in maximizing the effectiveness of therapeutic exposures (e.g., measuring and/or preventing the potential role as safety behavior). Furthermore, in order to investigate the mechanisms of physical and gaze behavior as an important feature in emotional learning and maintaining processes of social anxiety, eye-tracking indices are an interesting implicit method within fear conditioning.

Some study limitations should be noted. First, our non-clinical sample of mainly young students should be considered in generalizing the results to an additional context. Thus, to generalize our current findings we need a more diverse sample with different demographic and even clinical characteristics.

Second, technical failure in eye-tracking led to high data loss for seven participants, which may have impaired the likelihood of determining higher order interactions. Third, in future studies we should also define AOIs for the eyes and mouth region of the agents to investigate possible effects of emotional facial expression within a virtual social interaction.

Moreover, we want to note that the applied US (sound of spitting attended by an aversive air blast) during fear acquisition might not only signal potential social, but also physical harm and might induce general fear or even disgust. However, Grillon et al. (2004) revealed that the anticipation of an air blast was not aversive enough to induce a fear-potentiated startle reflex by unpredictable aversive events in humans. We think that being spat followed by the verbal rejection is rather a social relevant than a physical relevant US within a social situation. Thus, we assume that the triggered fear in our experiment is more a social fear of negative evaluation than a physical impairment. Furthermore, we did not receive any feedback regarding disgust, but unfortunately we used no standardized ratings or questionnaire of disgust, which should be measured in future studies. Moreover, it might be interesting to develop a measurement instrument to disentangle physical and social fear associated with the spit stimulus.

Furthermore, it is important to note that our power analysis based on the data from our prior study by Reichenberger et al. (2017) that contained two between subject conditions defined by social anxiety (LSA vs. HSA). In our current study, we included one more between subject condition defined by participants' gender (women vs. men). In detail, although our LSA and HSA group consisted of an equally distributed ratio of gender as mentioned above, the size of each condition is thus smaller. However, our main hypothesis does not relate to complex interaction effects of social anxiety and gender.

In conclusion, the current study demonstrated the advantages of VR in investigating physical and gaze behavior in social interactions in a highly standardized, experimentally controlled and ecological way. Our results show that HSA seemed to be more careful in approaching agents, fixated the CS+ earlier than the CS− agents at the first half of the fear acquisition, and showed increased attentional avoidance to CS+ compared to CS− agents during SFC compared to LSA participants. Further research could contribute to establishing an implicit and objective measure of physical and gaze behavior, which could potentially serve as a biomarker for SAD.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the University of Regensburg.

The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JR: study conception, data acquisition and analysis, and writing the manuscript. MP: data analysis and contribution to the manuscript. AM: study conception, data analysis, and contribution to the manuscript. All authors have approved the final version of the manuscript and its submission.

REFERENCES

Allan, S., and Gilbert, P. (1997). Submissive behaviour and psychopathology. *Br. J. Clin. Psychol.* 36, 467–488. doi: 10.1111/j.2044-8260.1997.tb01255.x

Becker, E. S., Rinck, M., Margraf, J., and Roth, W. T. (2001). The emotional stroop effect in anxiety disorders: general emotionality or disorder specificity? *J. Anxiety Disord.* 15, 147–159. doi: 10.1016/S0887-6185(01)00055-X

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., and Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychol. Inq.* 13, 103–124. doi: 10.1207/S15327965PLI1302-01

Bögels, S. M., and Mansell, W. (2004). Attention processes in the maintenance and treatment of social phobia: hypervigilance, avoidance and self-focused attention. *Clin. Psychol. Rev.* 24, 827–856. doi: 10.1016/j.cpr.2004.06.005

Çek, D., Sánchez, A., and Timpano, K. R. (2016). Social anxiety–linked attention bias to threat is indirectly related to post-event processing via subjective emotional reactivity to social stress. *Behav. Ther.* 47, 377–387. doi: 10.1016/j.beth.2016.02.007

Chen, N. T., and Clarke, P. J. (2017). Gaze-based assessments of vigilance and avoidance in social anxiety: a review. *Curr. Psychiatry Rep.* 19:59. doi: 10.1007/s11920-017-0808-4

Chen, Y. P., Ehlers, A., Clark, D. M., and Mansell, W. (2002). Patients with generalized social phobia direct their attention away from faces. *Behav. Res. Ther.* 40, 677–687. doi: 10.1016/S0005-7967(01)00086-9

Clark, D. M., and Ehlers, A. (2002). "Soziale phobie: eine kognitive perspektive," in *Soziale Phobie und Soziale Angststörung*, eds U. Stangier, and T. Fydrich (Göttingen: Hogrefe), 157–180.

Clark, D. M., and Wells, A. (1995). "A cognitive model of social phobia," in *Social Phobia: Diagnosis, Assessment and Treatment*, eds R. Heimberg, L. Michael, D. A. Hope, and F. R. Schneier (New York, NY: Guilford Press), 69–93.

Connor, K. M., Davidson, J. R., Churchill, L. E., Sherwood, A., Weisler, R. H., and Foa, E. (2000). Psychometric properties of the Social Phobia Inventory (SPIN). New self-rating scale. *Br. J. Psychiatry* 176, 379–386. doi: 10.1192/bjp.176.4.379

Dechant, M., Trimpl, S., Wolff, C., Mühlberger, A., and Shiban, Y. (2017). Potential of virtual reality as a diagnostic tool for social anxiety: a pilot study. *Comput. Human Behav.* 76, 128–134. doi: 10.1016/j.chb.2017.07.005

e Claudino, R. G., de Lima, L. K. S., de Assis, E. D. B., and Torro, N. (2019). Facial expressions and eye tracking in individuals with social anxiety disorder: a systematic review. *Psicol. Reflex. Crit.* 32:9. doi: 10.1186/s41155-019-0121-8

Felnhofer, A., Hlavacs, H., Beutl, L., Kryspin-Exner, I., and Kothgassner, O. D. (2019). Physical presence, social presence, and anxiety in participants with social anxiety disorder during virtual cue exposure. *Cyberpsychol. Behav. Soc. Netw.* 22, 46–50. doi: 10.1089/cyber.2018.0221

Foulsham, T., Walker, E., and Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vis. Res.* 51, 1920–1931. doi: 10.1016/j.visres.2011.07.002

Garner, M., Mogg, K., and Bradley, B. P. (2006). Orienting and maintenance of gaze to facial expressions in social anxiety. *J. Abnorm. Psychol.* 115, 760–770. doi: 10.1037/0021-843X.115.4.760

Gilboa-Schechtman, E., Foa, E. B., and Amir, N. (1999). Attentional biases for facial expressions in social phobia: the face-in-the-crowd paradigm. *Cogn. Emot.* 13, 305–318. doi: 10.1080/026999399379294

Grillon, C., Baas, J. P., Lissek, S., Smith, K., and Milstein, J. (2004). Anxious responses to predictable and unpredictable aversive events. *Behav. Neurosci.* 118, 916–924. doi: 10.1037/0735-7044.118.5.916

Heimberg, R. G., Brozovich, F. A., and Rapee, R. M. (2014). "A cognitive-behavioral model of social anxiety disorder: update and extension," in *Social Anxiety*, eds S. G. Hofmann, and P. M. DiBartolo (San Diego: Academic Press), 705–728. doi: 10.1016/b978-0-12-394427-6.00024-8

Hopkins, L. S., Schultz, D. H., Hannula, D. E., and Helmstetter, F. J. (2015). Eye movements index implicit memory expression in fear conditioning. *PLoS One* 10:e0141949. doi: 10.1371/journal.pone.0141949

Krohne, H. W., and Egloff, B. (1999). *Das Angstbewältigungs-Inventar (ABI)*. Frankfurt am Main: Swets Test Services GmbH.

Lange, W.-G., Heuer, K., Langner, O., Keijsers, G. P., Becker, E. S., and Rinck, M. (2011). Face value: eye movements and the evaluation of facial crowds in social anxiety. *J. Behav. Ther. Exp. Psychiatry* 42, 355–363. doi: 10.1016/j.jbtep.2011.02.007

Lazarov, A., Abend, R., and Bar-Haim, Y. (2016). Social anxiety is related to increased dwell time on socially threatening faces. *J. Affect. Disord.* 193, 282–288. doi: 10.1016/j.jad.2016.01.007

Maidenberg, E., Chen, E., Craske, M., Bohn, P., and Bystritsky, A. (1996). Specificity of attentional bias in panic disorder and social phobia. *J. Anxiety Disord.* 10, 529–541. doi: 10.1016/S0887-6185(96)00028-X

Mansell, W., Ehlers, A., Clark, D., and Chen, Y.-P. (2002). Attention to positive and negative social-evaluative words: investigating the effects of social anxiety, trait anxiety and social threat. *Anxiety Stress & Coping* 15, 19–29. doi: 10.1080/10615800290007263

McCall, C., Hildebrandt, L. K., Hartmann, R., Baczkowski, B. M., and Singer, T. (2016). Introducing the Wunderkammer as a tool for emotion research: unconstrained gaze and movement patterns in three emotionally evocative virtual worlds. *Comput. Human Behav.* 59, 93–107. doi: 10.1016/j.chb.2016.01.028

Michalska, K. J., Machlin, L., Moroney, E., Lowet, D. S., Hettema, J. M., Roberson-Nay, R., et al. (2017). Anxiety symptoms and children's eye gaze during fear learning. *J. Child Psychol. Psychiatry* 58, 1276–1286. doi: 10.1111/jcpp.12749

Mogg, K., and Bradley, B. P. (2002). Selective orienting of attention to masked threat faces in social anxiety. *Behav. Res. Ther.* 40, 1403–1414. doi: 10.1016/S0005-7967(02)00017

Mogg, K., Philippot, P., and Bradley, B. P. (2004). Selective attention to angry faces in clinical social phobia. *J. Abnorm. Psychol.* 113, 160–165. doi: 10.1037/0021-843X.113.1.160

Mühlberger, A., Wieser, M. J., and Pauli, P. (2008). Visual attention during virtual social situations depends on social anxiety. *CyberPsychol. Behav.* 11, 425–430. doi: 10.1089/cpb.2007.0084

Oh, C. S., Bailenson, J. N., and Welch, G. F. (2018). A systematic review of social presence: definition, antecedents, and implications. *Front. Robot. AI* 5:114. doi: 10.3389/frobt.2018.00114

Perowne, S., and Mansell, W. (2002). Social anxiety, self-focused attention, and the discrimination of negative, neutral and positive audience members by their non-verbal behaviours. *Behav. Cogn. Psychotherap.* 30, 11–23. doi: 10.1017/S1352465802001030

Rapee, R. M., and Heimberg, R. G. (1997). A cognitive-behavioral model of anxiety in social phobia. *Behav. Res. Ther.* 35, 741–756. doi: 10.1016/S0005-7967(97)00022-3

Reichenberger, J., Pfaller, M., Forster, D., Gerczuk, J., Shiban, Y., and Mühlberger, A. (2019). Men scare me more: gender differences in social fear conditioning in virtual reality. *Front. Psychol.* 10:1617. doi: 10.3389/fpsyg.2019.01617

Reichenberger, J., Porsch, S., Wittmann, J., Zimmermann, V., and Shiban, Y. (2017). Social fear conditioning paradigm in virtual reality: social vs. electrical aversive conditioning. *Front. Psychol.* 8:1979. doi: 10.3389/fpsyg.2017.01979

Rinck, M., Becker, E. S., Kellermann, J., and Roth, W. T. (2003). Selective attention in anxiety: distraction and enhancement in visual search. *Depress. Anxiety* 18, 18–28. doi: 10.1002/da.10105

Schofield, C. A., Johnson, A. L., Inhoff, A. W., and Coles, M. E. (2012). Social anxiety and difficulty disengaging threat: evidence from eye-tracking. *Cogn. Emot.* 26, 300–311. doi: 10.1080/02699931.2011.602050

Shechner, T., Jarcho, J. M., Wong, S., Leibenluft, E., Pine, D. S., and Nelson, E. E. (2017). Threats, rewards, and attention deployment in anxious youth and adults: an eye tracking study. *Biol. Psychol.* 122, 121–129. doi: 10.1016/j.biopsycho.2015.10.004

Shiban, Y., Reichenberger, J., Neumann, I. D., and Mühlberger, A. (2015). Social conditioning and extinction paradigm: a translational study in virtual reality. *Front. Psychol.* 6:400. doi: 10.3389/fpsyg.2015.00400

Singh, J. S., Capozzoli, M. C., Dodd, M. D., and Hope, D. A. (2015). The effects of social anxiety and state anxiety on visual attention: testing the vigilance–avoidance hypothesis. *Cogn. Behav. Ther.* 44, 377–388. doi: 10.1080/16506073.2015.1016447

Spurr, J. M., and Stopa, L. (2002). Self-focused attention in social phobia and social anxiety. *Clin. Psychol. Rev.* 22, 947–975. doi: 10.1016/S0272-7358(02)00107-1

Stangier, U., Heidenreich, T., Berardi, A., Golbs, U., and Hoyer, J. (1999). Die erfassung sozialer phobie durch social interaction anxiety scale (SIAS) und die social phobia scale (SPS). *Z. Klin. Psychol. Psychother.* 28, 28–36. doi: 10.1026/0084-5345.28.1.28

Stangier, U., and Steffens, M. (2002). *Social Phobia Inventory (SPIN)–Deutsche Fassung.* Frankfurt am Main: Psychologisches Institut der Universität Frankfurt am Main.

Stumpf, H., Angleitner, A., Wieck, T., Jackson, D., and Beloch-Till, H. (1985). *Deutsche Personality Research Form (PRF).* Boston, MA: Hogrefe.

Weeks, J. W., Howell, A. N., Srivastav, A., and Goldin, P. R. (2019). "Fear guides the eyes of the beholder": assessing gaze avoidance in social anxiety disorder via covert eye tracking of dynamic social stimuli. *J. Anxiety Disord.* 65, 56–63. doi: 10.1016/j.janxdis.2019.05.005

Wieser, M. J., Pauli, P., Alpers, G. W., and Mühlberger, A. (2009a). Is eye to eye contact really threatening and avoided in social anxiety?—an eye-tracking and psychophysiology study. *J. Anxiety Disord.* 23, 93–103. doi: 10.1016/j.janxdis.2008.04.004

Wieser, M. J., Pauli, P., Grosseibl, M., Molzow, I., and Mühlberger, A. (2010). Virtual social interactions in social anxiety—the impact of sex, gaze, and interpersonal distance. *Cyberpsychol. Behav. Soc. Netw.* 13, 547–554. doi: 10.1089/cyber.2009.0432

Wieser, M. J., Pauli, P., Weyers, P., Alpers, G. W., and Mühlberger, A. (2009b). Fear of negative evaluation and the hypervigilance-avoidance hypothesis: an eye-tracking study. *J. Neural. Transm.* 116, 717–723. doi: 10.1007/s00702-008-0101-100

Woody, S. R. (1996). Effects of focus of attention on anxiety levels and social performance of individuals with social phobia. *J. Abnorm. Psychol.* 105, 61–69. doi: 10.1037/0021-843x.105.1.61

The Current Research of Spatial Cognitive Evaluation and Training With Brain–Computer Interface and Virtual Reality

Yanhong Zhou[1,2†], Dong Wen[1,3†], Huibin Lu[1,4], Wang Yao[1,3], Yijun Liu[5], Wenbo Qian[1,4] and Jingpeng Yuan[1,3]*

[1] School of Information Science and Engineering, Yanshan University, Qinhuangdao, China, [2] School of Mathematics and Information Science and Technology, Hebei Normal University of Science and Technology, Qinhuangdao, China, [3] The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Yanshan University, Qinhuangdao, China, [4] The Key Laboratory of Information Transmission and Signal Processing of Hebei Province, Yanshan University, Qinhuangdao, China, [5] School of Science, Yanshan University, Qinhuangdao, China

INTRODUCTION

Spatial cognitive evaluation and training (SCET) is a rapidly growing research field (Chunyin et al., 2011) in cognitive study. SCET is also greatly significant in the diagnosis and rehabilitation of mild cognitive impairment (MCI), mainly because patients with MCI show symptoms of spatial cognitive impairment at an early stage (Allison et al., 2016; Laczó et al., 2016). For SCET, real-time and accurate quantification is the ultimate goal in evaluation (Lin et al., 2015); it is expected to have a strong sense of participation for the subjects in training, and the training content is closely related to their daily life (Bormans et al., 2016).

Virtual reality (VR) (Tu et al., 2017) and brain–computer interface (BCI) (Xu et al., 2013) are popular technologies in SCET. Training with VR meets experience and social needs of subjects and can be used as the main way of spatial cognitive training (SCT) (Serino et al., 2015; Bormans et al., 2016; Davis and Ohman, 2016; Migo et al., 2016; Tu et al., 2017; Zygouris et al., 2017). However, in these studies, it is difficult to understand the training effect in real time to the subjects and trainers, although they are very eager to observe the training effect timely so as to adjust the training state or program. BCI based on electroencephalogram (EEG) signals (Xu et al., 2018) is often used for real-time SCET and can be applied to real-time monitoring of brain activity on the premise of high temporal resolution (Lin et al., 2015; Han et al., 2017; Chen et al., 2018; Guevara et al., 2018; Pergher et al., 2018). Therefore, it is a good choice to combine BCI and VR (Lechner et al., 2014; Koo et al., 2015; De Tommaso et al., 2016; Donati et al., 2016; Vourvopoulos and I Badia, 2016) for SCET, and there are preliminarily applied studies (Bischof and Boulanger, 2003; Jaiswal et al., 2010; Kober and Neuper, 2011; Tarnanas et al., 2015), which suggested that BCI-VR was a recommendable method of SCET. However, this combination is still in its infancy, and more work needs to be done before conclusions can be confidently drawn.

This study will review the literature related to SCET with VR, BCI, and BCI-VR; discuss the potential advantages of BCI-VR in SCET and future problems to be solved; and put forward our opinions. It is expected that this analysis may provide valuable suggestions for the field of information technology in SCET.

This study used the Web of Science–Science Citation Index/Social Sciences Citation Index (WOS–SCI/SSCI) database, focused on the studies of BCI, VR, and BCI-VR in SCET. The following search keywords were used: "spatial cognitive evaluation (SCE)" or "spatial cognitive training (SCT)" in combination with "brain–computer interface (BCI)" or "virtual reality (VR)." The most recent search was conducted on March 21, 2019.

RESEARCH STATUS OF SCET WITH VR

It is a relatively independent cognitive factor for spatial cognitive ability (SCA), which can be improved by reasonable training (Chunyin et al., 2011). The water maze with VR, which has become the most classical experimental method in the field of SCET, was used to measure the SCA of rats (Morris, 1984). This method can be used to train the SCA of rats (Chunyin et al., 2011) and distinguish male and female participants (Astur et al., 1998). Currently, VR opens up a new way for the research of SCET and had been applied to some researches of SCET (Bormans et al., 2016; Tu et al., 2017), in which VR could help subjects to experience a strong sense of participation in training process, and the training environment with VR was not limited by the area.

In terms of SCE, VR is widely used in the evaluation of AD and MCI (Weniger et al., 2011). For AD patients, virtual hospital (Jiang and Li, 2007), virtual auditorium (Lange et al., 2007), virtual city (Zakzanis et al., 2009), virtual building (Cushman et al., 2008), and virtual environment of apartment (Davis and Ohman, 2016) can be used to evaluate the path finding and navigation ability. When implementing a virtual path learning task, preclinical AD showed disorder (Allison et al., 2016), and there are correlations between regional neurodegeneration of AD and ability of spatial terrain memory (Pengas et al., 2012). In addition, relocation tasks with VR can be used to assess the spatial memory of early AD (Shamsuddin et al., 2011, 2012; Caffo et al., 2012). For MCI patients, the missing ability to orientation in virtual space (Morganti et al., 2013) and virtual supermarket (Tu et al., 2015; Zygouris et al., 2017) can help us to screen them from all subjects (Zygouris et al., 2017). In addition, the virtual path navigation can be used to evaluate their visual spatial memory (Lesk et al., 2014), and virtual room location search can be used to detect their obstacles in spatial navigation (Serino et al., 2015).

In terms of SCT, there existed several VR games widely used by cognitive impairment patients. Virtual memory palaces game can be used to improve the quality of life and memory of AD (Bormans et al., 2016), virtual buildings navigation game can be applied to improve the driving skills and daily cognitive ability of AD patients (White and Moussavi, 2016), and the large virtual outdoor parks game can be used to train the navigation ability of patients with dementia (Flynn et al., 2003). In addition, other VR games can also be used to improve the spatial attention of cognitive impairment patients (Manera et al., 2016).

RESEARCH STATE OF SCET WITH THE CHANGE OF BCI SIGNALS

Previous studies have shown that spatial memory, spatial orientation, and navigation dominated by the hippocampus were important components of spatial cognition (Olton et al., 1979). They also have shown that the dynamic characteristics of EEG signals from the parietal lobe were directly related to the ability of spatial memory and navigation (Chiu et al., 2012). Therefore, it is feasible to explore the changes of SCA from the perspective of EEG signal analysis.

From the perspective of SCE, BCI signals are widely used in the evaluation of spatial navigation and memory. For spatial navigation, dynamic features of the EEG signal can respond to the change of spatial navigation (Lin et al., 2009, 2015). The navigation performance was related to the power modulation in theta or gamma frequency band (Bell, 2002; White et al., 2012; Chen et al., 2018), Especially, the phase reset of theta rhythm during spatial navigation caused the NT170 latency effect (Baker and Holroyd, 2013). Interestingly, female participants showed stronger theta oscillations during spatial navigation (Nishiyama et al., 2002). For spatial memory, the perception and action in the memory task are different by EEG network analysis (Protopapa et al., 2011). The power of EEG signals from AD patients on alpha and theta frequency bands was significantly higher relative to normal control when performing object-location memory tasks (Han et al., 2017). In addition, when the delay, peak amplitude, and root mean square of P300 act as the features in visual spatial memory paradigm, the classification accuracy could achieve up to 91.76% (Li et al., 2013).

From the perspective of SCT, almost all of the studies focused on the training of spatial memory, which can enhance the connection between nervous activity and the brain network in frontal, parietal, and occipital lobes of MCI patients (Hampstead et al., 2011). In addition, the Pearson correlation coefficient between the right prefrontal lobe regions decreased after spatial memory task, which indicated that the region had a higher participation in memory task (Plank et al., 2010; Guevara et al., 2018), Surprisingly, old people showed higher amplitude of P300 in the parietal lobe than the young on spatial memory training (Pergher et al., 2018),

Based on the above researches, BCI can provide support for objective data analysis in SCET and can effectively improve the scientific quantification performance of this research field, as well as the operability and repeatability of experiments.

RESEARCH STATUS OF SCET WITH BCI-VR

Research Value of BCI-VR

In recent years, the combination of BCI and VR has provided practical benefits for training and evaluation (Achanccaray et al., 2017). In VR training, BCI can be used to analyze brain activity synchronously in real time (Martišius and Damaševičius, 2016). Meanwhile, VR can provide a situation similar to the daily life, in which the brain activity could be evaluated by EEG analysis (Alchalcabi et al., 2017). Therefore, the fusion of VR and BCI can solve the subjective and nonreal-time evaluation of training effect in the training process. Recently, BCI-VR has been successfully used in some application fields, including steady-state visual evoked potential (Lechner et al., 2014) and motor imagery (I Badia et al., 2012; Vourvopoulos and I Badia, 2016) for stroke rehabilitation, which can overcome the limitation of traditional monitor refresh frequency (Calore et al., 2014; Koo et al., 2015) and increase the engagement of subjects (Koo and Choi, 2015). In addition, it has other applications, such as motion imaging for paraplegic rehabilitation (Donati et al., 2016), P300 for cognitive

FIGURE 1 | The schematic diagram for the system description of the SCET with BCI-VR.
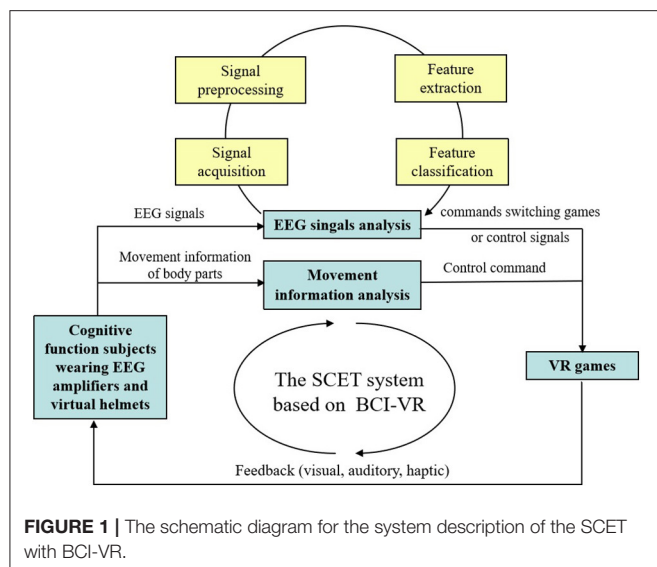
training of normal control (De Tommaso et al., 2016), autism training (Amaral et al., 2017), and hyperactivity training (Rohani and Puthusserypady, 2015).

Research Analysis of SCET With BCI-VR

At present, there are several studies that combine BCI and VR to exert their respective advantages in the field of SCET. For the spatial navigation, Bischof and Boulanger (2003) found that theta oscillation of EEG signals was related to the encoding and retrieval of spatial information when performing VR maze navigation tasks, and Silvia et al. compared the differences of EEG signals from male and female adults on theta frequency band with VR spatial navigation tasks (Kober and Neuper, 2011). For the spatial memory, Jaiswal et al. (2010) combined BCI with VR to observe differences in EEG activity during the coding and retrieval stages of spatial memory tasks. For other spatial task, Tarnanas et al. (2014a,b); Tarnanas et al. (2015) used VR day-out task to evaluate its predictive value of MCI. All of these studies have verified the rationality of the combination of BCI and VR in SCET research. **Figure 1** showed the schematic diagram for the system description of the SCET with BCI-VR.

DISCUSSION

Many studies suggested that VR can be used to the SCET (Bormans et al., 2016; Manera et al., 2016; Tu et al., 2017; Zygouris et al., 2017) and changed the old way of SCE with neuropsychological scales, subjective judgments, qualitative description, and so on (Manera et al., 2016). VR technology has obvious advantages in immersive space experience and internal space visualization in the field of architectural design. However, VR is still unable to evaluate the effects of SCT in real time like BCI. Many studies also suggested that the change of BCI signals can be used to the SCET (Han et al., 2017; Chen et al., 2018; Guevara et al., 2018; Pergher et al., 2018). BCI can be used to analyze EEG signals in SCET effectively and improve

the operability and repeatability of evaluation and training. However, BCI cannot supply an abundant environment for SCET compared to VR. Several studies in the field of SCET combined BCI and VR to exert their respective advantages (Jaiswal et al., 2010; Kober and Neuper, 2011; Tarnanas et al., 2015). Compared with BCI and VR used in SCET, BCI-VR not only can improve the effect of spatial cognition training but also can overcome the limitation of spatial cognition evaluation.

However, various challenges to be broken through currently exist in BCI-VR used in SCET. Firstly, effective interaction remains to be achieved in VR, so a new interaction model, which combines natural interaction in VR with BCI, needs to be studied. Secondly, wearing the wet electrode in BCI increases the experimental preparation time compared with the dry electrode in BCI. Currently, there is a lack of the dry electrode with high signal quality and good comfort. Therefore, the design and development of electrode in the future need to be strengthened. Thirdly, the SCT in the real environment requires a large space, but the range of activity scenes covered by existing VR equipment cannot meet this demand. Hence, the algorithm that can scale up or down the VR game scenes at will is very important.

CONCLUSIONS

In conclusion, this study reviewed the recent literature of VR, BCI, and BCI-VR used in SCET. Although these studies obtained promising results, more work needs to be done before conclusions can be confidently drawn. It is suggested that BCI and VR should be deeply integrated in order to give full play to their respective technical advantages. The current technical bottlenecks of BCI and VR should be broken through to provide personalized, comfortable, and real-time technical support for creating an effective environment for SCET. Therefore, in the future, for the SCET, it will be necessary to not only solve the technical problems of BCI-VR but also to take patients as important subjects, design personalized and adaptive BCI-VR games, and train the patients. In addition, the training effect had better be evaluated and fed back in real time in order to meet the individual needs of patients.

AUTHOR CONTRIBUTIONS

HL, YZ, WY, and YL designed the work. YZ, DW, and HL wrote this manuscript. YZ, DW, WY, YL, WQ, and JY collected and analyzed the literatures. DW, WQ, and JY revised the manuscript before firstly submitted. HL, WY, and YL revised important intellectual content, added one figure, approved the final version to be published, and revised English grammar and sentences.

FUNDING

REFERENCES

Achanccaray, D., Acuna, K., Carranza, E., and Andreu-Perez, J. (2017). "A virtual reality and brain computer interface system for upper limb rehabilitation of post stroke patients," in *2017 IEEE International Conference on Fuzzy Systems* (Naples), 1–5. doi: 10.1109/FUZZ-IEEE.2017.8015726

Alchalcabi, A. E., Eddin, A. N., and Shirmohammadi, S. (2017). "More attention, less deficit: wearable EEG-based serious game for focus improvement," in *2017 IEEE 5th International Conference on Serious Games and Applications for Health (SeGAH)* (Perth: IEEE), 1–8. doi: 10.1109/SeGAH.2017.7939288

Allison, S. L., Fagan, A. M., Morris, J. C., and Head, D. (2016). Spatial navigation in preclinical Alzheimer's disease. *J. Alzheimer's Dis.* 52, 77–90. doi: 10.3233/JAD-150855

Amaral, C. P., Simões, M. A., Mouga, S., Andrade, J., and Castelo-Branco, M. (2017). A novel Brain Computer Interface for classification of social joint attention in autism and comparison of 3 experimental setups: a feasibility study. *J. Neurosci. Methods* 290, 105–115. doi: 10.1016/j.jneumeth.2017.07.029

Astur, R. S., Ortiz, M. L., and Sutherland, R. J. (1998). A characterization of performance by men and women in a virtual Morris water task: a large and reliable sex difference. *Behav. Brain Res.* 93, 185–190. doi: 10.1016/S0166-4328(98)00019-9

Baker, T. E., and Holroyd, C. B. (2013). The topographical N170: electrophysiological evidence of a neural mechanism for human spatial navigation. *Biol. Psychol.* 94, 90–105. doi: 10.1016/j.biopsycho.2013.05.004

Bell, M. A. (2002). Power changes in infant EEG frequency bands during a spatial working memory task. *Psychophysiology* 39, 450–458. doi: 10.1111/1469-8986.3940450

Bischof, W. F., and Boulanger, P. (2003). Spatial navigation in virtual reality environments: an EEG analysis. *CyberPsychol. Behav.* 6, 487–495. doi: 10.1089/109493103769710514

Bormans, K., Roe, K., and De Wachter, D. (2016). Virtual memory palaces to improve quality of life in Alzheimer's disease. *Annu. Rev. CyberTher. Telemed.* 14, 227–232.

Caffo, A. O., De Caro, M. F., Picucci, L., Notarnicola, A., Settanni, A., Livrea, P., et al. (2012). Reorientation deficits are associated with amnestic mild cognitive impairment. *Am. J. Alzheimer's Dis. Other Dement.* 27, 321–330. doi: 10.1177/1533317512452035

Calore, E., Gadia, D., and Marini, D. (2014). "Eliciting steady-state visual evoked potentials by means of stereoscopic displays," in *Stereoscopic Displays and Applications XXV*: *International Society for Optics and Photonics* (San Francisco, CA), 901126. doi: 10.1117/12.2036936

Chen, D., Kunz, L., Wang, W., Zhang, H., Wang, W.-X., Schulze-Bonhage, A., et al. (2018). Hexadirectional modulation of theta power in human entorhinal cortex during spatial navigation. *Curr. Biol.* 28, 3310–3315.e3314. doi: 10.1016/j.cub.2018.08.029

Chiu, T. C., Gramann, K., Ko, L. W., Duann, J. R., Jung, T. P., and Lin, C. T. (2012). Alpha modulation in parietal and retrosplenial cortex correlates with navigation performance. *Psychophysiology* 49, 43–55. doi: 10.1111/j.1469-8986.2011.01270.x

Chunyin, S., Zhihong, H., Kaifeng, L., Xiaomin, L., and Wenjing, L. (2011). The effect of training style to spatial learning and memory of rat. *J. Shanxi Med. Univ.* 42, 539–541.

Cushman, L. A., Stein, K., and Duffy, C. J. (2008). Detecting navigational deficits in cognitive aging and Alzheimer disease using virtual reality. *Neurology* 71, 888–895. doi: 10.1212/01.wnl.0000326262.67613.fe

Davis, R., and Ohman, J. (2016). Wayfinding in ageing and Alzheimer's disease within a virtual senior residence: study protocol. *J. Adv. Nurs.* 72, 1677–1688. doi: 10.1111/jan.12945

De Tommaso, M., Ricci, K., Delussi, M., Montemurno, A., Vecchio, E., Brunetti, A., et al. (2016). Testing a novel method for improving wayfinding by means of a P3b Virtual Reality Visual Paradigm in normal aging. *SpringerPlus* 5:1297. doi: 10.1186/s40064-016-2978-7

Donati, A. R., Shokur, S., Morya, E., Campos, D. S., Moioli, R. C., Gitti, C. M., et al. (2016). Long-term training with a brain-machine interface-based gait protocol induces partial neurological recovery in paraplegic patients. *Sci. Rep.* 6:30383. doi: 10.1038/srep30383

Flynn, D., Van Schaik, P., Blackman, T., Femcott, C., Hobbs, B., and Calderon, C. (2003). Developing a virtual reality–based methodology for

people with dementia: a feasibility study. *CyberPsychol. Behav.* 6, 591–611. doi: 10.1089/109493103322725379

Guevara, M. A., Paniagua, E. I. C., González, M. H., Carrillo, I. K. S., Sepúlveda, M. L. A., Orozco, J. C. H., et al. (2018). EEG activity during the spatial span task in young men: differences between short-term and working memory. *Brain Res.* 1683, 86–94. doi: 10.1016/j.brainres.2018.02.004

Hampstead, B. M., Stringer, A. Y., Stilla, R. F., Deshpande, G., Hu, X., Moore, A. B., et al. (2011). Activation and effective connectivity changes following explicit-memory training for face–name pairs in patients with mild cognitive impairment: a pilot study. *Neurorehabil. Neural Repair* 25, 210–222. doi: 10.1177/1545968310382424

Han, Y., Wang, K., Jia, J., and Wu, W. (2017). Changes of EEG spectra and functional connectivity during an object-location memory task in Alzheimer's disease. *Front. Behav. Neurosci.* 11:107. doi: 10.3389/fnbeh.2017.00107

I Badia, S. B., Morgade, A. G., Samaha, H., and Verschure, P. (2012). Using a hybrid brain computer interface and virtual reality system to monitor and promote cortical reorganization through motor activity and motor imagery training. *IEEE Trans. Neural Syst. Rehabil. Eng.* 21, 174–181. doi: 10.1109/TNSRE.2012.2229295

Jaiswal, N., Ray, W., and Slobounov, S. (2010). Encoding of visual–spatial information in working memory requires more cerebral efforts than retrieval: evidence from an EEG and virtual reality study. *Brain Res.* 1347, 80–89. doi: 10.1016/j.brainres.2010.05.086

Jiang, C.-F., and Li, Y.-S. (2007). "Virtual hospital-a computer-aided platform to evaluate the sense of direction," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Lyon: IEEE), 2361–2364. doi: 10.1109/IEMBS.2007.4352801

Kober, S. E., and Neuper, C. (2011). Sex differences in human EEG theta oscillations during spatial navigation in virtual reality. *Int. J. Psychophysiol.* 79, 347–355. doi: 10.1016/j.ijpsycho.2010.12.002

Koo, B., and Choi, S. (2015). "SSVEP response on oculus rift in brain-computer interface (BCI)," in *The 3rd International Winter Conference on Brain-Computer Interface* (Sabuk: IEEE), 1–4. doi: 10.1109/IWW-BCI.2015.7073028

Koo, B., Lee, H.-G., Nam, Y., and Choi, S. (2015). "Immersive BCI with SSVEP in VR head-mounted display," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine And Biology Society (EMBC)* (Milan: IEEE), 1103–1106. doi: 10.1109/EMBC.2015.7318558

Laczó, J., Parizkova, M., Moffat, S. D., Vyhnalek, M., Andel, R., and Hort, J. (2016). The effect of early-stage Alzheimer's disease on spatial navigation strategies: a pilot study. *Alzheimer's Dement.* 12, 565–566. doi: 10.1016/j.jalz.2016.06.1106

Lange, B. S., Rizzo, A., Astur, R., and Parsons, T. (2007). Virtual reality visuospatial and wayfinding assessment for mild dementia of the Alzheimer's type. *Alzheimer's Dement.* 3:S135. doi: 10.1016/j.jalz.2007.04.178

Lechner, A., Ortner, R., and Guger, C. (2014). "Feedback strategies for BCI based stroke rehabilitation: evaluation of different approaches," in *2nd International Conference on NeuroRehabilitation* (Aalborg: Springer), 507–512. doi: 10.1007/978-3-319-08072-7_75

Lesk, V. E., Shamsuddin, S. N. W., Walters, E. R., and Ugail, H. (2014). Using a virtual environment to assess cognition in the elderly. *Virtual Real.* 18, 271–279. doi: 10.1007/s10055-014-0252-2

Li, X., Yan, Y., and Wei, W. (2013). Identifying patients with poststroke mild cognitive impairment by pattern recognition of working memory load-related ERP. *Comput. Math. Methods Med.* 2013:658501. doi: 10.1155/2013/658501

Lin, C.-T., Chiu, T.-C., and Gramann, K. (2015). EEG correlates of spatial orientation in the human retrosplenial complex. *Neuroimage* 120, 123–132. doi: 10.1016/j.neuroimage.2015.07.009

Lin, C.-T., Yang, F.-S., Chiou, T.-C., Ko, L.-W., Duann, J.-R., and Gramann, K. (2009). "EEG-based spatial navigation estimation in a virtual reality driving environment," in *2009 Ninth IEEE International Conference on Bioinformatics and BioEngineering* (Taichung: IEEE), 435–438. doi: 10.1109/BIBE. 2009.65

Manera, V., Chapoulie, E., Bourgeois, J., Guerchouche, R., David, R., Ondrej, J., et al. (2016). A feasibility study with image-based rendered virtual reality in patients with mild cognitive impairment and dementia. *PLoS ONE* 11:e0151487. doi: 10.1371/journal.pone.0151487

Martišius, I., and Damaševičius, R. (2016). A prototype SSVEP based real time BCI gaming system. *Comput. Intell. Neurosci.* 2016:3861425. doi: 10.1155/2016/3861425

Migo, E., O'daly, O., Mitterschiffthaler, M., Antonova, E., Dawson, G., Dourish, C., et al. (2016). Investigating virtual reality navigation in amnestic mild cognitive impairment using fMRI. *Aging Neuropsychol. Cogn.* 23, 196–217. doi: 10.1080/13825585.2015.1073218

Morganti, F., Stefanini, S., and Riva, G. (2013). From allo-to egocentric spatial ability in early Alzheimer's disease: a study with virtual reality spatial tasks. *Cogn. Neurosci.* 4, 171–180. doi: 10.1080/17588928.2013.854762

Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *J. Neurosci. Methods* 11, 47–60. doi: 10.1016/0165-0270(84)90007-4

Nishiyama, N., Mizuhara, H., Miwakeichi, F., and Yamaguchi, Y. (2002). "Theta episodes observed in human scalp EEG during virtual navigation-spatial distribution and task dependence," in *Proceedings of the 9th International Conference on Neural Information Processing, ICONIP'02* (Singapore: IEEE), 428–432.

Olton, D. S., Becker, J. T., and Handelmann, G. E. (1979). Hippocampus, space, and memory. *Behav. Brain Sci.* 2, 313–322. doi: 10.1017/S0140525X00062713

Pengas, G., Williams, G., Acosta-Cabronero, J., Ash, T., Hong, Y., Izquierdo-Garcia, D., et al. (2012). The relationship of topographical memory performance to regional neurodegeneration in Alzheimer's disease. *Front. Aging Neurosci.* 4:17. doi: 10.3389/fnagi.2012.00017

Pergher, V., Wittevrongel, B., Tournoy, J., Schoenmakers, B., and Van Hulle, M. M. (2018). N-back training and transfer effects revealed by behavioral responses and EEG. *Brain Behav.* 8:e01136. doi: 10.1002/brb3.1136

Plank, M., Müller, H. J., Onton, J., Makeig, S., and Gramann, K. (2010). "Human EEG correlates of spatial navigation within egocentric and allocentric reference frames," in *International Conference on Spatial Cognition* (Portland, OR: Springer), 191–206. doi: 10.1007/978-3-642-14749-4_18

Protopapa, F., Mylonas, D., Spiliotis, K., Siettos, C., Smyrnis, N., and Evdokimidis, I. (2011). "Dynamic analysis of EEG signals during spatial working memory used for either perception discrimination or planning of action," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Boston, MA: IEEE), 5896–5899. doi: 10.1109/IEMBS.2011.6091458

Rohani, D. A., and Puthusserypady, S. (2015). BCI inside a virtual reality classroom: a potential training tool for attention. *EPJ Nonlin. Biomed. Phys.* 3:12. doi: 10.1140/epjnbp/s40366-015-0027-z

Serino, S., Morganti, F., Di Stefano, F., and Riva, G. (2015). Detecting early egocentric and allocentric impairments deficits in Alzheimer's disease: an experimental study with virtual reality. *Front. Aging Neurosci.* 7:88. doi: 10.3389/fnagi.2015.00088

Shamsuddin, S. N. W., Ugail, H., Lesk, V., and Walters, E. (2012). "VREAD: a virtual simulation to investigate cognitive function in the elderly," in *2012 International Conference on Cyberworlds* (Darmstadt: IEEE), 215–220. doi: 10.1109/CW.2012.37

Shamsuddin, S. W., Lesk, V., and Ugail, H. (2011). "Virtual environment design guidelines for elderly people in early detection of dementia," in *Proceedings of International Conference on Computer and Information Science* (Sanya: Citeseer), 751–755.

Tarnanas, I., Laskaris, N., Tsolaki, M., Muri, R., Nef, T., and Mosimann, U. P. (2015). On the comparison of a novel serious game and electroencephalography biomarkers for early dementia screening. *Adv. Exp. Med. Biol.* 821, 63–77. doi: 10.1007/978-3-319-08939-3_11

Tarnanas, I., Tsolaki, M., Nef, T., Müri, R. M., and Mosimann, U. P. (2014a). Can a novel computerized cognitive screening test provide additional information for early detection of Alzheimer's disease? *Alzheimer's Dement.* 10, 790–798. doi: 10.1016/j.jalz.2014.01.002

Tarnanas, I., Tsolakis, A., and Tsolaki, M. (2014b). "Assessing virtual reality environments as cognitive stimulation method for patients with MCI," in *Technologies of Inclusive Well-Being*, eds A. L. Brooks, S. Brahnam, and L. C. Jain (Berlin; Heidelberg: Springer), 39–74. doi: 10.1007/978-3-642-45432-5_4

Tu, S., Spiers, H. J., Hodges, J. R., Piguet, O., and Hornberger, M. (2017). Egocentric versus allocentric spatial memory in behavioral variant frontotemporal dementia and Alzheimer's disease. *J. Alzheimer's Dis.* 59, 883–892. doi: 10.3233/JAD-160592

Tu, S., Wong, S., Hodges, J. R., Irish, M., Piguet, O., and Hornberger, M. (2015). Lost in spatial translation–A novel tool to objectively assess spatial disorientation in Alzheimer's disease and frontotemporal dementia. *Cortex* 67, 83–94. doi: 10.1016/j.cortex.2015.03.016

Vourvopoulos, A., and I Badia, S. B. (2016). Motor priming in virtual reality can augment motor-imagery training efficacy in restorative brain-computer interaction: a within-subject analysis. *J. Neuroeng. Rehabil.* 13:69. doi: 10.1186/s12984-016-0173-2

Weniger, G., Ruhleder, M., Lange, C., Wolf, S., and Irle, E. (2011). Egocentric and allocentric memory as assessed by virtual reality in individuals with amnestic mild cognitive impairment. *Neuropsychologia* 49, 518–527. doi: 10.1016/j.neuropsychologia.2010.12.031

White, D. J., Congedo, M., Ciorciari, J., and Silberstein, R. B. (2012). Brain oscillatory activity during spatial navigation: theta and gamma activity link medial temporal and parietal regions. *J. Cogn. Neurosci.* 24, 686–697. doi: 10.1162/jocn_a_00098

White, P. J., and Moussavi, Z. (2016). Neurocognitive treatment for a patient with Alzheimer's disease using a virtual reality navigational environment. *J. Exp. Neurosci.* 10, 129–135. doi: 10.4137/JEN.S40827

Xu, M., Qi, H., Wan, B., Yin, T., Liu, Z., and Ming, D. (2013). A hybrid BCI speller paradigm combining P300 potential and the SSVEP blocking feature. *J. Neural Eng.* 10:026001. doi: 10.1088/1741-2560/10/2/026001

Xu, M., Xiao, X., Wang, Y., Qi, H., Jung, T.-P., and Ming, D. (2018). A brain computer interface based on miniature event-related potentials induced by very small lateral visual stimuli. *IEEE Trans. Biomed. Eng.* 65, 1166–1175. doi: 10.1109/TBME.2018.2799661

Zakzanis, K. K., Quintin, G., Graham, S. J., and Mraz, R. (2009). Age and dementia related differences in spatial navigation within an immersive virtual environment. *Med. Sci. Monit.* 15, CR140–CR150.

Zygouris, S., Ntovas, K., Giakoumis, D., Votis, K., Doumpoulakis, S., Segkouli, S., et al. (2017). A preliminary study on the feasibility of using a virtual reality cognitive training application for remote detection of mild cognitive impairment. *J. Alzheimer's Dis.* 56, 619–627. doi: 10.3233/JAD-160518

Implicit Maternal Intuition Confidence Is Associated With Maternal Well-Being Across Cultures

Wendi L. Gardner[1]*, Katie N. Rotella[2] and Janeta Nikolovski[2]

[1] *Department of Psychology, Northwestern University, Evanston, IL, United States,* [2] *Johnson & Johnson Consumer, Inc., Skillman, NJ, United States*

The transition to motherhood involves the experience of each individual mother and child, as well as the burden of cultural expectations. Social desirability demands may impede self reports of difficulties during the transition to motherhood when using traditional explicit measures. One core component of maternal role attainment is a mother's confidence in her own intuitive knowledge of her child. This brief report presents two studies that examine a "low technology" implicit measure of maternal intuition confidence that is based within a more general decision confidence paradigm. Study 1 examined the association of both implicit and explicit maternal intuition confidence with depressive symptoms, life satisfaction, and maternal identity satisfaction in a United States sample of mothers. The implicit measure contributed to variance in each of the outcome measures, above and beyond an explicit measure. Study 2 explored the association of implicit maternal intuition confidence with life satisfaction and maternal identity satisfaction in Brazil, China, India, the United States and the United Kingdom. Across all samples, implicit maternal intuition confidence was significantly associated with satisfaction with life. However, it was significantly associated with maternal identity satisfaction only in the two individualistic countries (the United States and the United Kingdom), but not in the three collectivist countries.

Keywords: implicit measures, parenting, intuition, well-being, culture

INTRODUCTION

The capacity of implicit measures to tap thoughts or feelings that yield distinct information from explicit questioning has been a boon for research on difficult topics ranging from racial prejudice to suicidality (Sleek, 2018). The majority of implicit measures rely on the assessment of computerized reaction time under cognitive conflict (IAT, Greenwald and Banaji, 1995) or after priming (AMP, Payne et al., 2005), but physiological assessments such as facial EMG (Roddy et al., 2010) and the startle blink response (Mahaffey et al., 2011) have also proven valuable.

Not all researchers who could benefit from utilizing implicit measures may have the opportunity, resources, or expertise to use technologically advanced procedures, however. The current research explores a "low technology" procedure, one that capitalizes on the experiential processes driving decision confidence under uncertainty (Schwartz, 2004). We examine the utility of this procedure to serve as an implicit measure in a domain fraught with social desirability concerns, that of the transition to motherhood.

Mothers experience strong societal pressure to exemplify maternal "ideals," including unceasing warmth, omnipotent understanding, and continuous responsiveness to their infant (Hays, 1996). For example, in one survey of postpartum bonding (Wittkowski et al., 2007) zero out of ninety-six mothers self-reported ever feeling "resentful" of their infant, perhaps implying self-censorship of explicit responses that conflict with cultural ideals. Indeed, social desirability biases in mothers' explicit self-reports emerge during pregnancy (van Bussel et al., 2010) and last through their child's adolescence (Nederhof, 1985).

Implicit measures may thus be especially valuable when applied to self-assessments of maternal feelings and beliefs. Moreover, because research on maternal well-being is often conducted during "well-baby" visits to the pediatrician's office or home visits in early childhood intervention studies, an ideal index would sidestep social desirability concerns without requiring sophisticated measurement technology.

The current work focused on a form of maternal self-assessment especially conducive to implicit measurement, that of confidence in maternal intuition. One of the greatest challenges during the transition to motherhood is learning to understand the unique signals of an infant. In the classic model of "becoming a mother" (Mercer, 2004), feelings of competence in knowing the baby's feelings and preferences (maternal intuition) are critical to both maternal well-being and the attainment of a satisfying maternal identity. Confidence in intuitive understanding of one's infant is distinct from confidence in effective parenting behaviors, yet it is the intuitive understanding that is almost universally perceived as being central to motherhood. Mesman et al. (2015) found striking convergence across twenty-six cultural groups that both accurate understanding and sensitive responsiveness to the child were core components of the "ideal mother." Given theoretical importance to maternal well-being and identity, in addition to vulnerability to social desirability concerns, an *implicit* assessment of maternal intuition confidence seems especially valuable.

Fortunately, several "low tech" implicit measurement techniques exist (see Sekaquaptewa et al., 2010, for review), including techniques designed to assess confidence in intuitive judgments. Some judgments (e.g. "What percentage of days in July are above 85 degrees?") require one's "best guess" based on limited or uncertain explicit knowledge. *Confidence* in that judgment then reflects both the feelings of confidence experienced when making the decision, and beliefs concerning one's own competence in the domain (Schwartz, 2004). Kamat (2011) capitalized on this finding to develop a measure of "general intuition confidence" designed to tap implicit feelings of self-certainty. In a two-part procedure, participants are first asked a question for which explicit knowledge of the correct answer is unavailable (e.g. "What percentage of Americans do you think would prefer to travel to Italy for vacation over Spain?"), forcing them to give the answer that intuitively feels right. Second, participants report confidence in that decision. This procedure, repeated over several decisions and confidence judgments, assesses a participants general confidence in their intuition. Hamilton et al. (2011), using a similar two-step

procedure, demonstrated that decision confidence measures (but not the decisions themselves) were influenced by experimental manipulations of identity confusion. The current work applies the two-step procedure (Kamat, 2011), specifically to the domain of "maternal intuition confidence." Each of the initial decisions require intuitive knowledge of one's own infant (e.g. "When your baby cries, what percentage of the time is it due to physical discomfort versus psychological distress"), mothers then rate their confidence in each decision.

The aim of this research was to examine the use of this decision confidence assessment as a potential implicit measure of maternal intuition confidence. Because prior work proposed that maternal intuition confidence was a cornerstone to subjective maternal satisfaction and maternal identity development (Mercer, 2004), we examined the associations of this measure with those outcomes. We also compared those associations with an explicit assessment of maternal intuition confidence (Study 1), and examined potential cultural differences in the relationship of implicit maternal intuition confidence to life satisfaction and maternal identity by sampling mothers from five different countries (Study 2).

STUDY 1

Study 1 explored whether an implicit confidence-in-intuition measure used in previous studies to assess general decision confidence could be applied to the more specialized context of maternal intuition confidence. We developed a measure that tapped mothers feelings of confidence in their decisions concerning knowledge of their baby's needs/preferences in four different childcare situations. We also assessed implicit general intuition confidence and explicit parental intuition confidence to explore the relationship among the three measures, as well as to test the hypothesis that the implicit measure of maternal intuition confidence would hold unique value in explaining variability in depressive symptoms, satisfaction with life, and maternal identity.

Method
Participants

Two hundred and seventy-two North American mothers of babies under the age of 24 months were recruited to take an online survey using Amazon's Mechanical Turk platform. Age of mothers ranged from 18 to 44 ($M = 28$). Seventy percent of participants identified as European American, with 12% identifying as Black, 12% as LatinX or Hispanic, and 6% as Asian American. The sample was roughly equal in new and experienced mothers, with 51% reporting being first-time mothers, 37% reporting having one other child, and 11% reporting having two or more additional children. Eighteen percent of mothers reported their baby was under the age of 6 months, 41% between 6 and 12 months, and 41% between 12 and 24 months.

Materials
All materials for this study may be accessed at https://osf.io/38s6t/

Implicit General Intuition Confidence

A two-part procedure (Kamat, 2011) for each of two items first presented participants with a question that participants would have no knowledge of the correct answer, thus must give the answer that intuitively feels right (e.g. "What percentage of Americans do you think would prefer to travel to Italy for vacation over Spain?" 1 = *100% Italy*; 2 = *90% Italy, 10% Spain*…11 = *100% Spain*). After each, participants then answer "How confident are you in your prior answer? " 1 = *Not at all* to 7 = *Extremely*. Confidence scores are averaged to create the index of implicit general intuition confidence (M = 3.99, SD = 1.44).

Implicit Maternal Intuition Confidence

We adapted the general intuition confidence procedure to be specific to parenting, using 4 two-part responses. In each response mothers were first asked a question concerning understanding their baby in one of four key situations (crying, feeding, sleep, and play – e.g. "When your baby cries, what percentage of the time is it due to physical discomfort versus psychological distress" 1 = *100% of crying is due to physical discomfort*; 2 = *90% physical, 10% psychological*…11 = *100% of the time is due to psychological distress*.). Each item was immediately followed by asking "How confident are you in your prior answer?" 1 = *Not at all* to 7 = *Extremely*. The four "confidence" responses were averaged to create the index of implicit maternal intuition confidence (M = 5.62, SD = 0.78).

Explicit Maternal Intuition Confidence

Adapted from the Parental Self Efficacy Scale (Gibaud-Wallston and Wandersman, 1978) e.g. "If anyone can find the answer to what is troubling my baby, I am the one" 1 = *Strongly disagree*; 6 = *Strongly agree* (M = 4.88, SD = 1.03).

Depressive Symptoms

Ten-item Center for Epidemiologic Studies Short Depression Scale. e.g. "In the past week I felt depressed." 0 = *Not at all* to 4 = *Nearly every day*. Scores are sums (M = 16.99, SD = 5.33). The item "My sleep was disrupted" was not used in the index, given sleep disruption is common and non-diagnostic in new parents.

Satisfaction with Life

Five-item scale (Diener et al., 1985). e.g. "In most ways, my life is close to ideal." 1 = *Strongly Disagree* to 7 = *Strongly Agree* (M = 5.19, SD = 1.29).

Maternal Identity Satisfaction

Adapted from Sellers et al. (1998). e.g. "Overall, being a mother is the most fulfilling aspect of my life" 1 = *Not at all true of me* to 7 = *Extremely true of me* (M = 5.87, SD = 1.22).

Procedure

After written informed consent, participants completed a survey including the measures listed above. The implicit maternal intuition confidence assessment was presented in counterbalanced order with the explicit maternal intuition confidence and implicit general intuition confidence measures. Satisfaction with life, depressive symptomology, and maternal identity satisfaction were then presented in counterbalanced order.

Results and Discussion

Hallmarks of implicit measures include (1) low correlations with their explicit counterparts, (2) distinct antecedents and/or consequences from explicit measures, explaining unique variance in associated outcomes. We first examined the correlation between the implicit and explicit maternal intuition measures, and then examined whether and how the experience of the mother (first time vs. experienced mother) and the stage of the baby (under 6 mos, 6–12 mos, 12–24 mos) were related to each. Finally, we examined whether implicit maternal intuition confidence would show unique associations with subjective well-being as well as with maternal identity satisfaction.

Implicit and explicit measures of maternal intuition confidence showed a small but significant correlation, r (272) = 0.22, p < 0.001., implying they were related, but non-redundant. Additionally, and as expected, implicit maternal intuition confidence was correlated with implicit general intuition confidence, r (272) = 0.23, p < 0.001, whereas explicit maternal intuition confidence was unrelated to implicit general intuition confidence, r (272) = −0.03, p = 0.62.

Prior research has revealed that the transition to motherhood is equally difficult for first and subsequent babies (e.g. Krieg, 2007), in part because each individual infant brings unique challenges. Qualitative studies of maternal intuition confidence also suggest that a mother's intuitive confidence in understanding her baby's signals grows over time with that individual baby (e.g. Mercer, 2004). Analyses examining whether prior experience as a mother influenced implicit maternal intuition confidence were conducted using a two Experience (first-time mother; experienced mother) ANCOVA with implicit general intuition confidence as a covariate in order to focus solely on maternal intuition confidence. No relationship was found with prior experience as a mother and implicit maternal intuition confidence (M_{new} = 5.61; SD = 0.79, $M_{experienced}$ = 5.63, SD = 0.77, p = 0.84), Explicit maternal intuition confidence also did not differ as a function of maternal experience (M_{new} = 4.94, SD = 1.01; $M_{experienced}$ = 4.81, SD = 1.05; p = 0.31). However, a three Baby-age (<6 mos; 6–12 mos; 12–24 mos) ANCOVA with implicit general intuition confidence as a covariate supported the notion that more time with the baby is important and/or that older babies are easier to understand, as implicit maternal intuition confidence significantly increased across baby-age (Ms = 5.40, 5.52, 5.82; SDs = 0.82, 0.74, 0.77) F (2, 271) = 7.06, p < 0.001. Interestingly, an ANOVA examining explicit intuition confidence did not show similar effects (Ms = 4.81, 4.87, 4.90; SDs = 1.06, 1.06, 1.01) F < 1, p = 0.88.

We conducted three linear regression analyses predicting depressive symptomology, satisfaction with life, and maternal identity satisfaction from implicit and explicit maternal intuition confidence. For each of these analyses, implicit general intuition confidence was entered on the first step, then implicit and explicit intuition confidence entered simultaneously in the second step. As **Table 1** highlights, implicit maternal intuition confidence was uniquely and significantly associated with all three outcome

TABLE 1 | Implicit and explicit maternal intuition confidence as predictors of maternal well-being and satisfaction, controlling for implicit general intuition confidence.

	Depressive symptomology	Satisfaction with life	Maternal identity satisfaction
Implicit maternal intuition confidence	$\beta = -0.13$, t (268) $= -2.09$, $p = 0.037$	$\beta = 0.18$, t (268) $= 2.89$, $p = 0.004$	$\beta = 0.16$, t (268) $= 2.63$, $p = 0.009$
Explicit maternal intuition confidence	$\beta = -0.14$, t (268) $= -2.29$, $p = 0.023$	$\beta = 0.09$, t (268) $= 1.57$, $p = 0.12$	$\beta = 0.33$, t (268) $= 5.71$, $p = 0.000$

variables, over and above the contributions of explicit maternal intuition confidence. Explicit maternal intuition confidence was significantly related to depressive symptomology and maternal identity satisfaction and was positively but non-significantly related to satisfaction with life. Taken in combination, the results of Study 1 reveal that implicit and explicit maternal intuition confidence were only modestly correlated, that implicit but not explicit maternal intuition confidence was sensitive to baby-age, and that each carried unique explanatory variance in understanding well-being as assessed by both depressive symptoms and satisfaction with life, as well as in maternal identity satisfaction.

STUDY 2

Study 2 provided a replication and extension of Study 1. We had the opportunity to include the measures of implicit maternal and general intuition confidence as well as satisfaction with life and maternal identity satisfaction in a large survey examining childcare practices in first-time mothers across five cultures (Brazil, China, India, the United Kingdom, and the United States). The United Kingdom and United States samples provide replication opportunities for the associations of implicit maternal intuition confidence in the well-being and identity measures found in United States mothers in Study 1, as both the United States and the United Kingdom represent individualist countries. Because Brazil, India, and China are all considered collectivist countries (Triandis, 1995), these samples provide a valuable extension to assess whether implicit maternal intuition confidence is similarly associated with maternal satisfaction and identity across both individualist and collectivist countries.

Competing hypotheses exist for cultural specificity vs. universality of the relationship between implicit intuition confidence and maternal well-being. Prior research has shown that internal affective information is less important to overall life satisfaction in collectivist countries (e.g. Suh et al., 1998), implying that implicit maternal intuition may not be as strongly associated with well-being in collectivist as compared to individualist countries. On the other hand, the universality of both the need to understand non-verbal infants and of the perception that accurate responsiveness toward one's baby reflects ideal motherhood (Mesman et al., 2015) may suggest that high maternal intuition confidence would be similarly beneficial for life satisfaction across countries.

Hypotheses are clearer for the potential role of intuition confidence in maternal identity. Because individualist versus collectivist countries differ in the way in which identity is defined as either reflecting internal traits and feelings versus externally validated roles in relationships and

groups (Markus and Kitayama, 1991; Gardner et al., 1999), implicit maternal intuition confidence may be more strongly related to maternal identity satisfaction in individualist than collectivist countries.

Method
Participants
Seven hundred and forty-five mothers were recruited for a paid survey through HCD Research, Inc. One hundred forty-three mothers participated from Brazil, 151 from China, 151 from India, 150 from the United Kingdom, and 150 from the United States. All participants were first-time mothers of infants under 6 months of age and were all themselves under the age of 35. Participants were prescreened for post-partum depression, and only non-depressed mothers were invited to participate.

Materials
Implicit general intuition confidence and implicit maternal intuition confidence, satisfaction with life, and identity satisfaction were identical to the measures in Study 1, with the exception that the initial implicit general intuition measure questions were changed to be appropriate to an international sample (e.g. "What percentage of *mothers in your country* do you think would prefer to travel to Italy for vacation over Spain?"). All materials were translated into the participants' language in a multi-step translation and editing process involving two native speakers (for each language) providing independent translations that were then checked for consistency (or corrected) by a third native speaker before the translated survey was finalized.

Procedure
After being prescreened for eligibility and completing written informed consent, participants completed the items above as part of a larger survey on childcare experiences.

Results and Discussion
Table 2 provides descriptive data from each country. Brazil, China, and India represent collectivist countries, and the United Kingdom and the United States represent individualist countries. To examine whether cultural differences in implicit maternal intuition confidence emerged, a two Country type (individualist, collectivist) ANCOVA with implicit general intuition confidence as a covariate was conducted. Mothers from individualistic countries had significantly lower implicit maternal intuition confidence scores than mothers from collectivist countries ($M_{ind} = 5.48$, $SD = 0.89$; $M_{coll} = 5.82$, $SD = 0.71$), F $(1,742) = 33.86$, $p < 0.001$.

Our core hypothesis concerned whether implicit maternal intuition confidence was differentially associated with satisfaction with life and maternal identity satisfaction in individualist versus

TABLE 2 | Descriptive statistics by country.

	Implicit maternal intuition confidence	Implicit general intuition confidence	Satisfaction with life	Maternal identity satisfaction
Brazil	$M = 5.80, SD = 0.81$	$M = 5.21, SD = 1.24$	$M = 5.35, SD = 1.16$	$M = 5.86, SD = 1.24$
China	$M = 6.00, SD = 0.63$	$M = 5.85, SD = 0.78$	$M = 5.26, SD = 0.96$	$M = 3.17, SD = 1.87$
India	$M = 5.94, SD = 0.65$	$M = 5.88, SD = 0.88$	$M = 6.08, SD = 0.61$	$M = 5.67, SD = 1.25$
United Kingdom	$M = 5.26, SD = 0.89$	$M = 4.76, SD = 1.42$	$M = 5.05, SD = 0.96$	$M = 5.44, SD = 1.35$
United States	$M = 5.42, SD = 0.83$	$M = 4.72, SD = 1.42$	$M = 5.37, SD = 1.09$	$M = 5.59, SD = 1.50$

collectivist countries. For these analyses, individualist-collectivist country type was entered in the first step, along with implicit general intuition confidence and the interaction term, maternal intuition confidence and its interaction with individualist-collectivist country type were then entered simultaneously in the second step. Implicit maternal intuition confidence was positively associated with satisfaction with life and was not qualified by an interaction with culture type. Follow-up analyses revealed that implicit maternal intuition confidence was positively and significantly associated with satisfaction with life (βs ranging from 0.21 in the United Kingdom to 0.35 in China) in each of the five nations studied. In contrast, an interaction between implicit maternal intuition confidence and culture type emerged for identity satisfaction, $t(739) = 2.85$, $p = 0.005$. As **Table 3** illustrates, the positive association between implicit maternal intuition confidence and maternal identity satisfaction replicated in individualist countries; no significant association emerged in collectivist countries.

GENERAL DISCUSSION

These two studies demonstrated that a "low technology" implicit measure of maternal intuition confidence was reliably associated with aspects of maternal well-being. Results from Study 1 with mothers from the United States revealed that implicit, but not explicit maternal intuition confidence increased with baby age, and that implicit and explicit measures provided unique variance in predicting depressive symptomology, satisfaction with life, and maternal identity satisfaction. Study 2 provided an opportunity to replicate the associations of implicit maternal intuition confidence with satisfaction with life and maternal identity satisfaction in mothers

TABLE 3 | Implicit maternal intuition confidence as a predictor of maternal well-being and satisfaction in individualist and collectivist countries, controlling for implicit general intuition confidence.

	Satisfaction with life	Maternal identity satisfaction
INDIVIDUALIST		
Implicit maternal intuition confidence	$\beta = 0.29, t(297) = 4.95,$ $p = 0.000$	$\beta = 0.28, t(297) = 5.01,$ $p = 0.000$
COLLECTIVIST		
Implicit maternal intuition confidence	$\beta = 0.21, t(442) = 3.90,$ $p = 0.000$	$\beta = -0.02, t(442) = -0.41,$ $p = 0.684$

from individualistic countries, and to extend the findings to examine mothers from collectivist countries. The positive association between implicit maternal intuition confidence and satisfaction with life was replicated across all countries, regardless of individualism or collectivism. In contrast, the positive association with maternal identity satisfaction was only replicated in mothers from individualist cultures.

Different findings of the influence of culture for well-being versus identity may point to both the universally distressing nature of feeling one doesn't understand one's baby and the culturally distinct processes that underlay identity development. Because members of individualist cultures derive their sense of self from internal attributes and feelings, it seems sensible that implicit maternal intuition confidence plays a role in how satisfied mothers from those cultures feel with their identity. In contrast, members of collectivist cultures derive their sense of self more strongly from others' perceptions that they are fulfilling ascribed roles; it is possible that internal feeling states are not as important to maternal identity satisfaction, despite being important for overall life satisfaction. For mothers in collectivist cultures, identity may be scaffolded more upon external social aspects, such as being seen as adequate in the role by close others. In contrast, for mothers in individualist cultures identity may be more closely tied to self-feelings. This interpretation is speculative; these results must be replicated before we can be confident that cultural distinctions in the development of maternal identity satisfaction exist. Nonetheless, this work is the first, to our knowledge, to compare this aspect of Mercer's (2004) model in such a wide sample of countries.

Whereas cultural differences in the association of maternal intuition confidence to identity satisfaction may reflect previously seen patterns in the importance of internal versus external aspects of identity (Markus and Kitayama, 1991), it is more difficult to speculate on the differences in maternal intuition confidence scores themselves. Although greater self-confidence is typically seen in individualist as compared to collectivist countries (Heine, 2005), specific confidence estimates for intuitive judgments are often higher in collectivist countries (Yates and de Oliveira, 2016). The differences observed here should be replicated. Additionally, because information on some cultural samples was limited (e.g. no information on racial background in Brazil), national differences should also be interpreted with caution.

Other limitations of the studies include their cross-sectional nature. The current findings suggesting a relationship between baby-age and implicit maternal intuition confidence, as well as associations of implicit maternal intuition confidence and

well-being are both consistent with key aspects of Mercer's (2004) maternal role attainment theory -namely that mothers grow increasingly more confident in their abilities to decipher the signals and needs of their infants across time, and this confidence undergirds satisfaction with motherhood. However, longitudinal studies examining the factors underlaying the growth of maternal intuition confidence (such as increased clarity of the baby's communications) and alterations in maternal well-being and identity across time are needed to fully explore these dynamics. Additionally, the interplay between maternal intuition confidence (the feeling of knowing what one's baby is thinking/feeling) and more behavioral competencies such as parental efficacy (confidence one can care for and soothe one's baby) should be investigated as they grow over time, and as they relate to emotional well-being and identity satisfaction during the transition to parenthood. Parental efficacy has also been linked to depressive symptoms. However, unlike maternal intuition which is hypothesized to be developed anew with each individual child, parental efficacy has been found to increase with subsequent pregnancies (Leahy-Warren et al., 2012). Finally, it is likely that the relationship between maternal intuition confidence and well-being is bidirectional. Positive emotions both broaden attention and encourage flexible behavioral responses (Fredrickson, 2013), qualities potentially useful in developing maternal intuition confidence. It is thus possible that an upward spiral of maternal intuition confidence boosting well-being, and well-being in-turn boosting maternal intuition confidence could characterize the most positive transitions to motherhood.

Although implicit measures of maternal intuition confidence may be useful, in part, through bypassing unwillingness to report being less than an "ideal mother"(Mesman et al., 2015), we are not interpreting the modest correlation between implicit and explicit maternal intuition confidence as necessarily reflecting conscious deception. Implicit measures also tap into mental processes that individuals are unable (not just unwilling) to report, and it remains unclear how much the current measure reflects concerns less accessible to consciousness, as opposed to unreported due to social desirability. Future research examining the experiences and situations that support or threaten implicit maternal intuition confidence would provide valuable insight into this question. Additionally, future work using implicit measures to examine even more complicated and socially fraught maternal feelings (e.g. resentment toward their infant; Wittkowski et al., 2007) could allow novel examinations of factors underlying more negative experiences of parenting, such as burn-out.

Finally, the current research focused specifically on mothers, excluding fathers and other caregivers. In the majority of societies, mothers are the primary caregivers of infants and spend more time on infant-care than fathers even when both parents work outside the home (Raley et al., 2012). However, the study of parental intuition confidence should not be limited to mothers. Recent research has shown that, for fathers, the importance of parenthood to identity (Pew Research Center, 2015) and the prevalence of depression following the birth of an infant (DaCosta et al., 2019) are both increasing. Models of

the transition to parenthood that were initially developed on mothers should be extended to include the experience of fathers – including examining the role of parental intuition confidence in well-being and identity.

Despite the limitations of the current work, results suggest a set of relatively simple and "low tech" decision confidence questions can reveal important psychological differences in maternal self-doubt vs. assurance. Due to relative ease of administration, this measure could be used in a broader range of environments (e.g. paper and pencil, verbal interview) than those that rely on response time or physiology. Implicit measures may be especially valuable when studying the transition to parenthood given both cognitive exhaustion and social desirability concerns that may prevent explicit self-report measures from capturing the full range of mothers' experiences.

DATA AVAILABILITY STATEMENT

The datasets generated and all materials used for these studies are available by request from the corresponding author.

ETHICS STATEMENT

All research was conducted in accordance with the ethical standards of the American Psychological Association. All participants gave written informed consent in accordance with the Declaration of Helsinki. The protocol for Study 1 was approved by the Institutional Review Board of Northwestern University, and for Study 2 by the Institutional Review Board of Integreview.com to ensure national ethics committee requirements for data collection in participating countries.

AUTHOR CONTRIBUTIONS

WG was responsible for data collection for Study 1, analyzed the data for Studies 1 and 2 and prepared the draft and final manuscript. KR and JN were responsible for overseeing data collection for Study 2 and provided feedback on analyses as well as on draft manuscripts. All authors provided critical feedback and helped shape the overall research.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

DaCosta, D., Danieli, C., Abrahamowicz, M., Dasgupta, K., Sewitch, M., Lowensteyn, I., et al. (2019). A prospective study of postnatal depressive symptoms and associated risk factors in first-time fathers. *J. Affect. Disord.* 249, 371–377. doi: 10.1016/j.jad.2019.02.033

Diener, E. D., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The satisfaction with life scale. *J. Pers. Assess.* 49, 71–75. doi: 10.1207/s15327752jpa4901-13

Fredrickson, B. L. (2013). Positive emotions broaden and build. *Adv. Exp. Soc. Psychol.* 47, 1–53. doi: 10.1016/B978-0-12-407236-7.00001-2

Gardner, W. L., Gabriel, S., and Lee, A. Y. (1999). "I" value freedom but "We" value relationships: self-construal priming mirrors cultural differences in judgment. *Psychol. Sci.* 10, 321–326. doi: 10.1111/1467-9280.00162

Gibaud-Wallston, J., and Wandersman, L. P. (1978). Development and utility of the parenting sense of competence scale. *Paper Presented at the Meeting of the American Psychological Association*, Toronto.

Greenwald, A. G., and Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychol. Rev.* 102, 4–27. doi: 10.1037/0033-295X.102.1.4

Hamilton, R., Vohs, K. D., Sellier, A. L., and Meyvis, T. (2011). Being of two minds: switching mindsets exhausts self-regulatory resources. *Organ. Behav. Hum. Decis. Process.* 115, 13–24. doi: 10.1016/j.obhdp.2010.11.005

Hays, S. (1996). *The Cultural Contradictions of Motherhood*. New Haven, CT: Yale University Press.

Heine, S. J. (2005). Where is the evidence for pancultural self-enhancement? A reply to Sedikides, Gaertner, and Toguchi (2003). *J. Pers. Soc. Psychol.* 89, 531–538. doi: 10.1037/0022-3514.89.4.531

Kamat, P. (2011). *When Worlds Collide: The Self-Regulatory Costs of Switching Between Two Distinct Identities*. Masters thesis, Northwestern University, Evanston, IL.

Krieg, D. B. (2007). Does motherhood get easier the second-time around? examining parenting stress and marital quality among mothers having their first or second child. *Parenting* 7, 149–175. doi: 10.1080/15295190701306912

Leahy-Warren, P., McCarthy, G., and Corcoran, P. (2012). First-time mothers: social support, maternal parental self-efficacy and postnatal depression. *J. Clin. Nurs.* 21, 388–397. doi: 10.1111/j.1365-2702.2011.03701.x

Mahaffey, A. L., Bryan, A., Ito, T. A., and Hutchison, K. E. (2011). In search of the defensive function of sexual prejudice: exploring antigay bias through shorter lead and loner lead startle eye blink. *J. Appl. Soc. Psychol.* 41, 27–44. doi: 10.1111/j.1559-1816.2010.00700.x

Markus, H. R., and Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychol. Rev.* 98, 224–253. doi: 10.1037/0033-295X.98.2.224

Mercer, M. T. (2004). Becoming a mother versus maternal role attainment. *J. Nurs. Scholarsh.* 36, 226–232. doi: 10.1111/j.1547-5069.2004.04042.x

Mesman, J., van Ijzendoorn, M., Behrens, K., Carbonell, O., Cárcamo, R., Cohen-Paraira, I., et al. (2015). Is the ideal mother a sensitive mother? Beliefs about early childhood parenting in mothers across the globe. *Int. J. Behav. Dev.* 40, 385–397. doi: 10.1177/0165025415594030

Nederhof, A. (1985). Methods of coping with social desirability bias: a review. *Eur. J. Soc. Psychol.* 15, 263–280. doi: 10.1002/ejsp.2420150303

Payne, B. K., Cheng, C. M., Govorun, O., and Stewart, B. (2005). An inkblot for attitudes: affect misattribution as implicit measurement. *J. Pers. Soc. Psychol.* 89, 277–293. doi: 10.1037/0022-3514.89.3.277

Pew Research Center, (2015). *Parenting in America: Outlook, Worries, Aspirations Are Strongly Linked to Financial Situation*. Available at: http://www.pewsocialtrends.org/2015/12/17/parenting-in-america (accessed August 10, 2019).

Raley, S., Bianchi, S. M., and Wang, W. (2012). When do fathers care? Mothers' economic contribute and fathers' involvement in child care. *Am. J. Sociol.* 117, 1422–1459. doi: 10.1086/663354

Roddy, S., Stewart, I., and Barnes-Holmes, D. (2010). Anti-fat, pro-slim, or both? Using two reaction-time based measures to assess implicit attitudes to the slim and overweight. *J. Health Psychol.* 15, 416–425. doi: 10.1177/1359105309350232

Schwartz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *J. Consum. Psychol.* 14, 332–348. doi: 10.1207/s15327663jcp1404-2

Sekaquaptewa, D., Vargas, P., and von Hippel, W. (2010). "A practical guide to paper-and-pencil implicit measures of attitudes," in *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*, eds B. Gawronski, and B. K. Payne, (New York, NY: The Guilford Press), 140–155.

Sellers, R. M., Smith, M., Shelton, J. N., Rowley, S. J., and Chavous, T. M. (1998). Multidimensional model of racial identity: a reconceptualization of African American racial identity. *Pers. Soc. Psychol. Rev.* 2, 18–39. doi: 10.1207/s15327957pspr0201-2

Sleek, S. (2018). The bias beneath: two decades of measuring implicit associations. *Observer* 31, 11–14.

Suh, E., Diener, E., Oishi, S., and Triandis, H. C. (1998). The shifting basis of life satisfaction judgments across cultures: emotions versus norms. *J. Pers. Soc. Psychol.* 74, 482–493. doi: 10.1037/0022-3514.74.2.482

Triandis, H. C. (1995). *Individualism and Collectivism. New Directions in Social Psychology*. Boulder, CO: Westview Press.

van Bussel, J. C. H., Spitz, B., and Demyttenaere, K. (2010). Three self-report questionnaires of the early mother-to-infant bond: reliability and validity of the Dutch version of the MPAS, PBQ and MIBS. *Arch. Women Ment. Health* 13, 373–384. doi: 10.1007/s00737-009-0140-z

Wittkowski, A., Wieck, A., and Mann, S. (2007). An evaluation of two bonding questionnaires: a comparison of the mother-to-infant bonding scale with the postpartum bonding questionnaire in a sample of primiparous mothers. *Arch. Women Ment. Health* 10, 171–175. doi: 10.1007/s00737-007-0191-y

Yates, J. F., and de Oliveira, S. (2016). Culture and decision making. *Organ. Behav. Hum. Decis. Process.* 136, 106–118. doi: 10.1016/j.obhdp.2016.05.003

Questionnaire Measures and Physiological Correlates of Presence: A Systematic Review

Simone Grassini* and Karin Laumann

Department of Psychology, Norwegian University of Science and Technology, Trondheim, Norway

The published literature has produced several definitions for the sense of presence in a simulated environment, as well as various methods for measuring it. The variety of conceptualizations makes it difficult for researchers to interpret, compare, and evaluate the presence ratings obtained from individual studies. Presence has been measured by employing questionnaires, physiological indices, behavioral feedbacks, and interviews. A systematic literature review was conducted to provide insight into the definitions and measurements of presence in studies from 2002 to 2019, with a focus on questionnaires and physiological measures. The review showed that scholars had introduced various definitions of presence that often originate from different theoretical standpoints and that this has produced a multitude of different questionnaires that aim to measure presence. At the same time, physiological studies that investigate the physiological correlates of the sense of presence have often shown ambiguous results or have not been replicated. Most of the scholars have preferred the use of questionnaires, with Witmer and Singer's Presence Questionnaire being the most prevalent. Among the physiological measures, electroencephalography was the most frequently used. The conclusions of the present review aim to stimulate future structured efforts to standardize the use of the construct of presence, as well as inspire the replication of the findings reported in the published literature.

Keywords: virtual environment, presence, immersion, physiology, review

INTRODUCTION

Presence, which refers to the sense of "being there" in a simulated environment, is a critical concept in the discussion of new technologies and mediated environments (Cummings and Bailenson, 2016). In the academic and industrial communities, there is often an underlying assumption that the main goal for designing virtual environments (VEs, i.e., those environments generated by a computer that simulate some characteristic of reality) is to promote a sense of presence. Cummings and Bailenson (2016) noted that a heightened sense of presence enhances the user's capacity for interaction with the simulation. VEs have found applications in many fields, including clinical therapy, training, learning, and entertainment, due to their capacity to elicit a high degree of presence (Slater and Wilbur, 1997; Nunez and Blake, 2001; Tamborini and Skalski, 2006; Price and Anderson, 2007).

A relatively recent analysis (Cummings and Bailenson, 2016) noted that the systematic investigation of presence as a psychological phenomenon is quite new in the scientific literature. The comprehensive works of Biocca (1997, 1999) are among the earliest scientific efforts aiming

to discuss specific characteristics of the sense of presence (such as, e.g., the relationship between mind and embodiment in a virtual medium, the sense of physical presence in a simulated space, and the physical and social sense of presence). Only in the last two decades, starting with the work of Lee (2004), has there been an attempt to provide a more exhaustive explication of presence as a psychological phenomenon by introducing concepts such as social presence (the extent to which other entities presented in the VE "are there" from the user point of view) and self-presence (the sense of the user being able to perceive him/herself as part of the VE).

Researchers have conceptualized the sense of presence in different ways (e.g., Steuer, 1992; Slater and Wilbur, 1997; Witmer and Singer, 1998; Slater, 1999; McMahan, 2003). Furthermore, several terminologies (e.g., telepresence, mediated presence, virtual presence) have been used to refer to the same notion (Lee, 2004). It is also common that the terminologies "presence" and "immersion" are used as synonyms, primarily when used outside the area traditionally covered by psychology (see, e.g., Jennett et al., 2008; Cheng et al., 2015). Generally, the term presence applies to a broad family of phenomena primarily experienced during the use of VE, though this is also reported during the use of different types of displays and immersive media. VE users that experience a high sense of presence often report the feeling of being in a different place than the position in space occupied by their physical body (Slater et al., 1994). Some scholars have proposed that the concept of presence should be treated as a type of perceptual outsourcing or distal attribution (see Loomis, 1992).

It has been a challenge for researchers to develop a widely accepted and unified notion of the phenomenon. Nevertheless, Lee (2004) made a detailed effort, through a comprehensive clarification process, to provide a coherent understanding and a global description of presence (Makransky et al., 2017a). As indicated by the work of Lee (2004), the experiences in real life and those in VEs can be separated into three distinct domains: physical, social, and self. With the term "physical experience," Lee (2004) refers to the experience of the physical environment and the objects situated in the environment. The social experience domain refers to the experience of other entities with social value in the environment. The self-experience domain is applied to all those experiences that the user has of him/herself.

Furthermore, Lee (2004) hypothesized two different paths by which an experience can be defined as a virtual one. First, an experience can be considered virtual when an artificial technology mediates it. Such artificial technology allows the users to have an experience of a mediated version of the natural, physical space (as in the case of mixed reality). Another VE case is when the human-made equipment enables the experience of environments and entities that do not exist at all in the real world [as in the case of virtual reality (VR)].

Several studies have explored the sense of presence and its possible physiological correlates, but there is a lack of an overview and critical analysis of the various methods. Furthermore, an effort to replicate the findings reported in the literature is missing.

The present review intends to fill the gap through an analysis of the various methodologies used for indexing the sense of presence, with a focus on the questionnaire and physiological methods. Furthermore, it aims to update the results of the comprehensive review on questionnaire use for evaluating the sense of presence published by Hein et al. (2018), as well as previous outdated works (see Insko, 2003) that examined the physiological correlates of the sense of presence. The present investigation asked the following. (1) Which questionnaires and physiological methods have been used together to evaluate the sense of presence? (2) What are the advantages, possible problems, and criticalities of using these methodologies? The answers to these questions will hopefully promote the development of the field as well as the development of a more unified and coherent theoretical research framework for the scientific study of the sense of presence.

Terminological Clarifications

Defining what presence is and how related concepts can be differentiated from it has posed challenges for researchers. The constructs of immersion, involvement, and emotion are difficult to disentangle from presence itself (Slater, 2003). Several studies have considered the difference between these terms and related constructs (Nacke and Lindley, 2008; Cummings and Bailenson, 2016; Suvajdzc et al., 2018). Generally, immersion refers to the outcome of immersive technology when evaluated objectively (Nilsson et al., 2016). Briefly, a system is more immersive depending on the numbers and quality of delivering displays, in all modalities, and preservation of visual fidelity that is similar to the real world (Cummings and Bailenson, 2016). Technical parameters are related to the level of immersion. Some examples are the field of view, image latency, and frame rate of the image stream (Slater, 1999). On the other hand, presence has been defined as the subjective and psychological reaction to immersive environments (Fromberger et al., 2015). As already introduced in the previous section of the present article, some studies have systematically used the terminology "immersion" and "presence" as synonyms (e.g., Jennett et al., 2008; Amin et al., 2016; Papachristos et al., 2017; Lum et al., 2018). In the present review, the word "presence" will be preferred for the description of the psychological and subjective feeling of being in a VE.

Presence and involvement can be logically separated into two different psychological phenomena (Grabarczyk and Pokropski, 2016). As in the everyday experience, one can be involved in something but not believe that one is present in it (for example, while reading a book or watching a movie; Slater, 2003). Presence involves a variety of emotions, and thus it can be measured by examining the emotional reactions promoted by VE. As noted by Riva et al. (2007), what someone sees may engage him/her emotionally. However, emotional engagement does not qualify as presence *per se*. Various studies have discussed the connection between presence and emotion (Ijsselsteijn et al., 2000; Banos et al., 2008).

Many scholars have used the terms VR and VE synonymously. Although they share similarities, they differ in their context and history (see Luciani and Cadoz, 2007). In the present review, VE will be preferred to describe any immersive visual technology (IVT), while VR will be used to describe those modern IVTs that utilize head-mounted displays (HMD or VR headsets).

Presence and Human Performance

The importance of studying the sense of presence goes beyond mere scientific curiosity; such study is directly related to applicable research. In this regard, several published studies have revealed a positive relationship between presence and human performance (Baumgartner et al., 2006; Baus and Bouchard, 2017). Questionnaires evaluating the sense of presence (e.g., PQ and SUS) showed it to positively correlate with human performance, and the positive association between questionnaire scores and user performance has been directly used as an argument for a good predictive validity of the questionnaire measures.

As noted by Serafin et al. (2016), presence is essential in VEs because it easily influences the user's behavior, and this factor can be used to modulate the learning performance of users when VEs are used for training. For instance, if the VE is used to train professionals, such as surgeons and firefighters, presence will be vital, since they are expected to perform in the VE similarly to how they will perform in the real world. The study by Stevens and Kincaid (2015) investigated the relationship between performance and presence in virtual simulation training and showed that a high level of presence during virtual training correlates positively with an increase in user performance while performing the trained task in real life.

However, there are some instances in which presence can negatively affect performance (Sharples et al., 2008). A high sense of presence correlates with simulator/cybersickness (Lin et al., 2002), which is a collection of undesirable symptoms often reported by users exposed to VE. However, please note that cybersickness and presence are generally inversely correlated, as reported by a recently published literature review (Weech et al., 2019). Cybersickness negatively affects performance because it triggers disorientation, nausea, and oculomotor symptoms in the user, causing discomfort (Kennedy et al., 1993). Finally, presence may not facilitate every type of performance. For instance, some studies (Mania and Chalmers, 2001; Makransky et al., 2017b) found no association or an inverse correlation between presence and learning outcomes.

Moreover, the sense of presence may not be directly related to task performance: task performance may be more associated with the actual manipulation in the experimental scenario or related to technical characteristics of the human-computer interface (e.g., input lag, visual quality, et cetera). These characteristics, even if connected with the sense of presence (i.e., we can assume that better visual quality generally increases the sense of presence), do not directly represent the subjective sense of presence.

Performance in a VR scenario was directly investigated by a recent study (Rose and Chen, 2018) using different degrees of simulation vividness to modulate subjective presence in the simulation. Nevertheless, the study did not find a relationship between the quality of the simulation (and, consequently, the level of reported presence) and actual (objectively measured) task performance in the VR. Organizations that look to introduce the use of VR into their work routine for training or performance purposes are increasingly interested in studying the possible advantages and disadvantages of the technology at the current

stage of research. A recent work (Pallavicini et al., 2019) showed that the visualization of a VE using highly immersive media increases both presence and emotional response. Nevertheless, for some practical applications, especially in highly stressful work scenarios, a high degree of presence and emotional involvement in the simulation may not be a helpful factor, as emotionality may be disadvantageous for training activities or for task performance in VR.

Measures of Presence

Various studies have attempted to measure presence in a laboratory setting, and four main investigation methods (as noted already by Hein et al., 2018) can be distinguished in the literature: questionnaires, physiological measures, analysis of the user's behavior, and interviews. Baren and Ijsselsteijn (2004) have presented a comprehensive list of the methods (even though outdated). Questionnaires and physiology are two typologies of measures for the assessment of the sense of presence that are often used together. However, several factors (e.g., the many questionnaires used to measure presence and the lack of a standard for the analysis of physiological data), have made a comparison between different studies challenging. The present article tries to analyze the criticalities of using those two methodologies, focusing on the published literature on presence where questionnaires and physiological measures were reported together.

Use of Questionnaires

Questionnaires are the most frequently used method for the investigation of presence (see Hein et al., 2018). A standard design employed in studies using questionnaires is to make the experimental participant engage passively or actively in a VE and later ask the participant to answer a survey that evaluates his/her experience. Usually, the questions consist of ordinal scales (e.g., Witmer and Singer, 1998; Lessiter et al., 2001). Some presence questionnaires share similarities: for example, they use a Likert scale that ranges from 1 and 7 [such as the Presence Questionnaire (PQ), Igroup Presence Questionnaire (IPQ), and Slater-Usoh-Steed Questionnaire (SUS)]. Questionnaire instruments have several advantages compared to other methods of investigating presence. They are cheap, easy to administer, and they are applicable regardless of the VE used. Questionnaires do not require lengthy prior preparation, as other methodologies do, and do not require specialized skills or scientific instruments (as physiological measures). On the other hand, the numbers of questionnaires measuring presence, the variety of constructs explored by these questionnaires, and the overall lack of a standard definition for presence as a psychological construct may represent a problem on the use of this methodology. Furthermore, results from studies using different questionnaires may be difficult to compare (Kober and Neuper, 2013). Moreover, questionnaires assessing the degree of presence are susceptible to response bias: where a questionnaire poses queries about presence directly or indirectly, it may possibly load an answer that would not otherwise have reached the participants' conscious level (Szczurowski and Smith, 2017).

Use of Physiological Measures

There have been numerous attempts to use human physiological measures as indices for presence (e.g., Meehan et al., 2002; Arndt et al., 2016), and several different types of measures have been identified. These physiological indices can be coarsely divided into two families: brain-related and not brain-related. Among the brain-related measures, electroencephalography (EEG) is one of the most commonly used within the field of cognitive science and has found extensive use in relation to the sense of presence (e.g., Terkildsen and Makransky, 2019). EEG measures the electrical activity of the human brain in a passive and non-invasive way: many neurons disposed perpendicularly to the scalp and firing at the same time produce an electrical potential that is possible to measure from outside the scalp (Breedlove et al., 2010). EEG signals can be analyzed in several different ways. Continuous EEG signals can be divided into frequency bands (usually delta, theta, alpha, beta, and gamma), and those oscillatory neural activities can be interpreted, for example, in connection with human behavior and cognitive processes (Teplan, 2002). Presence has been often investigated using the EEG-based technique of event-related potentials (ERPs; see Kober and Neuper, 2012; Terkildsen and Makransky, 2019). ERPs represent brain activity generated as a response to an event (a stimulation that can be, e.g., visual or auditory). This activity is generally averaged across many samples (trials), in order to reduce signal noise and obtain a reliable estimate for the brain activity related to the response to the stimulation (Andreassi, 2010). The ERP methodology takes advantage of the good time resolution (milliseconds) of the EEG recording and it is widely used in cognitive research (attention, perception, consciousness, etc.). However, EEG has somewhat imprecise signal localization (low spatial resolution). EEG is sometimes also utilized to identify the physical sources of brain signals and connectivity among brain areas (see, e.g., Greicius et al., 2003), and this methodology has also been implemented for the spatial individuation of the physiological correlates of presence (Clemente et al., 2013b).

EEG offers several advantages. It is relatively affordable, even for unspecialized laboratories. The latest developments of portable consumer-grade equipment have made EEG equipment cost-effective and easy to operate. However, the setup of the equipment for the experimental phase is rather lengthy compared with other methods, and data analysis requires specialized expertise. Furthermore, the recording is somewhat sensitive to movements, and data quality could be impaired in VEs where movements are essential. For the latter reason, EEG is not adaptable to all kinds of VEs.

A further brain-related physiological measure that is used for the study of presence is functional magnetic resonance imaging (fMRI; Hoffman et al., 2003). fMRI is a brain imaging technique that allows for spatially precise (millimeters) identification of activity in the human brain. The fMRI scanner can identify the flow of oxygenated blood in the brain in a relatively short amount of time (seconds). This identification occurs due to blood-oxygenation-dependent imaging (BOLD), which highlights the activated brain areas. However, the spatial precision of fMRI comes at the cost of temporal resolution, which is much lower than that of EEG (Huettel et al., 2004). Furthermore, fMRI is very expensive and is generally used for clinical purposes, sometimes only available within hospital infrastructures. The machinery is complex to operate, and specialized medical staff are often required for its correct operation and to limit possible usage risks, further increasing its operational costs. Moreover, there are some non-negligible risks and restrictions for the participants of fMRI studies. For example, a participant cannot participate in fMRI studies if he/she has permanent metal prostheses on his/her body (or, e.g., a pacemaker), as the magnetic field of the scanner may interact with the metal. Eventual presentation of the stimuli (e.g., via VR goggles) needs to be mediated by MRI-scan compatible equipment, i.e., not interacting with magnetic fields. Additionally, the experiment is performed in an unnatural lying-down position, and the head of the subject needs to be immobilized, often causing discomfort. These limitations make this methodology challenging to apply in most, if not all, VEs.

Generally, brain-related physiological measures are more expensive and less adaptable compared to non-brain physiological measures. Several studies have explored the use of galvanic skin response (GSR), also known as electrodermal activity (EDA) or SC [sometimes in the literature as skin conductance responses (SCRs)], which measures how the electrical variations in the skin trigger eccrine sweat glands, a phenomenon that allows SC measurement. In the present article, the abbreviation "SC" will be used to refer to this methodology. The use of SC is well-documented in the literature on human emotion and cognition (Weber et al., 2009; Poels et al., 2012; Chalfoun and Dankoff, 2018). SC is associated, for example, with stress, excitement, engagement, and frustration, and arousal, among other factors (see e.g., Kurniawan et al., 2013).

Furthermore, stimuli that promote attentional processes and attention-demanding tasks relate to several characteristics of the SC signal. SC data can be analyzed in several ways (e.g., decomposing phasic-transient and tonic activity), even though researchers are still in the process of understanding the exact meanings of those different components of SC (see Nagai et al., 2004; Braithwaite et al., 2013; Dawson et al., 2017). SC is easy to set up and minimally invasive, while SC sensors are affordable. However, SC data are sensitive to movements (especially of the part of the body where the sensors are attached) and to all those activities that may modulate the activity of the eccrine sweat glands. Furthermore, SC may not be an optimal proxy for the measure of presence (based on the assumption that arousal/excitement correlates with a higher sense of presence), as SC modulation is highly dependent on the content of the immersive experience. SC can also be significantly modulated by external (often non-controlled) factors, like room temperature or environmental humidity (Boucsein, 2012).

Heart rate (HR) has also been investigated as a possible correlate of presence. In the present review, we do not focus on the various HR analysis methods. However, the most used methods in psychophysiological research are heart rate (HR) and heart-rate variability (HRV). HR is the calculation/estimation of the average heartbeats per timeframe (generally 1 min). HRV, instead, is the measure (in milliseconds) of the changes (i.e., the variability) between successive heartbeats. This time period

is called the R-R interval (inter-beat interval; van Ravenswaaij-Arts et al., 1993). The experimental methodology generally preferred for HR studies is electrocardiography (ECG or EKG). ECG records changes in electrical potential associated with the heartbeat (Goldberger, 1998). Due to its affordability compared to ECG, PPG (photoplethysmography) was also used to measure heart rhythm. The PPG can detect blood volume changes in the microvascular tissue, and it is often measured using a pulse oximeter, which illuminates the skin and can detect changes in the light absorption and, with that, indirectly measures the heart rhythm. Heart rate sensors are becoming increasingly cheap, especially PPG sensors. HR is generally less affected by movement than EEG and SC. On the other hand, PPG provides fewer analysis possibilities compared to ECG, giving only the number of heartbeats per time but without giving information on the beat components. However, ECG is more expensive, more challenging to operate, and more invasive than PPG.

Skin temperature (ST) is a rather simple experimental methodology in which a sensor records the temperature of the participant. Even though the use of skin temperature in psychological research date back many years (see Maslach et al., 1972) and has attracted attention in cognitive science research due to its cost-effectiveness and simplicity of use (Lara et al., 2018), this method has not found as many applications as the other methodologies listed above.

Electromyography (EMG) is a technique for recording the electrical activity produced by skeletal muscles. Its setup is easy and inexpensive. However, it finds limited application within the field of cognitive psychology. As subject muscle activity is strongly related to the EMG signal, it has been extensively used in clinical and sports medicine (Steele, 2012). However, due to its connection to subject behavior (muscle movements), it has generally been less used to study cognitive phenomena (with the exclusion of the quite widely used technique of facial electromyography; see Durso et al., 2012). Its robust connection with human behavioral responses makes this technique difficult to use in many VR contexts.

While physiological measures could be more objective indices of the level of presence experienced by a subject, there is no consensus on which measure is the best to use. Furthermore, many of these measures have methodological limitations (e.g., requiring the subject to be still, requiring a long preparation time, and being costly) and therefore cannot be applied for all situations and in all VEs.

Behavioral Measures

A third approach for measuring presence is behavioral. It is a sign of a higher level of presence when participants in a VE behave as if they are in a real environment. Some examples of this phenomenon include user behaviors related to conflicting multisensory cues that emanate from both real and simulated environments (Slater, 2002), postural sway (Freeman et al., 2017), and responses to simulated stimuli (Bouchard et al., 2008). Typically, behavioral measures require the introduction of features or tasks into the environment to elicit behavioral responses. For the behavioral approach to be applicable, the VE must present features that trigger either voluntary (e.g., pressing

of a button, as in the case of the evaluation of performance; see Slater and Wilbur, 1997), or involuntary (e.g., motion in reaction to an approaching object; see Sanchez-Vives and Slater, 2005) responses. However, these additional tasks or events may be detrimental to the measure of presence. These tasks may not be relevant for the VE, may disrupt the feeling of immersion, or may interact with the VE in uncontrolled ways.

Interviews

Hein et al. (2018), in their review, reported that only one study was found that made use of interviews without using a questionnaire. However, even in that case, the presence factors from the interview shared similarities with the typical items of the Presence Questionnaire (PQ, Witmer and Singer, 1998).

METHODS

The methodology utilized here was a systematic literature review of published research articles, following the guidelines stated in the Preferred Reporting Items for Systematic Reviews (PRISMA; see Liberati et al., 2009). The method used for the literature review was based on that reported in the PRISMA checklist. The first step defined the research questions and the setting of the study protocol. Second, relevant research manuscripts were identified, considering only papers authored during the last two decades. Those articles were then screened in different stages, based on their titles, abstracts, and keywords. Afterward, a sub-sample of articles were selected, and their full texts were evaluated. Finally, the information from the reviewed research was summarized and categorized into the topic of interest for the current review.

Inclusion Criteria and Search Strategy

The inclusion and exclusion criteria were established before the literature review. The inclusion criteria focused on the factors that it was important to evaluate when selecting an article to be part of this review. They limited the articles based on their research design, data, and focus. Research articles that directly investigated the sense of presence were considered for the analysis of measurement approaches to presence. Further criteria were applied to target the articles of interest. First, only papers published from January 2002 to August 2019 were included in order to review the most current literature. This time frame was chosen to not critically overlap with the analysis of a similar review article published in 2003 (Insko, 2003) and to report the current state-of-the-art research. Second, the reviewed manuscripts had to contain an empirical study in which there was some VE experienced by participants. Furthermore, only those studies that reported first-hand experimental data were selected.

The search for relevant papers was conducted using some of the most popular research engines for academic articles: Web of Science, Scopus, and Google Scholar. These scientific databases were chosen for their popularity and for their comprehensive and interdisciplinary nature, which fit the interdisciplinary topic of the present review. Only articles in English were considered for the present review. Only articles published in peer-reviewed

journals or conference proceedings were selected during the preliminary screening phase.

During the first stage, the keywords used for the search included "virtual reality" AND "presence," "virtual reality" AND "immersion," "virtual environment" AND "presence," and "virtual environment" AND "immersion." For Web of Science, the articles in the "Web of Science core collection" database were searched. The queries made by using the "Topic" field, which includes title, abstract, and keywords. Both published articles and conference proceedings were searched for. After the preliminary screening, the results obtained included 3,466 articles. To narrow down the results, the articles were filtered for the categories "Neurosciences," "Psychology multidisciplinary," "Psychology applied," "Psychology experimental," "Psychology," and "Psychology biological." After duplications from every query were eliminated, the number of articles included in the preliminary selection from Web of Science was 371.

For Scopus, the queries were run on the article title, abstract, and keywords of published scientific articles. After the first screening, 6,546 were retrieved. The subject areas "Psychology" and "Neurosciences" were selected to pre-screen those articles. After the filters were applied, 651 results were retrieved.

For Google Scholar, each combination of keywords was queried separately, and the first 10 pages of resulting titles were manually reviewed for each one of the queries. Each page contained 10 results, and the results were sorted by relevance. A total of 140 articles were selected.

The retrieved articles from the three search engines were then imported into Endnote software, and duplicate entries were removed. A total of 875 unique articles were identified. The titles of these articles were screened to distinguish papers fitting the topic of investigation, and 205 articles were selected in the preliminary screening phase. Preliminary screening of keywords, titles, and abstracts was performed to filter the articles further. From this further screening, 120 papers were selected. The full abstract of these articles was then evaluated for inclusion and exclusion criteria, and 59 articles were included for full-text evaluation. Of these articles, only those that used a combination of questionnaires and physiological methods were included. Furthermore, only those articles that clearly attempted to measure presence were selected. Studies that were judged to be clinically oriented and made use of specific subject samples (e.g., people with specific health conditions) were also excluded. Finally, 18 articles were included in the review. The scientist that worked on the selection of the article did not have any conflicts of interest in evaluating the inclusions. However, a degree of subjectivity in choosing the inclusion/exclusion criteria as well as selecting the articles fitting those criteria may have affected the selection of the articles included in the final phase of the review (e.g., the ability of the responsible author, S.G., to understand the methodologies in sufficient detail and evaluate their pertinence to the topic of the present review). This represents a limitation to the systematicity of the present review.

Table 1 provides a summary of the papers evaluated, including authors and year, measure(s) used for assessing presence, and the main findings reported. Furthermore, only papers that reported at least one physiological measure and one questionnaire

instrument for assessing presence were chosen. This design allowed the assessment of the relationship between self-report questionnaires and physiological measures. Articles that only used physiological measures and not questionnaires to assess presence were scarce in the literature (e.g., Bosse et al., 2016). They lacked the subjective self-report component and thus failed to clearly state their criteria for assessing the sense of presence. Finally, twenty articles were included in the main result body of the present literature review and are therefore presented and summarized in **Table 1**. The first author of the present review (S.G.) had the responsibility for selecting the articles that were included in each stage of selection. However, both authors participated in the discussion for identifying exclusion and inclusion criteria. A PRISMA flow diagram is presented to facilitate understanding of the review procedure (**Figure 1**).

RESULTS

The study results are presented in the following paragraphs. They are firstly analyzed regarding the questionnaires used, and in the following section of the review, regarding the physiological measures. A critical overview of the measures and the main findings of the analyzed articles is also presented.

Questionnaires

The most utilized method for evaluating presence is by self-reporting questionnaire (Hein et al., 2018). According to the articles selected for the current review, the most commonly used questionnaires were the PQ, SUS, ITQ, and MEC Spatial Presence Questionnaire (MEC-SPQ). The PQ was by far the most frequently used self-report measure in the analyzed literature, a result that confirms the findings of the literature review by Hein et al. (2018). Schubert et al. (2001) already noted a few years after its introduction that the PQ questionnaire was widely used in a variety of different fields, including studies on general presence, social presence, simulations, storytelling, and games. The SUS Questionnaire was the second most used questionnaire in the body of literature analyzed.

Other studies used less conventional questionnaires, such as the Modified Reality Judgment and Presence Questionnaire (MRJPQ; Anderson et al., 2017), IPQ (Schubert et al., 2001), University College of London Questionnaire (UCLQ; Usoh et al., 1999; Meehan, 2001), and Multimodal Presence Scale (MPS; Makransky et al., 2017a). Several studies preferred simpler scales (e.g., Hoffman et al., 2003), constituted by a single item that queried the general sense of presence. Some studies used a combination of questionnaires, and the results from different measures of presence correlated, as was the case for the PQ and SUS (Salanitri et al., 2016; Skarbez et al., 2017). One of the studies analyzed (Lee et al., 2017) employed a specific questionnaire (in addition to the PQ) that aimed to evaluate social presence (Social Presence Questionnaire, Bailenson et al., 2003). However, the results of the two questionnaires were not directly compared.

There are many other questionnaires that have had some resonance in the literature but were not presented in the current literature review because they are not associated with

TABLE 1 | Summary of reviewed articles, listed from the oldest to the newest (and alphabetically for publications from the same year).

Study (Author, Year)	Journal	N	Presence measure used	Details/Task	Main findings
Meehan et al. (2002)*	ACM Transaction on Graphics	Multiple experiments, from 92 to 132	Questionnaire (UCLQ) Physiological (HR, SC, ST)	Comparison of participants' physiological reactions when in a state of stressful virtual height and a non-threatening virtual room	Presence correlated with change in HR and, to a lesser extent, SC. There were no changes in ST.
Hoffman et al. (2003)*	CyberPsychology & Behavior	7	Questionnaire (scale from 0 to 10) Physiological (fMRI)	VR game that induced low or high presence (visual immersion)	Explored the potential of using fMRI to evaluate the sense of presence. Detailed results were not reported.
Baumgartner et al. (2006)	CyberPsychology & Behavior	23	Questionnaire (MEC-SPQ) Physiological (EEG, SC)	Participants asked to attend a simulation of a roller coaster	Increase in SC responses in the experimental vs. control condition. Event-related power decrease in alpha waves over the brain parietal cortex (cortical activation in the region). Signal localization analyses showed that spatial cues presented during the simulation elicited activity over the parietal cortex and the insula.
Baumgartner et al. (2008)*	Frontiers in Human Neuroscience	77	Questionnaire (MEC-SPQ) Physiological (fMRI)	Subjects presented with a simulation of a first-person roller coaster in different scenarios	Involvement of the right brain and, to a lesser extent, of the left dorsolateral prefrontal cortex (negative correlation with presence) as a neural correlate for presence in adults. Absence of such a mechanism in children.
Busscher et al. (2011)*	Journal of CyberTherapy and Rehabilitation	60 (exp 1) and 44 (exp 2)	Questionnaire (IPQ) Physiological (HR)	Presentation of a neutral virtual world and a virtual flight	HR decreased during the highly immersive VR presentation compared to both neutral presentation and real life.
Kim et al. (2012)*	2012 IEEE Virtual Reality Workshops (VRW)	53	Questionnaire (PQ) Physiological (SC)	Examined the effects of different kinds of VE technologies on human emotions and performance	The most immersive methodologies (CAVE and HMD) produced a higher sense of presence in the users vs. desktop display and increased SC. The CAVE system showed the highest modulation of SC.
Kober et al. (2012)	International Journal of Psychophysiology	30	Questionnaire (PQ, SUS, ITQ) Physiological (EEG)	A spatial navigation task performed using two different presentations: a highly immersive single-wall display (three-dimensional view) and a desktop display (two-dimensional view)	The more immersive media provoke a more intense sense of presence. Task-related power decrease in the alpha band (8–12 Hz) over the parietal cortex correlated with a stronger feeling of presence. A lower reported sense of presence correlated with enhanced brain connectivity between frontal and parietal brain areas.
Kober and Neuper (2012)	International Journal of Human-Computer Studies	52	Questionnaire (SUS, PQ, Short Feedback Questionnaire [SFQ]; Kizony et al., 2006) Physiological (EEG)	Navigation in a virtual city (first-person view)	Auditory event-related potentials (ERPs) elicited by VR-irrelevant tones (oddball paradigm) were indexes of presence experience. The reported increase in the sense of presence correlated with the allocation of attentional resources toward the simulated environment (as opposed to the real environment). Late negative slow waves were the best indicators for presence experience in VR, but not mismatch negativity (MMN) or earlier ERP components.

(Continued)

TABLE 1 | Continued

Study (Author, Year)	Journal	N	Presence measure used	Details/Task	Main findings
Poels et al. (2012)	Cyberpsychology, Behavior, and Social Networking	19	Questionnaire (PQ) Physiological (SC, EMG)	Investigated the emotion felt by gamers during a videogame-playing session	The measured level of arousal and subjective pleasure were good predictors of the user's sense of presence. SC and electromyography were used as an estimator for arousal.
Clemente et al. (2013a)	Interacting with Computers	14	Questionnaire (SUS) Physiological (fMRI)	Navigation task in a VE consisting of a clean bedroom with basic furniture and accessories	Several brain regions increase their activity during navigation in the VEs (frontal, parietal, and occipital). Activity in the dorsolateral prefrontal cortex was reduced during the navigation task. The centro-parietal cortex and insula were positively modulated by the navigation task; increased activation was related to the sense of presence.
Clemente et al. (2013b)	Studies in Health Technology Informatics	10	Questionnaire (SUS) Physiological (EEG)	Navigation task in a VE consisting of a virtual bedroom containing some standard items (bed, closet, desk, etc.)	Insula activation was correlated with the sense of presence. This brain area may be related to stimulus attentional allocation and self-awareness processes. Portable EMOTIV EPOC EEG was used.
Burns and Fairclough (2015)*	International Journal of Human-Computer Studies	20	Questionnaire (IEQ) Physiological (EEG)	Playing a computer game via two different visualization methods; modulation of sense of presence using different stimulus-presentation devices	The technology of visualization of the video game utilized did not affect the EEG signals recorded (ERPs). They used a dual-oddball task as in Kober and Neuper (2012). As in the latter study, the late ERP component was modulated by task involvement.
Anderson et al. (2017)*	Aerospace Medicine and Human Performance	18	Questionnaire (Modified Reality Judgment and Presence Questionnaire [MRJPQ]) Physiological (SC, HR)	Participants were asked to passively view a VE	Reductions in SC from baseline were greater at the end of the natural scenes compared to the control scenes. HR results were inconclusive.
Lee et al. (2017)*	IEEE 2017: Virtual Reality	41	Questionnaire (PQ, Social Presence Questionnaire; Bailenson et al., 2003) Physiological (SC)	Participants were asked to rate their sense of presence when experiencing various degrees of multi-sensory stimulation: "mute," "vibration," and "sound."	Increase specifically in the sense of "social presence" in the "vibration" condition compared to the "sound" condition. Increase in general sense of presence both in "vibration" and "sound" conditions compared to a mute presentation. SC was not different among conditions.
Makransky et al. (2017b)*	Learning and Instruction	52 (two groups of 28 and 24)	Questionnaire (rating 1 to 5) Physiological (EEG)	A teaching scene (science simulation) was presented using a traditional desktop display (PC) or a VR headset	The VR condition increased the sense of presence, but the learning effect was lower compared to the display presentation. Cognitive load (EEG activity based on the calculation reported by Berka et al., 2007) was higher in the VR condition.
Gromer et al. (2019)*	Frontiers in Psychology	49	Questionnaire (MEC-SPQ) Physiological (SC, HR)	Immersive simulation of height; participants were selected for being sensitive to height but not clinically phobic	Experiencing emotions in a VE gave a stronger feeling of being there. The emotional factors were detected by physiological measures.

(Continued)

TABLE 1 | Continued

Study (Author, Year)	Journal	N	Presence measure used	Details/Task	Main findings
Pallavicini et al. (2019)*	Simulation & Gaming	24	Questionnaire (SUS) Physiological (SC, HR)	Participants played a game (first-person shooter) using different visualization technologies to induce a lower or higher degree of immersion (desktop monitor vs. Oculus Gear VR)	No difference in "game difficulty" between desktop video and VR. The study participants showed enhanced emotional responses in VR. The reported sense of presence was also higher in the VR condition.
Terkildsen and Makransky (2019)	International Journal of Human-Computer Studies	34	Questionnaire (MPS) Physiological (SC, EEG)	Participants explored and interacted in the scenario of a game	In MMN, the N1 components were presence indicators. Skin response peaks/minimums were also found to be presence indicators.

The names of the questionnaires, as well as the names of the physiological measures utilized, are abbreviated; please consult the Appendix. Articles indicated with * made use of head-mounted displays or similar devices.
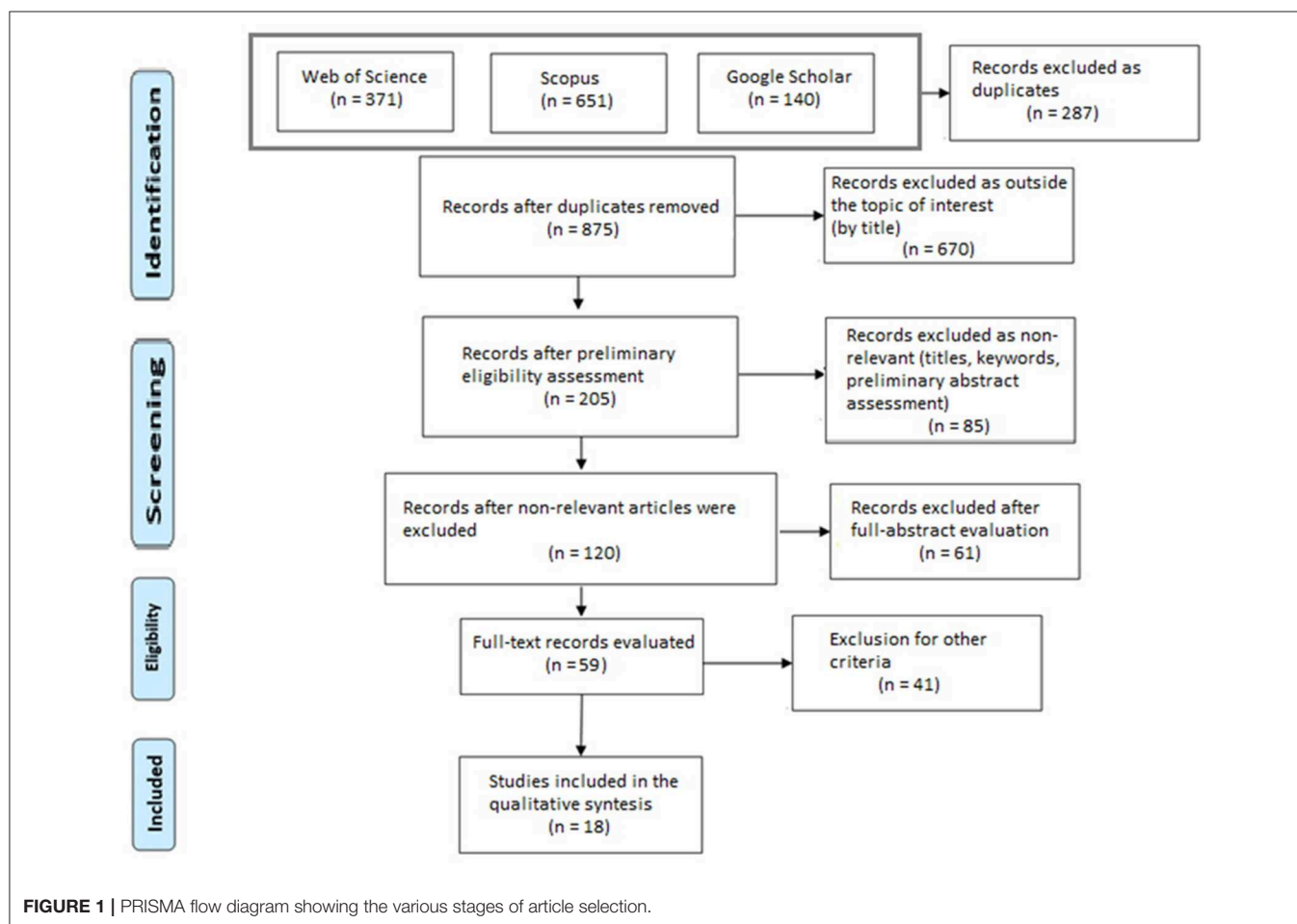
FIGURE 1 | PRISMA flow diagram showing the various stages of article selection.

physiological measures in the published research body. For example, the E^2I Questionnaire has been adopted in studies that emphasized the role that enjoyment has in presence (Lin et al., 2002). Following the same rationale, Frommel et al. (2017) explored enjoyment as a vital part of the VE, including enjoyment items in a modified version of the PQ questionnaire. Similarly,

there are other studies that considered more specific measures of what is referred to as "spatial presence" (MEC-SPQ; Vorderer et al., 2004). The sense of "spatial presence" was specifically defined as the feeling of being present within the physical body, as well as feeling able to interact in the VE (Lee, 2004). Sohre et al. (2017) also used a customized questionnaire to evaluate attention

(and, specifically, spatial presence). Witmer and Singer (1998) published the Immersive Tendencies Questionnaire (ITQ), which is a modification of the PQ they had designed earlier. The researchers used this questionnaire in combination with the PQ in several studies (Witmer and Singer, 1998; Skarbez et al., 2017).

Overall, the published research evidenced the use of a wide variety of questionnaires in association with physiological measures. Furthermore, when their items are more closely analyzed, these questionnaires were found to differ significantly in terms of scope and details. In principle, the questionnaires aim to quantify the same construct (presence), but intrinsic differences exist due to the designer's conceptualizations of presence, as noted by Slater (1999) in his critique of Witmer and Singers' PQ. Of the questionnaires presented in the reviewed literature, the majority aims to explore the concept of "physical presence" (PQ, SUS, IPQ, and MRJQ); however, the way physical presence is described varies among the instruments. The PQ questionnaire emphasizes the "involvement" and "immersion" characteristics of the simulated environment, while the SUS and IPQ are focused on the sense of "being there" (i.e., the sense that the experienced VE may be part of the reality). The MEC-SPQ questionnaire analyzes what is called "spatial presence," specifically in the framework of the MEC Two-Level Model of Spatial Presence (Vorderer et al., 2003). The latter model theorizes a clear separation between the constructs of presence, involvement, and attention, and factors such as involvement, self-location, and the possibility of performing real action in the VE. Other questionnaires are less clear about the theoretical model they follow. For instance, the Immersive Experience Questionnaire (IEQ) does not directly discuss a model in its theoretical formulation but elaborates more generally about theories of cognitive involvement, absorption, and flow (Jennett et al., 2008). In contrast, the MPS questionnaire clearly states its theoretical starting point as Lee's (2004) theory, as well as the types of presence assessed in its inventory (physical, social, and self-presence). A detailed overview of the questionnaires reviewed in the present review is given in **Table 2**.

With the proliferation of technologies that aim to simulate interactions between people and environments, there is a need to evaluate presence metrics to build measurements that can consistently predict the sense of presence of the users (based on a unified theory). Thus, developing questionnaires to measure presence properly is still an ongoing scientific endeavor. Factor analysis was used in the analyzed articles to test the construct validity of the MPS, IPQ, and IEQ specifically reported in that article. The predictive validity of the questionnaires is sometimes stated in the original article. As reported in **Table 2**, this factor was most commonly assessed by measuring the positive correlation with various types of task performance (PQ and SUS). However, the later work of Youngblut and Huie (2003) found inconclusive evidence for the ability of those questionnaires to predict task performance in VEs.

Items for many of the analyzed questionnaires were sometimes unclear or difficult to grasp for a naïve participant, e.g., "How is your memory of the scenario similar to being in real places?" (SUS; Slater and Steed, 2000) or "How compelling was your sense of objects moving throughout space?" (PQ; Witmer

and Singer, 1998). A recent effort to build a unitary scale for the measurement of various aspects of presence was made with the development of the MPS (Makransky et al., 2017a). This effort represents a step toward a standard and widely accepted measure of presence for various types of VEs. The scale attempts to assess the three major senses of presence in virtual environments (Lee, 2004): physical presence (like several previously reported scales, including PQ and SUS), social presence, and self-presence. Another major advantage of the MPS scale is its clear theoretical standpoint (Lee, 2004), which has been validated by multiple experiments according to tests performed using modern test theories (i.e., confirmatory factor analysis, as described in Brown, 2014) and the Item Response Theory (as described in Embretson and Reise, 2013).

It is worth noting that the current review may not adequately represent the entire literature published in the last two decades, as several factors may have influenced the selection criteria of the sample of articles included. Therefore, the prominence of the measures of presence that were distinguished may not be generalized outside the sample of articles analyzed here. However, our findings (e.g., the PQ as the most-used questionnaire) are coherent with previously published reviews (see Hein et al., 2018). **Figure 2** presents an overview of the prevalence of questionnaires used in association with physiological measures.

Physiological Measures

Psychophysiological studies use a variety of instruments to evaluate the relationship between psychological and physical states in both experimental and naturalistic settings. In tandem with modern brain image methods, psychophysiological approaches have enhanced the understanding of the interplay between physiological indices and several human cognitive processes.

Numerous studies aimed to find physiological measures that correlate with the self-reported sense of presence, using a variety of methodologies. The published research shows a variety of physiological measures that correlate with subjective self-report of presence. EEG was commonly analyzed using both spectral signal decomposition and ERPs. The ERP paradigms commonly assessed presence indirectly, specifically by studying the attentional processes allocated toward the simulated vs. the "external" real environment. In order to do so, the researchers used a traditional auditory dual-task oddball paradigm. In this paradigm, a series of frequent, repeated auditory tones are followed by less frequent tones of a different pitch, a phenomenon that generates a cognitive mismatch in the listener. This experimental design is often used in the study of subjective awareness, in combination with EEG (Näätänen et al., 2019). Burns and Fairclough (2015) used the Kober and Neuper (2012) paradigm in their study and found that task difficulty significantly increases immersion. They also replicated the connection between slow-wave amplitude in response to oddball auditory stimuli and sense of presence. However, the recent work of Terkildsen and Makransky (2019), who employed a very similar study design to both Kober and Neuper (2012) and Burns and Fairclough (2015), did not replicate their findings. Instead,

TABLE 2 | Summary of the characteristics of the most utilized questionnaires in the analyzed articles.

Name	Number of items	Explored concept	Theory/formulation	Subscales	Factors/Indicators/Dimensions	Reliability	Construct validity	Predictive validity
PQ	19	Physical presence	The sense of involvement and immersion are different, and both are necessary for developing the sense of presence. Involvement: focusing attention and energy on VE. Immersion: self-perceiving as a part of the VE stimulation.	Involvement/control, Natural, Auditory, Haptic, Resolution, Interface quality	Control, Sensory, Distractor, Realism	Cronbach's α = 0.88	Items based on the factors derived from a review of the literature. PQ items investigate presence as involvement and immersion. Positively related to ITQ (r = 0.24).	Positively related to task performance. Negatively related to simulator sickness. Validity was inconclusive when testing for performance (Youngblut and Huie, 2003).
SUS	6	Physical presence	Presence is treated as a mental state. Users are physically present in the VEs (from the formulation of Draper et al., 1998).	None	Being there, VE as more real or present than reality, Locality (VE as a visited place)	Not reported	Not reported by the author of the questionnaire. Positively related to PQ (r = 0.51).	Negatively related to the number of errors in the VE task. Validity was inconclusive when testing for performance (Youngblut and Huie, 2003).
MEC-SPQ	3 versions: 4, 6, or 8 items per subscale	Spatial presence	The MEC Two-Level Model of Spatial Presence (Vorderer et al., 2003); distinction among presence, involvement, and attention.	Attention Allocation, Spatial Situation Model, Spatial Presence: Self Location, Spatial Presence: Possible Actions, Suspension of Disbelief, Higher Cognitive Involvement, Domain Specific Interest, and Visual-Spatial Imagery	Spatial Presence: Self Location, Spatial Presence: Possible Actions, Cognitive Involvement	Assessed separately for each scale, Cronbach's α from 0.78 to 0.94.	Not reported	Not reported by the author. Positively correlated with user performance (Morrison et al., 2009)
IPQ	14 (using items from the SUS and PQ)	Physical presence	The sense of presence emerges from the creation of a spatial-functional mental model of the VE. Cognitive processes related to this model are (1) construction of the representation of body action as real possible actions within the VE and (2) suppression of external sensory inputs that are incompatible with the VE.	Spatial Presence, Involvement, Realness	Transportation, presence as immersion in the VE; presence as the "realism" of the VE	Cronbach's α = 0.85 and 0.87 (from two preliminary studies)	Data collected using this questionnaire showed a similar factor structure compared to other questionnaires. Confirmatory factor analysis of the model that consists of three factors confirmed the validity of the measure (Schubert et al., 2001).	Ability to distinguish multiple levels of presence (van Schaik et al., 2004; Riecke and Schulte-Pelkum, 2015)

(Continued)

51

TABLE 2 | Continued

Name	Number of items	Explored concept	Theory/formulation	Subscales	Factors/Indicators/ Dimensions	Reliability	Construct validity	Predictive validity
ITQ	18	Involvement	The items assess the tendency to become involved in activities and the ability to focus on one specific activity. Investigates both involvement and immersion.	Not reported	Focus, involvement, tendency to play games	Cronbach's α = 0.75	Based on the items from the analysis of previous literature. Correlates with PQ.	Not reported
IEQ	31	Mixed: physical presence and theories of cognitive involvement	Built on theories of flow, cognitive absorption, and presence. Five dimensions of cognitive absorption (Agarwal and Karahanna, 2000). Operationalized based on available items from the literature.	Basic attention, temporal dissociation, transportation, challenge, emotional involvement, enjoyment	Cognitive Involvement, Real World Dissociation, Challenge, Emotional Involvement, Control	Not reported	Factor analysis (principal component analysis)	Not reported
MRJPQ	15	Physical presence	Items build to evaluate the extent to which the user feels physically in the virtual world and an emotional impact from the simulated scenario	Not reported for the modified version	Not reported for the modified version	Cronbach's α = 0.88	Not reported for the modified version.	Not reported for the modified version.
MPS	15	Mixed: physical, social, and self-presence	Lee's (2004) theory of presence, division of presence into three sub-dimensions: physical, social, and self.	Physical presence, Social Presence, Self-Presence	Physical realism, not paying attention to the real environment, sense of being in the VE, not aware of the physical mediation, sense of coexistence, human realism, not aware of the artificiality of social interaction, not aware of the social mediation, sense of bodily connectivity, the sense of bodily extension	Good reliability measured in different experiments for each of the subscales in different studies (using the Person separation index and Cronbach's alpha)	Based on the items from the analysis of previous literature. The scale was validated using confirmatory factor analysis and item response theory.	Not reported by the authors. Physiological index of attentional allocation (Terkildsen and Makransky, 2019).

*This overview is based on the original publication of the scales and published literature. Please see the **Appendix** for full questionnaire names.*
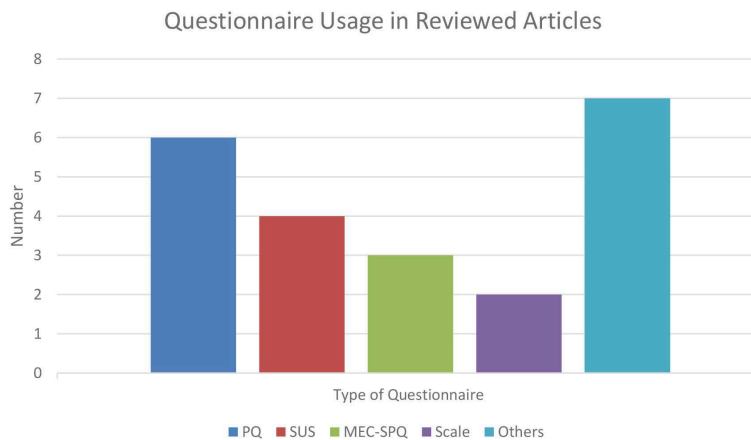
FIGURE 2 | Usage of presence questionnaires in the reviewed articles. See the **Appendix** for full questionnaire names.

these authors reported an earlier ERP component correlated with presence (N1 and MMN). Clemente et al. (2014) used EEG to measure presence for navigation in virtual environments, using consumer-grade EEG equipment (Emotive EPOC EEG).

Only a limited number of fMRI studies (three) were found in the literature (one of which did not report detailed findings; Hoffman et al., 2003), reporting only partially coherent results.

Among the non-brain physiological indices, presence was found to positively correlate with SC activity (Baumgartner et al., 2006; Kim et al., 2012), but Meehan et al. (2002) reported only a weak association and suggested that HR is a better correlate for presence. Furthermore, Terkildsen and Makransky (2019) found that the sense of presence may differently modulate different SC components: SC peaks/minima are good predictors of the sense of presence, while SC response magnitude shows no association with it. Another study (Lee et al., 2017) found no association between presence and SC (using the sum of skin responses for each of the three conditions employed in their experiment).

HR was found to be one of the most reliable indices of presence among the physiological indices, and it can be compared directly with other physiological measures (Meehan et al., 2002), except in the case of the study of Busscher et al. (2011), which found negative correlation between HR and presence in a simulated flight scenario. The authors of the latter article proposed that the average lower heart rate during the virtual flight (compared to a neutral VE condition), could reflect participant coping mechanisms. A similar reduction in HR has been observed in other studies presenting stressful stimuli (Bosch et al., 2001; Busscher et al., 2010). Furthermore, analysis of HR variability (Anderson et al., 2017) was shown not to be a reliable index of the sense of presence. Some studies analyzed in this review (e.g., Pallavicini et al., 2019) explicitly did not use HR to study presence *per se*, but rather, they used it to index connected factors, such as emotional responses.

One study investigated ST as a possible index of presence (Meehan et al., 2002); even though the authors expected skin temperature to be an index for enhanced presence in their more highly arousing experimental condition (according to previous investigations using sense of height as exposure; see Andreassi,

2010), there was no connection between the phenomena. Unfortunately, the use of ST as a possible index of presence was not found in other articles included in the present review. More studies are needed to gain a better understanding of ST as an index for presence. One study also combined EMG with SC to assess the sense of presence (Poels et al., 2012). The results of the EMG confirmed those from SC. However, in the context of the study of Poels et al. (2012), EMG was used as a measure of arousal and not of presence. Unfortunately, such a finding was not replicated nor attempted by other studies, and therefore it is challenging to give an interpretation to it. For an overview of the number of uses of physiological measures in the reviewed articles, see **Figure 3**. The prevalence of the psychophysiological indices reported in the present review may be characteristic of the sample of studies retrieved and may be difficult to generalize when also considering studies that have been excluded from the current review.

DISCUSSION

In the present review, articles from the last two decades that examined both questionnaires and physiological measures for the assessment of the sense of presence in VEs were analyzed. The PQ was the most often-used instrument. This review showed that a wide range of questionnaires had been developed and utilized for assessing presence. Furthermore, it was noted that those questionnaires, even if they aimed to measure the same constructs, are rooted in several different theorizations of presence. Indeed, they have a very different number of items and use a variety of distinct sub-scales. These scales, however, seem to generally correlate with one another (e.g., PQ with SUS and PQ with ITQ). Construct validity of the scales was generally assessed by reviewing the already existing literature; sometimes, factor analysis (IEQ or MPS) was utilized to assess it. Human performance was used directly to validate the effectiveness of the most-used scales (but see Youngblut and Huie, 2003).

A wide range of physiological measures was used together with questionnaires, attempting to identify physiological correlates of the sense of presence. The most prevalent measure

FIGURE 3 | Physiological measures in the reviewed studies. SC, skin conductance; EEG, electroencephalogram; HR, heart rate; fMRI, functional magnetic resonance imaging; ST, skin temperature; EMG, electromyography. As noted from the chart, EEG is the most commonly used physiological measure, followed by SC.
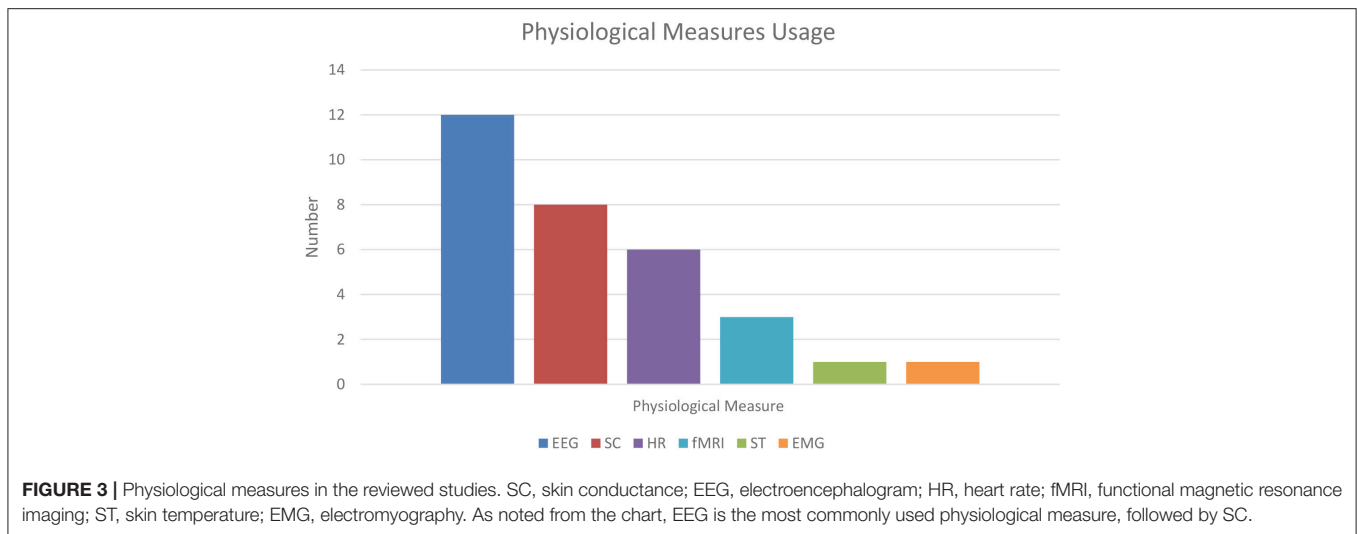
in the retrieved literature was EEG. The attempts to use SC and HR or to exploit other cognitive phenomena during EEG recordings to assess the sense of presence have revealed mixed results. Even though many studies explored the sense of presence, many criticalities were identified in the present literature review. The study of presence is made complex by the degree of subjectivity of the construct, the different modalities of inducing presence (and their evolution over the years), and the difficulties related to the communication of the feeling of presence (Slater, 2004; Sanchez-Vives and Slater, 2005; Oh and Rosakranse, 2014; Cummings and Bailenson, 2016).

Presence is the sensation of being in the place presented in a VE (Nash et al., 2000). It is characterized by the illusion that the virtual events are real, and it is fundamental to determine the extent of presence in those simulated environments (Slater et al., 1994). The "illusion" or "feeling" of presence is highly subjective, and there is no developed and generally accepted method for its evaluation. As shown in the present review, researchers have used markedly different means to measure the phenomenon, such as questionnaires featuring widely different items and a variety of physiological responses. Furthermore, findings from the physiological responses were obtained using a variety of methodologies and diverse phenomena within the same methodology (e.g., brain wave oscillation analysis vs. event-related potentials; peak amplitude analysis vs. overall skin conductance).

Measuring presence using questionnaires is justified by the fact that it is a subjective concept (Nichols et al., 2000). As such, questionnaires are commonly considered the most appropriate measure for a person's subjective experience. Previous studies demonstrated that experiments for evaluating presence using questionnaires employ manipulative aspects of VE. In most cases, the responses are based on Likert scales (Nichols et al., 2000; Lessiter et al., 2001; Slater et al., 2010). As noted in the study by Lessiter et al. (2001), a variety of questionnaires are utilized, and they are often based on different definitions of presence. The present review attempted to clarify and summarize the characteristics of the most used questionnaires (see **Table 2**). As

already evidenced in the Results section, those questionnaires utilized many different theorizations of the construct of presence, for example, the theory of multidimensionality of presence of Lee (2004) employed by the MPS, and the theories of cognitive flow (see, e.g., Agarwal and Karahanna, 2000) used by the MRJPQ. This has created an astonishing number of measures for presence; see the criticism of the proliferation of questionnaires assessing presence by Nordin et al. (2014) and the comprehensive (but outdated) list of presence questionnaires presented by Baren and Ijsselsteijn (2004).

There are several benefits with psychophysiological measures, in case reliable ways to measure the sense of presence with only physiological measures are identified. These benefits include the possibility of continuous data collection and, therefore, the ability to study the association of physiological activity with the effect of a contemporary stimulus presentation in experimental paradigms. Psychophysiology can also help in the investigation of phenomena that are difficult to capture using self-report methods. Thus, this modality will provide a more accurate analysis of the physiological state of a person compared to questionnaires or bare behavior.

Consumer-oriented, cost-effective, small, and easy to handle EEG systems are now available and have already been used for the study of the sense of presence (Clemente et al., 2013b; Tarrant et al., 2018). However, even though validation against medical- or research-grade equipment has been undertaken (see, e.g., Badcock et al., 2013), these consumer-oriented EEGs have been shown to record a less reliable signal compared to their—generally more expensive—professional counterparts (see Duvinage et al., 2012, 2013). Therefore, results from studies that utilize these tools may generally be more inaccurate and difficult to replicate. Furthermore, EEG (with regards to the study of presence) is often used by employing very different experimental paradigms and methodologies that are impossible to compare directly. Additionally, as in the case of the recent study of Terkildsen and Makransky (2019), some published results are difficult to replicate, even when employing almost identical paradigms. However, it is worth noting that Terkildsen and

Makransky (2019) used the MPQ to assess subjectively reported presence, a measure that shows essential differences compared to the questionnaire used in previous studies that employed the same paradigm (SUS and PQ in Kober and Neuper, 2012; IEQ in Burns and Fairclough, 2015). On the positive side, Tauscher et al. (2019) studied the methodological feasibility and analyzed the possibility of using EEG equipment combined with VR headsets. They validated the possibility of combining these methods and reported a relatively small and predictable interference (easy to clean) from the VR headset with the recorded signal. As Millán et al. (2010) noted, EEG is widely implemented in the development of brain-computer interface (BCI), and therefore it is expected that in the next decades the use of this methodology will grow exponentially in the tech sector and possibly in the study of brain activity during immersive experiences.

The few fMRI studies reported in the literature appear to generally agree on the involvement (negative correlation) of the dorsolateral prefrontal cortex in the experience of the sense of presence and on the role of the insula. However, Clemente et al. (2013a) reported the inactivation of the dorsolateral prefrontal cortex in a different, inferior location compared to Baumgartner et al. (2008). Related brain areas were also shown to activate in EEG studies (Clemente et al., 2013b). However, the source localization attempt carried out in the EEG study of Clemente et al. (2013b) was conducted with a sub-optimal configuration: the researchers used data from low-density (14-channel) consumer-oriented EEG equipment to perform signal source-localization analysis (see Staljanssens et al., 2017; Michel and Brunet, 2019) and with a small participant sample (10). Further studies are needed for a more precise spatial localization of the phenomenon and to confirm it. The activation of the insula is related to cognition and behavior, as for emotion, regulation of the body's homeostasis, perception, motor control of hands and eyes, self-awareness, cognitive functions, and interpersonal experience (Karnath et al., 2005; Craig, 2009; Clemente et al., 2014). Self-awareness, sense of agency, and sense of body ownership are essential in this context, as directly linked with the sense of presence. Attentional and behavioral components are crucial for the development of the sense of presence, such as increasing the ability to understand the dynamics, predict, and interact with the VE (Sjölie, 2012). Clemente et al. (2013a) acknowledged that their interpretation of the data was speculative and that the brain areas activated during their VE navigation task are possibly not directly related to the sense of presence *per se*.

The reviewed literature showed that little effort had been made to replicate the physiological indices proposed for the feeling of presence, despite the ambiguous results sometimes reported and the variety of different methods used for indexing the phenomenon. When replication was deliberately attempted, the early results were not always successful (Terkildsen and Makransky, 2019). SC and HR studies generally showed more consistent findings, with a higher level of immersion correlating with greater SC or increased HR. However, Busscher et al. (2011) reported a negative correlation between HR and presence and proposed this to be a result of the "conservation-withdrawal" (see Bosch et al., 2003) coping response to a more arousing VE experience. If this interpretation is correct, HR may be very sensitive to the content of the simulation, and its use for the study of presence may be impaired.

While fMRI and EEG are more likely, in the future, to pinpoint specific brain-related activity patterns or areas directly involved (or highly correlated) with the sense of presence (as in the somewhat related field of awareness; see Koivisto and Grassini, 2016; Koivisto et al., 2016, 2017; Jimenez et al., 2018), SC and HR are more likely to identify secondary effects of presence, as well as experiences modulated by or together with presence, such as, for example, arousal, emotion, and stress (Poels et al., 2012; Chalfoun and Dankoff, 2018). Furthermore, there are myriad data collection and data analysis modalities for SC (SCR, SCL, and GSR peak amplitude/number, just to cite a few), and this makes the comparison of the various methodologies that employ SC almost impossible.

Methodological problems in the study of presence include over-reliance on the use of the self-evaluating questionnaires and the difficulty in understanding and elaborating the questionnaire items for the naïve participants involved in the experiments. Additionally, questionnaire methods only allow presence in VEs to be evaluated in a non-invasive manner and only after the experience, while attempting to do it during the experience comes at the cost of breaking the participant's sense of presence. Thus, an approach to indirectly measure presence via, for instance, non-invasive and reliable physiological measurements, would be ideal for understanding the ongoing presence level of the users without the need to query them about their subjective feelings and interrupting their immersive experience to collect data.

Physiological measures may provide numerous advantages compared to other types of measures in the future, for example, continuous, non-invasive, real-time, and relatively objective assessment (Kavikanga et al., 2011). Unfortunately, these measures are difficult to use alone in the current state of research due to the lack of verified and easily replicable physiological indices of presence that are highly correlated with self-reports. Furthermore, several problems are present in the use of physiological measures compared to questionnaire instruments to assess presence. Recording equipment is very sensitive to motion (EEG, SC, and HR), and, in fMRI, motion by the subject is not allowed at all due to technical constraints. Therefore, not all experimental scenarios will be suitable for those methods.

Moreover, VE content can affect the recorded data and thus constitute an uncontrollable variable when comparing among different experimental settings. It may be challenging to isolate the phenomena of "presence" *per se*, as many other cognitive or perceptive factors (e.g., emotional charge of the environment, arousal of the subject, and image quality) may profoundly influence the physiological data. Only systematic replicative efforts of published studies, as well as future investigations that aim to isolate the phenomenon of presence and reduce the influence of confounding factors, can help to identify reliable physiological correlates of presence. An optimal physiological correlate of presence should be able to discriminate the level of the user's presence with a high correlation with self-reported questionnaires and independently from the content of the VE experience.

Immersive media are currently used as a popular source of entertainment (Williams and Mascioni, 2017). It is, therefore, important to comprehensively evaluate the user experience, and the sense of presence is often used as an index of quality for virtual environments. Additionally, understanding and measuring presence has become necessary due to the various applications of virtual reality in different fields, as noted by Schuemie et al. (2001). Furthermore, several studies demonstrated that presence could be a crucial factor to consider when using VE outside the entertainment context, for example, as a training tool in work environments or to increase the performance of users (Baumgartner et al., 2006; Baus and Bouchard, 2017).

CONCLUSIONS

The present review analyzed the body of scientific literature on the measurement of the sense of presence published during the last two decades. The analyzed studies were selected from those reporting both physiological and questionnaire data. In the introduction of the present review, emphasis was placed on clarifying the most common definitions and propositions for presence and related concepts. Furthermore, the review replicated and updated the results of the comprehensive review on questionnaires used for evaluating the sense of presence published by Hein et al. (2018), as well as previous work (Insko, 2003) that examined the physiological correlates of the sense of presence.

Overall, there was no standard measurement method for presence, even though there is a growing body of literature that compares various measurement constructs. One of the most important findings reported in this review is the reliance on the use of questionnaires (and the diversity of questionnaires). Considering that in the coming years, VR technology users will probably increase in number, there is a need for research on standard practices and standardization in the area to understand the effect of those technologies on the end-user, as well as to help in the development of better ones in the future.

At the current state of research, no physiological measure has collected enough evidence to be considered "good enough" to be reliably used alone, without the user giving their subjective evaluation of the experience. The nowadays quite outdated study of Meehan et al. (2002) compared several non-brain physiological measures and showed that HR may be a better measure than SC and ST. Other studies have also shown the generally acceptable reliability of HR for evaluating presence. However, the measure was found not to be reliable when HR variability analysis was attempted (Anderson et al., 2017) and was shown to be possibly too sensitive to the content of the VE to be successfully used in emotional/arousing VE scenarios (Busscher

et al., 2011). A combination of different types of measures (e.g., questionnaire, behavioral, and physiological), may be, in the current state, the better approach for properly evaluating the sense of presence in a VE.

Physiological measures, especially SC and EEG, showed widespread use. However, despite contrasting results in the literature, little effort has been invested in systematically replicating already published studies. In this regard, when talking about new technology, it is essential to acknowledge that technical progress evolves very quickly, and several of the reviewed studies were conducted more than a decade ago. Thus, their findings may have still-relevant but limited implications for recent technology (modern high-definition, more immersive VEs). As such, more studies should be performed based on current technology and devices. Furthermore, some studies should aim to replicate old findings using new devices. This line of research can help with the development of a more comprehensive theoretical framework for presence and related constructs.

Future research should also focus on the role of presence in shaping human performance. The role of presence in new media and its relationship with user performance may be of great interest for applied research in the field of IVTs. More empirical research is needed to understand better how the sense of presence (and related factors, e.g., emotional involvement) may represent a positive or a negative factor for human performance, depending on different uses of IVTs. The authors of the present review wish to stimulate the replication of the published scientific research on the physiological correlates of presence, especially in those cases where the published studies have reported conflicting or inconclusive results. At the same time, this review underlines the theoretical problems with the definition of presence as a psychological construct.

AUTHOR CONTRIBUTIONS

SG was responsible for retrieving and reviewing the current literature, writing the drafts of the article, including figures and tables, and revise all the previously submitted versions of the manuscript. KL provided funding for the project, extensively commented on the earlier drafts of the manuscript, and revised the manuscript at every stage of its development. Both authors agreed on and commented on the final version of the manuscript prior to submission.

FUNDING

REFERENCES

Agarwal, R., and Karahanna, E. (2000). Time flies when you're having fun: cognitive absorption and beliefs about information technology usage. *MIS Q.* 24, 665–694. doi: 10.2307/3250951

Amin, A., Gromala, D., Tong, X., and Shaw, C. (2016). "Immersion in cardboard VR compared to a traditional head-mounted display," in *International Conference on Virtual, Augmented and Mixed Reality* (Cham: Springer), 269–276. doi: 10.1007/978-3-319-39 907-2_25

Anderson, A. P., Mayer, M. D., Fellows, A. M., Cowan, D. R., Hegel, M. T., and Buckey, J. C. (2017). Relaxation with immersive natural scenes presented using virtual reality. *Aerosp. Med. Hum. Perform.* 88, 520–526. doi: 10.3357/AMHP.4747.2017

Andreassi, J. L. (2010). *Psychophysiology: Human Behavior and Physiological Response, 5th Edn.* New York, NY: Psychology Press. doi: 10.4324/9780203880340

Arndt, S., Brunnstrom, K., Cheng, E., Engelke, U., Möller, S., and Antons, J. (2016). Review on using physiology in quality of experience. *Soc. Imaging Sci. Technol.* 9, 1–9. doi: 10.2352/ISSN.2470-1173.2016.16.HVEI-125

Badcock, N. A., Mousikou, P., Mahajan, Y., De Lissa, P., Thie, J., and McArthur, G. (2013). Validation of the Emotiv EPOC® EEG gaming system for measuring research quality auditory ERPs. *Peer J* 1:e38. doi: 10.7717/peerj.38

Bailenson, J. N., Blascovich, J., Beall, A. C., and Loomis, J. M. (2003). Interpersonal distance in immersive virtual environments. *Pers. Soc. Psychol. Bull.* 29, 819–833. doi: 10.1177/0146167203029007002

Banos, R., Botella, C., Quero, S., Garcia-Palacious, A., and Alcaniz, M. (2008). Presence and emotions in virtual environments: the influence of stereoscopy. *Cyberpsychol. Behav.* 11, 1–8. doi: 10.1089/cpb.2007.9936

Baren, J. V., and Ijsselsteijn, W. (2004). *Measuring Presence: A Guide To Current Measurement Approaches*. Project Report. OmniPres project IST-2001-39237.

Baumgartner, T., Speck, D., Wettstein, D., Masnari, O., Beeli, G., and Jäncke, L. (2008). Feeling present in arousing virtual reality worlds: prefrontal brain regions differentially orchestrate presence experience in adults and children. *Front. Hum. Neurosci.* 2:8. doi: 10.3389/neuro.09.008.2008

Baumgartner, T., Valko, L., Esslen, M., and Jäncke, L. (2006). Neural correlate of spatial presence in an arousing and noninteractive virtual reality: an EEG and psychophysiology study. *Cyberpsychol. Behav.* 9, 30–45. doi: 10.1089/cpb.2006.9.30

Baus, O., and Bouchard, S. (2017). Exposure to an unpleasant odour increases the sense of presence in virtual reality. *Virtual Real.* 21, 59–74. doi: 10.1007/s10055-016-0299-3

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., et al. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviat. Space Environ. Med.* 78(5 Suppl.), B231–B244. Available online at: http://www.ingentaconnect.com/content/asma/asem/2007/00000078/A00105s1/art00032

Biocca, F. (1997). The cyborg's dilemma: progressive embodiment in virtual environments. *J. Comput. Mediat. Commun.* 3:JCMC324. doi: 10.1111/j.1083-6101.1997.tb00070.x

Biocca, F. (1999). Chapter 6 The cyborg's dilemma: progressive embodiment in virtual environments. *Hum. Factors Inform. Technol.* 13, 113–144. doi: 10.1016/S0923-8433(99)80011-2

Bosch, J. A., De Geus, E. J., Kelder, A., Veerman, E. C., Hoogstraten, J., and Amerongen, A. V. N. (2001). Differential effects of active versus passive coping on secretory immunity. *Psychophysiology* 38, 836–846. doi: 10.1111/1469-8986.3850836

Bosch, J. A., de Geus, E. J., Veerman, E. C., Hoogstraten, J., and Amerongen, A. V. N. (2003). Innate secretory immunity in response to laboratory stressors that evoke distinct patterns of cardiac autonomic activity. *Psychosom. Med.* 65, 245–258. doi: 10.1097/01.psy.0000058376.50240.2d

Bosse, S., Acqualagna, L., Samek, W., Porbadnigk, A., Curio, G., and Blankertz, B. (2016). Assessing perceived image quality using steady-state visual evoked potentials and spatio-spectral decomposition. *IEEE Trans. Circuits Syst. Video Technol.* 28, 1694–1706. doi: 10.1117/12.2193684

Bouchard, S., St-Jackques, J., Robillard, G., and Renaud, P. (2008). Anxiety increases the feeling of presence in virtual reality. *Presence Teleoper. Virtual Environ.* 17, 376–391. doi: 10.1162/pres.17.4.376

Boucsein, W. (2012). *Electrodermal Activity*. Boston, MA: Springer Science; Business Media. doi: 10.1007/978-1-4614-1126-0

Braithwaite, J. J., Watson, D. G., Jones, R., and Rowe, M. (2013). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology* 49, 1017–1034. doi: 10.1111/j.1469-8986.2012.01384.x

Breedlove, S. M., Watson, N. V., and Rosenzweig, M. R. (2010). *Biological Psychology*. Sunderland, MA: Sinauer Associates.

Brown, T. A. (2014). *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford Publications.

Burns, C., and Fairclough, S. (2015). Use of auditory event related potentials to measure immersion during a computer game. *Int. J. Hum. Comput. Stud.* 73, 107–144. doi: 10.1016/j.ijhcs.2014.09.002

Busscher, B., De Vliegher, D., Ling, Y., and Brinkman, W. P. (2011). Physiological measures and self-report to evaluate neutral virtual reality worlds. *J. Cyberther. Rehabil.* 4, 1–18.

Busscher, B., van Gerwen, L. J., Spinhoven, P., and de Geus, E. J. (2010). Physiological reactivity to phobic stimuli in people with fear of flying. *J. Psychosom. Res.* 69, 309–317. doi: 10.1016/j.jpsychores.2009.12.005

Chalfoun, P., and Dankoff, J. (2018). "Developing actionable biometric insights for production teams," in *Games User Research* (Oxford, UK: Oxford University Press), 281–299. doi: 10.1093/oso/9780198794844.003.0017

Cheng, M. T., She, H. C., and Annetta, L. A. (2015). Game immersion experience: its hierarchical structure and impact on game-based science learning. *J. Comput. Assist. Learn.* 31, 232–253. doi: 10.1111/jcal.12066

Clemente, M., Rey, B., Rodríguez-Pujadas, A., Barros-Loscertales, A., Baños, R. M., Botella, C., et al. (2013a). An fMRI study to analyze neural correlates of presence during virtual reality experiences. *Interact. Comput.* 26, 269–284. doi: 10.1093/iwc/iwt037

Clemente, M., Rodriguez, A., Rey, B., and Alcaniz, M. (2013b). Measuring presence during navigation in a virtual environment using EEG. *Stud. Health Technol. Inform.* 196, 136–140. doi: 10.3233/978-1-61499-282-0-136

Clemente, M., Rodríguez, A., Rey, B., and Alcañiz, M. (2014). Assessment of the influence of navigation control and screen size on the sense of presence in virtual reality using EEG. *Expert Syst. Appl.* 41, 1584–1592. doi: 10.1016/j.eswa.2013.08.055

Craig, A. D. (2009). How do you feel–now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* 10, 59–70. doi: 10.1038/nrn2555

Cummings, J. J., and Bailenson, J. N. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media Psychol.* 19, 272–309. doi: 10.1080/15213269.2015.1015740

Dawson, M. E., Schell, A. M., and Filion, D. L. (2017). "The electrodermal system," in *Cambridge Handbooks in Psychology*, eds J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson (Cambridge, UK: Handbook of psychophysiology; Cambridge University Press), 217–243.

Draper, J. V., Kaber, D. B., and Usher, J. M. (1998). Telepresence. *Hum. Factors* 40, 354–375. doi: 10.1518/001872098779591386

Durso, F. T., Geldbach, K. M., and Corballis, P. (2012). Detecting confusion using facial electromyography. *Hum. Factors* 54, 60–69. doi: 10.1177/0018720811428450

Duvinage, M., Castermans, T., Dutoit, T., Petieau, M., Hoellinger, T., Saedeleer, C. D., et al. (2012). A P300-based quantitative comparison between the Emotiv Epoc headset and a medical EEG device. *Biomed. Eng.* 765, 2012–2764. doi: 10.2316/P.2012.764-071

Duvinage, M., Castermans, T., Petieau, M., Hoellinger, T., Cheron, G., and Dutoit, T. (2013). Performance of the Emotiv Epoc headset for P300-based applications. *Biomed. Eng. Online* 12:56.

Embretson, S. E., and Reise, S. P. (2013). *Item Response Theory for Psychologists*. London, UK: Psychology Press.

Freeman, D., Reeve, S., Robinson, A., Ehlers, A., Clark, D., Spanlang, B., et al. (2017). Virtual reality in the assessment, understanding, and treatment of mental health disorders. *Psychol. Med.* 47, 2393–2400. doi: 10.1017/S003329171700040X

Fromberger, P., Meyer, S., Kempf, C., Jordan, K., and Muller, J. (2015). Virtual viewing time: the relationship between presence and sexual interest in androphilic and Gynephilic men. *PLoS ONE* 10:e0127156. doi: 10.1371/journal.pone.0127156

Frommel, J., Sonntag, S., and Weber, M. (2017). "Effects of controller-based locomotion on player experience in a virtual reality exploration game," in *Proceedings of the 12th International Conference on the Foundation of Digital Games* (New York, NY). doi: 10.1145/3102071.3102082

Goldberger, A. L. (1998). *Clinical Electrocardiography: A Simplified Approach*. New York, NY: C. V. Mosby.

Grabarczyk, P., and Pokropski, M. (2016). Perception of affordances and experience of presence in virtual reality. *AVANT* 3, 25–44. doi: 10.26913/70202016.0112.0002

Greicius, M. D., Krasnow, B., Reiss, A. L., and Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 100, 253–258. doi: 10.1073/pnas.0135058100

Gromer, D., Reinke, M., Christner, I., and Pauli, P. (2019). Causal Interactive links between presence and fear in virtual reality height exposure. *Front. Psychol.* 10:141. doi: 10.3389/fpsyg.2019.00141

Hein, D., Mai, C., and Hußmann, H. (2018). "The usage of presence measurements in research: a review," in *Proceedings of the International Society for Presence Research Annual Conference (Presence)* (Prague: The International Society for Presence Research).

Hoffman, H. G., Richards, T., Coda, B., Richards, A., and Sharar, S. R. (2003). The illusion of presence in immersive virtual reality during an fMRI brain scan. *Cyberpsychol. Behav.* 6, 127–131. doi: 10.1089/109493103321640310

Huettel, S. A., Song, A. W., and McCarthy, G. (2004). *Functional Magnetic Resonance Imaging.* Sunderland, MA: Sinauer Associates.

Ijsselsteijn, W., Ridder, H., Freeman, J., and Avons, S. (2000). Presence: concept, determinants, and measurement. *Int. Soc. Opt. Photonics* 3959, 520–530. doi: 10.1117/12.387188

Insko, B. E. (2003). "Measuring presence: subjective, behavioral and physiological methods," in *Studies in New Technologies and Practices in Communication. Being There: Concepts, Effects and Measurements of User Presence in Synthetic Environments,* eds G. Riva, F. Davide, and W. A. IJsselsteijn (Amsterdam: IOS Press), 109–119.

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., et al. (2008). Measuring and defining the experience of immersion in games. *Int. J. Hum. Comput. Stud.* 66, 641–661. doi: 10.1016/j.ijhcs.2008.04.004

Jimenez, M., Grassini, S., Montoro, P. R., Luna, D., and Koivisto, M. (2018). Neural correlates of visual awareness at stimulus low vs. high-levels of processing. *Neuropsychologia* 121, 144–152. doi: 10.1016/j.neuropsychologia.2018.11.001

Karnath, H. O., Baier, B., and Nägele, T. (2005). Awareness of the functioning of one's own limbs mediated by the insular cortex? *J. Neurosci.* 25, 7134–7138. doi: 10.1523/JNEUROSCI.1590-05.2005

Kavikanga, J., Chanel, G., Cowley, B., Ekman, I., Salminen, M., and Jarvela, S. (2011). A Review of the use of psychophysiological methods in game research. *J. Gaming Virtual Worlds* 3, 181–199. doi: 10.1386/jgvw.3.3.181_1

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

Kim, K., Rosenthal, M. Z., Zielinski, D., and Brady, R. (2012). "Comparison of desktop, head mounted display, and six wall fully immersive systems using a stressful task," *Paper Presented at the Virtual Reality Short Papers and Posters (VRW)* (Piscataway, NJ: IEEE), 143–144. doi: 10.1109/VR.2012.6180922

Kizony, R., Weiss, P. L. T., Shahar, M., and Rand, D. (2006). TheraGame: A home based virtual reality rehabilitation system. *Int. J. Disabil. Hum. Dev.* 5, 265–270. doi: 10.1515/IJDHD.2006.5.3.265

Kober, S., and Neuper, C. (2012). Using auditory event-related EEG potentials to assess presence in virtual reality. *Int. J. Hum. Comput. Stud.* 70, 577–587. doi: 10.1016/j.ijhcs.2012.03.004

Kober, S., and Neuper, C. (2013). Personality and presence in virtual reality: does their relationship depend on the used presence measure? *Int. J. Hum. Comput. Interact.* 56, 1763–1767. doi: 10.1080/10447318.2012.668131

Kober, S. E., Kurzmann, J., and Neuper, C. (2012). Cortical correlate of spatial presence in 2D and 3D interactive virtual reality: an EEG study. *Int. J. Psychophysiol.* 83, 365–374. doi: 10.1016/j.ijpsycho.2011.12.003

Koivisto, M., and Grassini, S. (2016). Neural processing around 200 ms after stimulus-onset correlates with subjective visual awareness. *Neuropsychologia* 84, 235–243. doi: 10.1016/j.neuropsychologia.2016.02.024

Koivisto, M., Grassini, S., Salminen-Vaparanta, N., and Revonsuo, A. (2017). Different electrophysiological correlates of visual awareness for detection and identification. *J. Cogn. Neurosci.* 29, 1621–1631. doi: 10.1162/jocn_a_01149

Koivisto, M., Salminen-Vaparanta, N., Grassini, S., and Revonsuo, A. (2016). Subjective visual awareness emerges prior to P3. *Eur. J. Neurosci.* 43, 1601–1611. doi: 10.1111/ejn.13264

Kurniawan, H., Maslov, A. V., and Pechenizkiy, M. (2013). "Stress detection from speech and galvanic skin response signals," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems* (Porto: IEEE), 209–214. doi: 10.1109/CBMS.2013.6627790

Lara, T., Molina, E., Madrid, J. A., and Correa, Á. (2018). Electroencephalographic and skin temperature indices of vigilance and inhibitory control. *Psicol. J.* 39, 223–260. doi: 10.2478/psicolj-2018-0010

Lee, K. M. (2004). Presence, explicated. *Commun. Theory* 14, 27–50. doi: 10.1111/j.1468-2885.2004.tb00302.x

Lee, M., Bruder, G., and Welch, G. (2017). Exploring the effect of vibrotactile feedback through the floor on social presence in an immersive virtual environment. *Virtual Real.* 2017, 105–111. doi: 10.1109/VR.2017.7892237

Lessiter, J., Freeman, J., Koegh, E., and Davidoff, J. (2001). A cross-media presence questionnaire: the sense of presence inventory. *Presence Teleop. Virtual Environ.* 10, 282–297. doi: 10.1162/105474601300343612

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med.* 6:e1000100. doi: 10.1371/journal.pmed.1000100

Lin, J., Duh, H., Parker, D., Abi-Rached, H., and Furness, T. (2002). Effects of field of view on presence, enjoyment, memory, and simulator sickness in a virtual environment. *Proc. IEEE Virtual Real.* 2002, 164–171. doi: 10.1109/VR.2002.996519

Loomis, J. M. (1992). "Presence and distal attribution: phenomenology, determinants, and assessment," in *Human Vision, Visual Processing, and Digital Display III, Vol. 1666* (San Jose, CA: International Society for Optics and Photonics), 590–595.

Luciani, A., and Cadoz, C. (2007). *Enaction and Enactive Interfaces: A Handbook of Terms.* Grenoble: Enactive System Books.

Lum, H. C., Greatbatch, R., Waldfogle, G., and Benedict, J. (2018). "How immersion, presence, emotion, & workload differ in virtual reality and traditional game mediums," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 62* (Los Angeles, CA: SAGE Publications), 1474–1478. doi: 10.1177/1541931218621334

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

Makransky, G., Terkildsen, T. S., and Mayer, R. E. (2017b). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learn. Instr.* 60, 225–236. doi: 10.1016/j.learninstruc.2017.12.007

Mania, K., and Chalmers, A. (2001). The effects of levels of immersion on memory and presence in virtual environments: a reality centered approach. *Cyberpsychol. Behav.* 4, 247–264. doi: 10.1089/109493101300117938

Maslach, C., Marshall, G., and Zimbardo, P. G. (1972). Hypnotic control of peripheral skin temperature: a case report. *Psychophysiology* 9, 600–605. doi: 10.1111/j.1469-8986.1972.tb00769.x

McMahan, A. (2003). "Immersion, engagement, and presence: a method for analyzing 3-D video games," in *The Video Game Theory Reader,* eds M. J. P. Wolf and P. Bernard (New York, NY: Routledge), 67–86.

Meehan, M. (2001). *Physiological reaction as an objective measure of presence in virtual environments* (Doctoral Dissertation). University of North Carolina at Chapel Hill.

Meehan, M., Insko, B., Whitton, M., and Brooks, F. (2002). Physiological measures of presence in stressful virtual environments. *ACM Trans. Graph.* 21, 645–652. doi: 10.1145/566654.566630

Michel, C. M., and Brunet, D. (2019). EEG Source Imaging: a practical review of the analysis steps. *Front. Neurol.* 10:325. doi: 10.3389/fneur.2019.00325

Millán, J. D. R., Rupp, R., Müller-Putz, G. R., Murray-Smith, R., Giugliemma, C., Tangermann, M., et al. (2010). Combining brain-computer interfaces and assistive technologies: state-of-the-art and challenges. *Front. Neurosci.* 4:161. doi: 10.3389/fnins.2010.00161

Morrison, A., Oulasvirta, A., Peltonen, P., Lemmela, S., Jacucci, G., Reitmayr, G., et al. (2009). "Like bees around the hive: a comparative study of a mobile augmented reality map," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA: ACM), 1889–1898. doi: 10.1145/1518701.1518991

Näätänen, R., Kujala, T., and Light, G. (2019). *Mismatch Negativity: A Window to the Brain.* Oxford, UK: Oxford University Press. doi: 10.1093/oso/9780198705079.001.0001

Nacke, L., and Lindley, C. (2008). Boredom, immersion, flow- a pilot study investigating player experience. *Measurement* 24, 1–5.

Nagai, Y., Critchley, H. D., Featherstone, E., Trimble, M. R., and Dolan, R. J. (2004). Activity in ventromedial prefrontal cortex covaries with sympathetic skin conductance level: a physiological account of a "default mode" of brain function. *Neuroimage* 22, 243–251. doi: 10.1016/j.neuroimage.2004.01.019

Nash, E. B., Edwards, G. W., Thompson, J. A., and Barfield, W. (2000). A review of presence and performance in virtual environment. *Int. J. Hum. Comput. Int.* 12, 1–41. doi: 10.1207/S15327590IJHC1201_1

Nichols, S., Haldane, C., and Wilson, J. R. (2000). Measurement of presence and its consequences in virtual environments. *Int. J. Hum. Comput. Stud.* 52, 471–491. doi: 10.1006/ijhc.1999.0343

Nilsson, N., Nordahl, R., and Serafin, S. (2016). Immersion revisited: a review of existing definitions of immersion and their relation to different theories of presence. *Hum. Technol.* 12, 108–134. doi: 10.17011/ht/urn.201611174652

Nordin, A. I., Denisova, A., and Cairns, P. (2014). Too many questionnaires: measuring player experience whilst playing digital games. *Paper Presented at the The Seventh York Doctoral Symposium on Computer Science and Electronics* (York).

Nunez, D., and Blake, E. (2001). "Cognitive presence as a unified concept of virtual reality effectiveness," in *Proceedings of the 1st International Conference on Computer Graphics, Virtual Reality and Visualisation*, ed S. N. Spencer (New York, NY: ACM), 115–118. doi: 10.1145/513867.513892

Oh, Y., and Rosakranse, C. (2014). "Measuring presence: the use of trends of five canonical presence questionnaires from 1998-2012," in *Proceedings of the 15th International Workshop on Presence* (Wien: facultas.wuv).

Pallavicini, F., Pepe, A., and Minissi, M. E. (2019). Gaming in virtual reality: what changes in terms of usability, emotional response and sense of presence compared to non-immersive video games? *Simul. Gaming* 50, 136–159. doi: 10.1177/1046878119831420

Papachristos, N. M., Vrellis, I., and Mikropoulos, T. A. (2017). "A comparison between oculus rift and a low-cost smartphone VR headset: Immersive user experience and learning," in *Proceedings IEEE 17th International Conference on Advanced Learning Technologies (ICALT)* (Timişoara), 477–481. doi: 10.1109/ICALT.2017.145

Poels, K., Ijsselsteijn, W., and Kort, Y. (2012). Pleasure to play, arousal to stay: the effect of player emotions on digital game preferences and playing time. *Cyberpsychol. Behav. Soc. Netw.* 15, 1–6. doi: 10.1089/cyber.2010.0040

Price, M., and Anderson, P. (2007). The role of presence in virtual reality exposure therapy. *J. Anxiety Disord.* 21, 742–751 doi: 10.1016/j.janxdis.2006.11.002

Riecke, B. E., and Schulte-Pelkum, J. (2015). "An integrative approach to presence and self-motion perception research," in *Immersed in Media*, eds M. Lombard, F. Biocca, J. Freeman, W. IJsselsteijn, and R. Schaevitz (Cham: Springer), 187–235.

Riva, G., Mantovani, F., Capideville, C., Preziosa, A., Morganti, F., and Villani, D. (2007). Affective interactions using virtual reality: the link between emotion and presence. *Cyberpsychol. Behav.* 10, 45–56. doi: 10.1089/cpb.2006.9993

Rose, T., and Chen, K. B. (2018). "Effect of levels of immersion on performance and presence in virtual occupational tasks," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 62* (Philadelphia, PA: Sage), 2079–2083. doi: 10.1177/1541931218621469

Salanitri, D., Lawson, G., and Waterfield, B. (2016). "The relationship between presence and trust in virtual reality," in *Proceedings of the European Conference on Cognitive Ergonomics* (New York, NY). doi: 10.1145/2970930.2970947

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

Schubert, T., Friedmann, F., and Regenbrecht, H. (2001). The Experience of presence: factor analytic insights. *Presence Teleop. Virtual Environ.* 10, 266–281. doi: 10.1162/105474601300343603

Schuemie, M., Staaten, P., Krijn, M., and Mast, C. (2001). Research on presence in virtual reality: a survey. *Cyperpsychol. Behav.* 4, 183–190. doi: 10.1089/109493101300117884

Serafin, S., Erkut, C., Kojs, J., Nilsson, N. C., and Nordahl, R. (2016). Virtual reality musical instruments: state of the art, design principles, and future directions. *Comput. Music J.* 40, 22–40. doi: 10.1162/COMJ_a_00372

Sharples, S., Cobb, S. V., Moody, A., and Wilson, J. R. (2008). Virtual reality induced symptoms and effects (VRISE): comparison of head mounted display (HMD), desktop and projection display systems. *Displays* 29, 58–69. doi: 10.1016/j.displa.2007.09.005

Sjölie, D. (2012). Presence and general principles of brain functions. *Interact. Comput.* 24, 193–202. doi: 10.1016/j.intcom.2012.04.004

Skarbez, R., Brooks, F. P., and Whitton, M. C. (2017). "Immersion and coherence in a visual cliff environment," in *IEEE Virtual Reality (VR)* (Los Angeles, CA: IEEE), 397–398. doi: 10.1109/VR.2017.7892344

Slater, M. (1999). Measuring presence: a response to the Witmer and Singer presence questionnaire. *Presence Teleop. Virtual Environ.* 8, 560–565. doi: 10.1162/105474699566477

Slater, M. (2002). Presence and the sixth sense. *Presence Teleoper. Virtual. Environ.* 11, 435–439. doi: 10.1162/105474602760204327

Slater, M. (2003). A note on presence terminology. *Presence Connect.* 3, 1–5.

Slater, M. (2004). How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence Teleop. Virtual Environ.* 13, 484–493. doi: 10.1162/1054746041944849

Slater, M., Spanlang, B., and Corominas, D. (2010). Simulating virtual environments within virtual environments as the basis for a psychophysics of presence. *ACM Transactions on Graphics (TOG)* 29:92. doi: 10.1145/1778765.1778829

Slater, M., and Steed, A. (2000). A virtual presence counter. *Presence Teleop. Virtual Environ.* 9, 413–434. doi: 10.1162/105474600566925

Slater, M., Usoh, M., and Steed, A. (1994). Depth of presence in virtual environments. *Presence Teleop. Virtual Environ.* 3, 130–144. doi: 10.1162/pres.1994.3.2.130

Slater, M., and Wilbur, S. (1997). A framework for immersive virtual environments five: speculations on the role of presence in virtual environments. *Presence Teleop. Virtual Environ.* 6, 603–616. doi: 10.1162/pres.1997.6.6.603

Sohre, N., Mackin, C., Interrante, V., and Guy, S. (2017). *Evaluating Collision Avoidance Effects on Discomfort in Virtual Environments.* (Los Angeles, CA: IEEE), 1–5. doi: 10.1109/VHCIE.2017.7935624.

Staljanssens, W., Strobbe, G., Holen, R. V., Birot, G., Gschwind, M., Seeck, M, et al. (2017). Seizure onset zone localization from ictal high-density EEG in refractory focal epilepsy. *Brain Topogr.* 30, 257–271. doi: 10.1007/s10548-016-0537-8

Steele, C. (ed.). (2012). *Applications of EMG in Clinical and Sports Medicine.* BoD-Books on Demand.

Steuer, J. (1992). Defining virtual reality: dimensions determining telepresence. *J. Commun.* 42, 73–93. doi: 10.1111/j.1460-2466.1992.tb00812.x

Stevens, J. A., and Kincaid, J. P. (2015). The relationship between presence and performance in virtual simulation training. *Open J. Model. Simul.* 3, 41–48. doi: 10.4236/ojmsi.2015.32005

Suvajdzc, M., Bihorac, A., Parisa, R., and Joel, T. O. (2018). Virtual reality and human consciousness: the use of immersive environments in delirium therapy. *Technoetic Arts* 16, 75–83. doi: 10.1386/tear.16.1.75_1

Szczurowski, K., and Smith, M. (2017). "Measuring presence: hypothetical quantitative framework," in *23rd International Conference on Virtual System & Multimedia (VSMM)* (Dublin: IEEE), 1–8. doi: 10.1109/VSMM.2017.8346261

Tamborini, R., and Skalski, P. (2006). "The role of presence in the experience of electronic games," in *Playing Video Games: Motives, Responses, and Consequences*, eds P. Vorderer and J. Bryant (New York, NY: Routledge), 225–240.

Tarrant, J., Viczko, J., and Cope, H. (2018). Virtual reality for anxiety reduction demonstrated by Quantitative EEG: a pilot study. *Front. Psychol.* 9:1280. doi: 10.3389/fpsyg.2018.01280

Tauscher, J. P., Schottky, F. W., Grogorick, S., Bittner, P. M., Mustafa, M., and Magnor, M. (2019). "Immersive EEG: evaluating electroencephalography in virtual reality," in *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (Osaka: IEEE), 1794–1800. doi: 10.1109/VR.2019.8797858

Teplan, M. (2002). Fundamentals of EEG measurement. *Meas. Sci. Rev.* 2, 1–11.

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

Usoh, M., Arthur, K., Whitton, M. C., Bastos, R., Steed, A., Slater, M., et al. (1999). "Walking> walking-in-place> flying, in virtual environments," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (New York, NY), 359–364.

van Ravenswaaij-Arts, C. M., Kollee, L. A., Hopman, J. C., Stoelinga, G. B., and van Geijn, H. P. (1993). Heart rate variability. *Ann. Intern. Med.* 118, 436–447. doi: 10.7326/0003-4819-118-6-199303150-00008

van Schaik, P., Turnbull, T., van Wersch, A., and Drummond, S. (2004). Presence within a mixed reality environment. *Cyberpsychol. Behav.* 7, 540–552. doi: 10.1089/cpb.2004.7.540

Vorderer, P., Hartmann, T., and Klimmt, C. (2003). "Explaining the enjoyment of playing video games: the role of competition," in *Proceedings of the Second International Conference on Entertainment Computing* (Pittsburgh, PA: Carnegie Mellon University), 1–9.

Vorderer, P., Wirth, W., Gouveia, F., Biocca, F., Saari, T., and Jiancke, F. (2004). *Mec Spatial Presence Questionnaire: Short Documentation and Instructions for Application.* Report to the European Community, Project Presence.

Weber, R., Behr, K., Tamborini, R., Ritterfeld, U., and Mathiak, K. (2009). What do we really know about first-person shooter games? An event related high resolution content analysis. *J. Comput. Mediat. Commun.* 14, 1016–1037. doi: 10.1111/j.1083-6101.2009.01479.x

Weech, S., Kenny, S., and Barnett-Cowan, M. (2019). Presence and cybersickness in virtual reality are negatively related: a review. *Front. Psychol.* 10:158. doi: 10.3389/fpsyg.2019.00158

Williams, K., and Mascioni, M. (2017). *The Out-of-Home Immersive Entertainment Frontier: Expanding Interactive Boundaries in Leisure Facilities.* New York, NY: Routledge.

Witmer, B. G., and Singer, M. J. (1998). Measuring presence in virtual environments: a presence questionnaire. *Presence* 7, 225–240.

Youngblut, C., and Huie, O. (2003). "The relationship between presence and performance in virtual environments: results of a VERTS study," in *IEEE Virtual Reality* (Los Angeles, CA), 277–278.

APPENDIX

Table of Abbreviations		
Abbreviation	**Explanation**	**Category**
IVT	Immersive Visual Technology	General
VE	Virtual Environment	
VR	Virtual Reality	
E2I	E^2I Questionnaire (Lin et al., 2002)	Questionnaires
IEQ	Immersive Experience Questionnaire (Jennett et al., 2008)	
IPQ	Igroup Presence Questionnaire (Schubert et al., 2001)	
ITC-SoPI	ITC-sense of presence inventory (Lessiter et al., 2001)	
ITQ	Immersive Tendencies Questionnaire (Witmer and Singer, 1998)	
MEC-SPQ	MEC Spatial Presence Questionnaire (Vorderer et al., 2004)	
MPS	Multimodal Presence Scale (Makransky et al., 2017a)	
MRJPQ	Modified Reality Judgment and Presence Questionnaire (Anderson et al., 2017)	
PQ	Presence Questionnaire (Witmer and Singer, 1998).	
SUS	Slater-Usoh-Steed Questionnaire (Slater and Steed, 2000)	
UCL	University College of London Questionnaire (Usoh et al., 1999; Meehan, 2001)	
BOLD	Blood-oxygen-level-dependent (signal)	Physiology
ECG/EKG	Electrocardiography	
EDA	Electrodermal Activity	
EEG	Electroencephalography	
EMG	Electromyography	
ERP(s)	Event-related potential(s)	
fMRI	Functional Magnetic Resonance Imaging	
GSR	Galvanic Skin Responses	
HR	Heart Rate	
PPG	Photoplethysmography	
SC	Skin Conductance	
ST	Skin Temperature	

Embodied Emotion Regulation: The Influence of Implicit Emotional Compatibility on Creative Thinking

Li Wu[1], Rong Huang[1], Zhe Wang[1], Jonathan Nimal Selvaraj[2], Liuqing Wei[1], Weiping Yang[1] and Jianxin Chen[1]*

[1] *Department of Psychology, Faculty of Education, Hubei University, Wuhan, China,* [2] *College of Life Science, Hubei University, Wuhan, China*

The regulatory effect of embodied emotion on one's general emotion and the impact of the compatibility or incompatibility of the two types of emotion on creative thinking are still debatable. The purpose of this study is to investigate these issues experimentally. In Experiment 1, participants completed an explicit positive and negative emotion test [Positive and Negative Affect Schedule (PANAS)] and an implicit positive and negative emotion test [Implicit Positive and Negative Affect Test (IPANAT)] twice on a computer after emotional video priming was used to induce negative emotions and facial expression manipulation was performed to induce embodied positive or negative emotions. It was found that maintaining the expression of a suppressed smile was helpful in regulating negative emotions ($p = 0.047$). Specifically, the implicit negative emotions induced by facial expression manipulation had a positive regulating effect on the implicit negative emotions induced by the video (T1, $M = 47.813$; to T2, $M = 44.188$). In Experiment 2, the positive or negative emotions of the participants were induced using emotional videos, and facial expression manipulation was used to induce their embodied positive or negative emotions. Then, the participants completed a creative test by completing alternative use tasks (AUTs) and Chinese character riddles. The AUT fluency score in the emotionally compatible group was significantly higher than that in the emotionally incompatible group ($p = 0.032$), but while experiencing negative emotions, the emotionally compatible group had a significantly higher originality score and insight in Chinese character riddle score than the emotionally incompatible group ($p = 0.017$, $p = 0.004$). Therefore, embodied negative emotion has a significant regulating effect on implicit negative emotion. The compatibility of emotion activated by facial expression and viewing a video contributes to creative thinking, whereas the incompatibility of emotion hinders creative thinking. The compatibility of emotion under positive emotions improved thinking fluency, whereas under negative emotions, it activated originality and insight in creative thinking. The influence of such emotional compatibility on creative thinking may be due to the regulating effect of embodied emotions on implicit emotions induced by emotional stimuli.

Keywords: embodied emotions, emotional compatibility, emotional incompatibility, implicit emotions, creative thinking

INTRODUCTION

Creative thinking refers to the cognitive ability of an individual to create and develop new, valuable things. Creative thinking can also be defined as the cognitive ability to generate ideas, insights, and solutions that are original and flexible (Amabile, 1983; Craft et al., 2013). Fluency, originality, flexibility, and elaboration act as supportive roles in accelerating the creative thinking cognition process (Runco and Acar, 2012). The interactive role of emotions in the creative thinking process is one of the most intriguing research topics on creative thinking (Baas et al., 2008; Ding et al., 2015; Mastria et al., 2019).

Although the existence of a close relationship between emotions and creative thinking has been reported widely, there is no clear consensus on how these connections exist in the cognitive process. Studies have reported that individuals with positive emotions activated positive information connections and cognitive flexibility to enhance the fluency, originality, and insight; but individuals' negative emotions had negative impact on the creative thinking process (Isen et al., 1986; Fernández-Abascal and Díaz, 2013; Mastria et al., 2019). However, some studies supported that negative emotions can stimulate an individual's exploration of the real environment to improve the creative thinking process (Damian and Robins, 2012; Eastwood et al., 2012; Van Tilburg and Igou, 2012). Baas et al. (2013) proposed the dual pathway to creativity model, where approach-related traits such as positive affectivity and power motivation can accelerate creativity, as they enhance cognitive flexibility, but avoidance-related traits such as negative affectivity under the right circumstances do increase creativity, as they enhance cognitive persistence.

Embodied emotion theory indicates that emotional expression, perception, processing, and understanding are closely related to individuals' physical arousal. Various aspects of cooperative relationships between physical actions and cognitive processes do exist in effective manner (Niedenthal, 2007; Glenberg et al., 2013). The muscle feedback signal transmitted by the body state triggers a unique neural activation pattern in the brain. These neural activation patterns represent unconscious emotions, and the embodied emotions belong to the category of implicit emotions (Strack and Deutsch, 2004; Wiers and Stacy, 2006; Neal and Chartrand, 2011). On the basis of the theory of embodied cognition, researchers have explored advanced cognitive activities, that is, the embodiment of creative thinking (Stanciu, 2015). On comparing the positive and negative emotional facial expressions, the positive and happy affect accelerated the flow of creative divergent thinking, whereas negative facial affect had no significant impact on creative thinking (Fernández-Abascal and Díaz, 2013).

An individuals' emotions in daily life are quite complex. Emotions can be divided into explicit emotions, which the individual is aware of, and implicit emotions, which the individual is not aware of. Explicit and implicit emotions can form a compatible combination (i.e., the two emotional valences are consistent) and an incompatible combination (i.e., the two emotional valences are opposite). Studies have shown that different emotions being superimposed on each other will have a regulating effect on emotions overall (Gyurak et al., 2011; Koole and Rothermund, 2011). Riskind (1984) suggested that when an explicit negative emotion (failure experience) and an implicit negative emotion (slumped posture) formed a compatible emotion interaction, the feelings of helplessness and depression and motivation deficits are reduced, thereby playing a positive role in regulating negative emotions. Veenstra et al. (2017) reported that after inducing negative emotions, the stooped posture group showed mood recovery to a lesser extent than the upright and control groups, and the stooped posture created more negative thoughts than the other postures. The conflicting results of these two studies indicated that the intrinsic mechanism involved in the compatibility of embodied emotion and general emotion on emotion regulation needs further research.

Emotional compatibility or incompatibility has an important effect on creative thinking. A study showed that when participants' emotional states were incompatible with the induced mood (low in depression with induced negative emotion), higher creative thinking was observed than in the compatible condition (low in depression with induced positive emotion) (Forgeard, 2011). Huang and Galinsky (2011) asked participants to complete the category association task under the incompatible conditions of recalling a happy thing when frowning and a sad thing when smiling and surprisingly found that participants showed greater innovation with new and unconventional answers for the problems. However, the theory proposed by Leung et al. (2014) believed that when an individual's neuro-emotional traits and experienced emotions form a consistent experience, this consistency could promote performance of creative tasks. Another study also indicated that the compatibility between body posture and emotion was beneficial for creativity because the implicit emotions elicited by body posture matched explicit emotions, which in turn positively enhanced emotions on creativity, that is, through the flexibility or persistence pathway (Hao et al., 2017). The emotion matching hypothesis proposed that the emotions expressed by body postures had a special guiding effect on the self-regulation and information processing of the individual's original emotions. This emotion matching effect was conducive to the regulation of emotions, which was consistent with the protective effect of emotions on individuals in the process of biological evolution (Riskind, 1982). This compatibility of emotion could generate a sense of coherence, which enhanced processing fluency and speeded up decision making of advanced cognition (Alter and Oppenheimer, 2006).

Based on earlier studies, research on the relationship between emotional compatibility and creative thinking is essential and needs further exploration. Many studies have ignored the combined effects of embodied emotions and emotions evoked by external emotional materials, and the impact of the interaction of these two emotions on high-level cognitive activities (such as creative thinking). In view of these concepts, this study explored the regulating effect of the interaction between embodied emotions (by facial expression manipulation technology) and emotions (by video activation). We also explored the impact of positive and negative emotional compatibility on the originality, fluency, and insight in creative thinking.

MATERIALS AND METHODS

Experiment 1

In Experiment 1, we attempted to explore the regulating effect of the interaction between embodied emotions (activated by facial expression manipulation technology) and emotions (activated by emotional videos).

Participants

We calculated the sample size that was necessary to achieve 90% power (which is required to detect an adequate effect) *a priori* using G*Power 3.1 (Faul et al., 2007). Based on $f = 0.40$ (Cohen, 1988), the power calculation yielded a recommended sample size of 45 participants. So 60 participants were selected initially, and they were undergraduate students from Hubei University. The participants were randomly divided into three groups (smile group, smile suppression group, and control group). Data from 12 participants who did not maintain their facial expressions as required were excluded. Thus, the final sample for the present experiment included 48 undergraduates with an average age of 20.46 years ($SD = 2.19$ years). The smile group, smile suppression group, and control group each had 16 participants. In the sample study, 13 students were male. All the students in the sample study were native Chinese speakers with normal hearing, normal vision or corrected vision, and no limb disabilities. Participants gave written informed consent prior to the experiment. The study was approved by the Institutional Ethics Committee at Hubei University.

Experimental Design

A 3 (expression control: smile group, smile suppression group, and control group) × 2 [emotion measurement: Time 1 (T1), Time 2 (T2)] mixed experimental design was used. The dependent variables were the scores and reaction times (RTs) of the Chinese version of the Implicit Positive and Negative Affect Test (IPANAT) and the Chinese version of the Positive and Negative Affect Schedule (PANAS), which were measured twice.

Experimental Materials

Emotion inductions

Participants watched short emotion-appropriate videos to induce corresponding emotion as per the procedures in previous studies (Forgas and East, 2008; Forgeard, 2011). The negative emotion-appropriate clips (3 min 13 s) were excerpted from movie clips "My Brothers and Sisters" (Jin et al., 2009). The film tells the story of an original happy family: because of the changes in the family, overnight, children become orphans and were separated.

Implicit positive and negative affect test

The IPANAT (Quirin et al., 2009) has a total of 36 entries and requires participants to evaluate the degree of association between six meaningless artificial compound words (SAFME, VIKES, TUNBA, TALEP, BELNI, and SUKOV) and six emotional words (six meaningless artificial compound words × 6 emotional words = 36 combinations) on a 4-point scale (1 = completely inconsistent; 2 = somewhat consistent; 3 = more consistent; and 4 = exact match); then, the scores of the six emotional words are calculated. The scores for the positive emotional words (cheerful, happy, and energetic) and negative emotional words (helpless, tense, and inhibited) indicate implicit positive affect (IPA) and implicit negative affect (INA), respectively. The higher the factor score is, the higher the implicit emotion score. A previous study showed that the Chinese version of the IPANAT holds better reliability and validity (Bao and Fu, 2018).

Positive and negative affect schedule

The PANAS (Watson et al., 1988) includes two factors, positive affect (PA) and negative affect (NA), each of which has nine entries. It requires subjects to indicate to what extent they are currently experiencing the emotions described by the presented words on a 5-point scale (1 = very slight or none at all; 5 = very strong). Studies have shown that the Chinese version of the PANAS has good reliability and validity (Bao and Fu, 2018).

Procedure

The experimental design was programmed with E-Prime 2.0 software to record the keys and RT data in the emotional test. Each participant completed the experiment in a separate compartment. There were 36 trials for implicit emotion using the Chinese version of the IPANAT. For each trial, the black "+" fixation point was presented for 250 ms, and then the matching task detection interface of a single artificial word and emotion word was presented. According to their first reaction, participants rated the degree of correspondence between the artificial words presented on the screen and the emotional words from level 1 to level 4 by pressing number keys 1, 2, 3, and 4. The Chinese version of the explicit emotion PANAS was administered similarly to the implicit emotion test, with a total of 18 trials and five rating levels that corresponded to number keys 1, 2, 3, 4, and 5 on the keyboard. To conceal the purpose of the current experiment, the participants were told before the experiment that it was an experiment on intuition.

The experimental steps are shown in **Figure 1**. The participants had 2 min of practice under the guidance of the instructor before the formal test. The formal test required the participants to watch negative emotion videos for negative emotion activation, and then the first measurement (T1) of implicit emotions and explicit emotions was performed after the video watching was completed. Then according to the instructions presented on the computer screen, the instructor assisted the participants in completing the facial expression manipulation by using the facial muscle control paradigm to initiated embodied emotions (Wiswede et al., 2009; **Figure 2**). The positive emotion group held chopsticks horizontally in their teeth, which caused them to shape their faces into smiling, happy facial expressions. The negative emotion group held chopsticks vertically in their lips, which caused them to suppress their smiles and showed sad facial expressions. The control group did not control their facial expressions. Finally, the participants were asked to complete the measurement of implicit emotion and explicit emotion again while maintaining the standard expression control (T2).

FIGURE 1 | The trial structure of Experiment 1.

FIGURE 2 | Facial expression manipulation. **(A)** smile group: biting chopsticks across the teeth. **(B)** Smile suppression group: with vertical chopsticks on the lips.
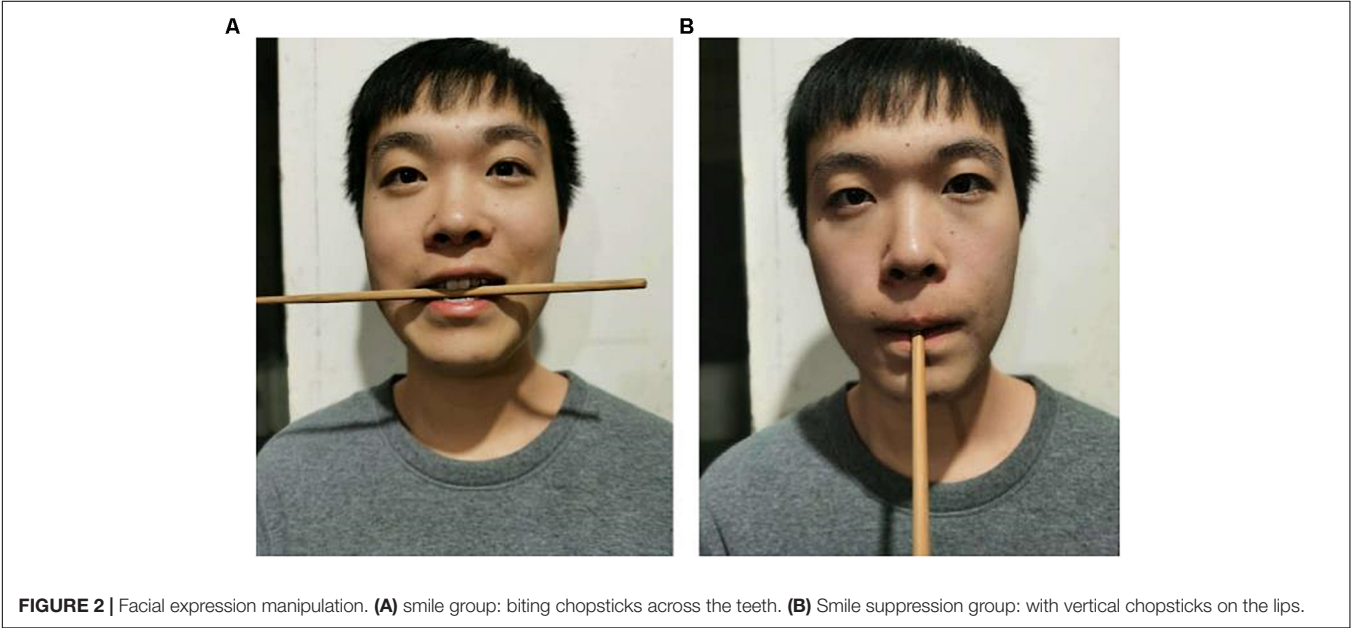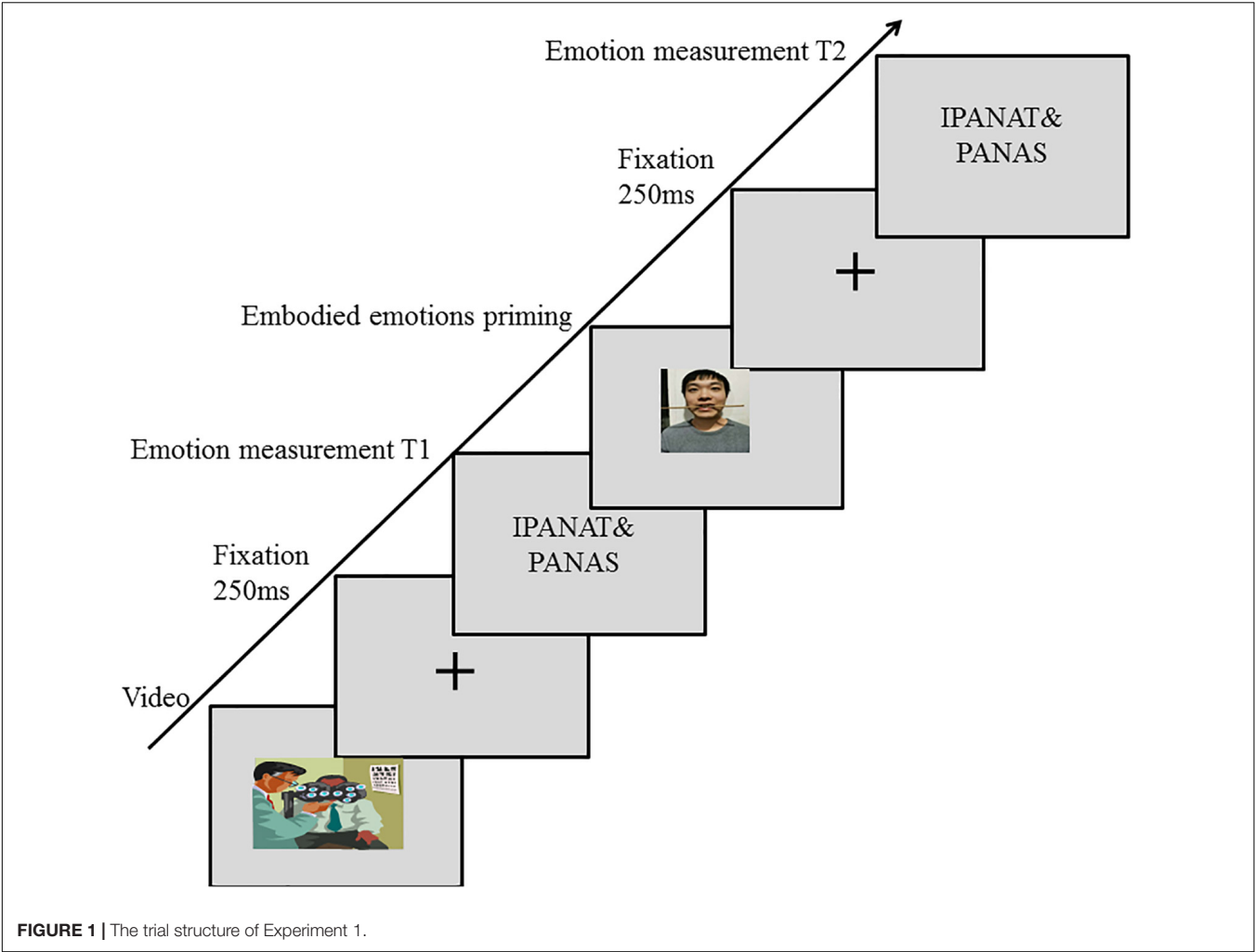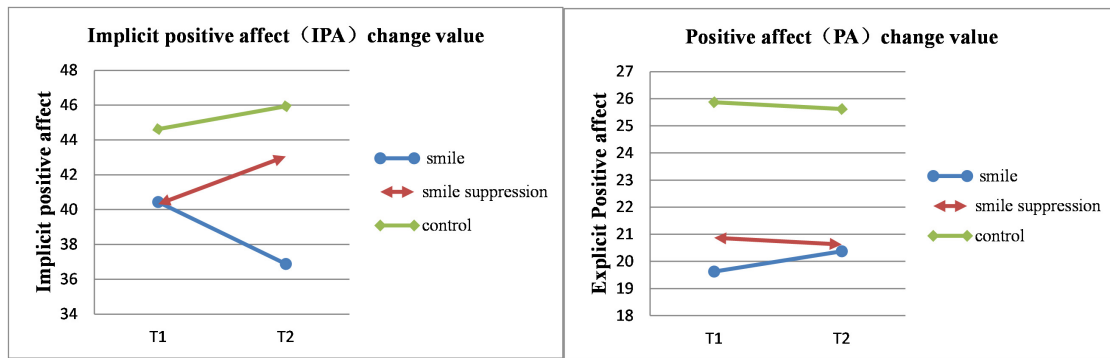
FIGURE 3 | Left: Change value of IPA. Right: Change value of PA.

Results and Discussion

Comparison of positive emotion score changes

It was necessary to determine whether there was a regulating effect of embodied emotions induced by the manipulation of different facial expression on positive emotions. The descriptive statistical results are shown in **Figure 3**. The 2 * 3 ANOVA with the time point (time: T1 vs. T2) of IPA and PA as within-subject factors and facial expression (expression control: smile vs. smile suppression group vs. control) as a between-subject factor was performed (**Figure 4**). The results showed that the main effects of measurement time [$F(1, 45) = 0.028$, $p = 0.868$] and facial expression [$F(2, 45) = 2.038$, $p = 0.142$] for IPA were not significant, but the interaction effect between measurement time and facial expression was significant [$F(2, 45) = 3.675$, $p = 0.033$, $\eta_p^2 = 0.140$]. A further simple-effects analysis found that there were significant changes in IPA between the pretest and posttest in the smile group. The IPA score decreased significantly [$F(1, 45) = 4.260$, $p = 0.045$] from T1 ($M = 40.438$, $SD = 7.375$) to T2 ($M = 36.875$, $SD = 8.221$). There was no significant difference in mood changes between the smile group [$F(1, 45) = 2.540$, $p = 0.118$] and the control group [$F(1, 45) = 0.580$, $p = 0.451$]. In the smile group, the participants' IPA showed a decreasing trend, which means the embodied positive emotions activated by the smile facial expression manipulation did not play a positive role in regulating the negative emotions induced by the video but instead significantly increased the implicit negative emotions. However, the main effects of measurement time [$F(1, 45) = 0.016$, $p = 0.899$], facial expression [$F(2, 45) = 3.063$, $p = 0.057$], and the interaction effect [$F(2, 45) = 0.261$, $p = 0.771$] for PA were not significant. This suggests that expression manipulation had no regulating effect on the explicit emotion of the subjects.

Comparison of negative emotion score changes

It was necessary to determine whether there was a regulating effect of the manipulation of different facial expressions on negative emotions. The descriptive statistical results are shown in **Figure 5**. We performed a 2 * 3 ANOVA of INA and NA with time point (time: T1 vs. T2) as a within-subject factor and facial expression (expression control: smile vs. smile suppression group vs. control) as a between-subject factor (**Figure 6**). The results showed that the main effects of measurement time [$F(1,$

$45) = 0.007$, $p = 0.935$] and facial expression [$F(2, 45) = 2.320$, $p = 0.110$] for INA were not significant. However, the interaction effect of measurement time and facial expression was significant [$F(2, 45) = 3.570$, $p = 0.036$, $\eta_p^2 = 0.137$]. A further simple-effects analysis found that there were significant changes in INA between the pretest and posttest in the smile suppression group. The INA score decreased significantly [$F(1, 45) = 4.190$, $p = 0.047$] from T1 ($M = 47.813$, $SD = 7.530$) to T2 ($M = 44.188$, $SD = 10.297$). There was no significant difference in the mood changes between the smile group [$F(1, 45) = 2.640$, $p = 0.101$] and the control group [$F(1, 45) = 0.320$, $p = 0.575$]. In the smile suppression group, the participants' INA showed a decreasing trend, which means that the embodied negative emotions caused by the suppressing smile facial expression manipulation played a significant positive role in regulating the negative emotions induced by video, which in turn significantly alleviated implicit negative emotion. The main effects of measurement time [$F(1, 45) = 4.480$, $p = 0.040$, $\eta_p^2 = 0.091$] and facial expression [$F(2, 45) = 3.507$, $p = 0.038$, $\eta_p^2 = 0.135$] for NA were significant. This indicates that facial expression manipulation and PANAS scores are more sensitive to negative emotional states than to positive emotional states. The interaction effect [$F(2, 45) = 0.782$, $p = 0.464$] for NA was not significant.

Reaction time

The laboratory measurements of the Chinese version of IPANAT were similar to those of the implicit association tests (IATs) in the field of implicit cognition. In Experiment 1, artificial words were considered target concepts, and emotional words were treated as attribute concepts. When artificial words and emotional words appeared in pairs, the participants pressed keys to determine the degree of correspondence of emotional words and artificial words (e.g., SAFME-happy and SAFME-sad). The key RT was considered, and the key value was excluded. Data on implicit emotional reactions in the T1 and T2 measurements used the D-value calculations proposed by Greenwald et al. (2003) (**Table 1**).

1. The mean and standard deviation of the RT of each group at T1 were calculated. The RT for positive emotional words (e.g., SAFME-happy) in each group
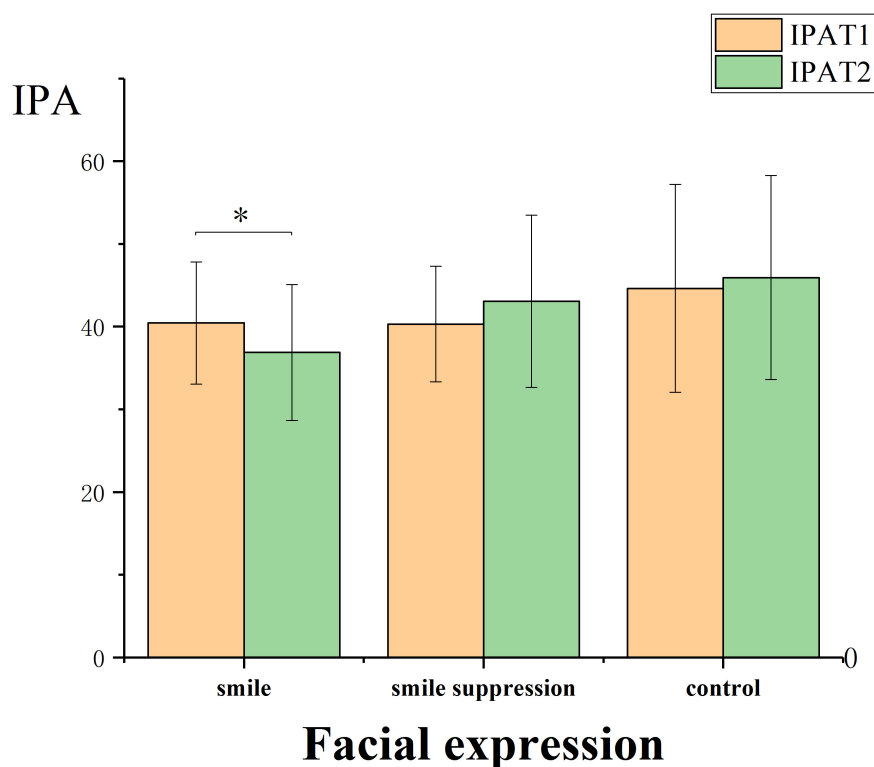
FIGURE 4 | IPA of participants controlled by different facial expressions.
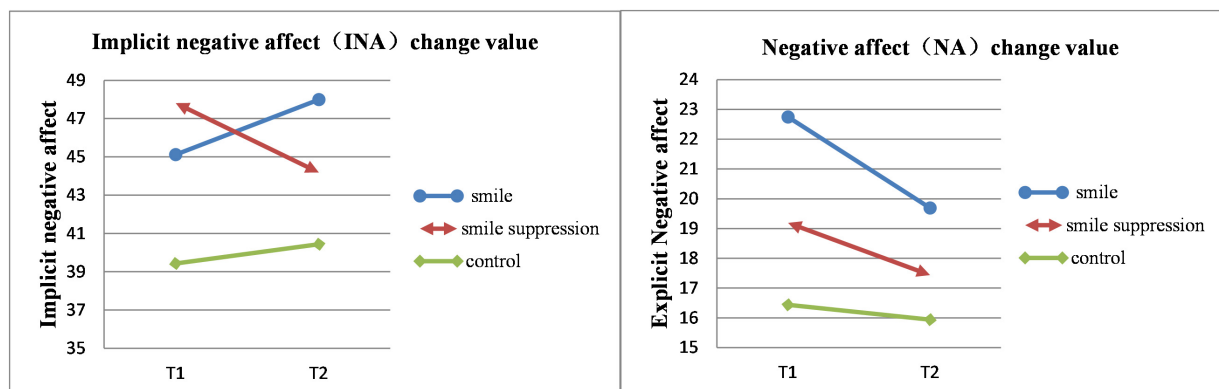
FIGURE 5 | Left: Change value of INA. Right: Change value of NA.

was longer than that for negative emotional words (e.g., SAFME-sad). The study showed that participants in each group preferred negative emotion words at T1 and T2 and that they were in a negative emotional state, which also indicated that emotion induction was effective. According to the difference between a compatible task and an incompatible task in IAT, *the artificial word–negative emotional word* pair was regarded as a compatible task, whereas *the artificial word–positive emotional word* pair was regarded as an incompatible task.

2. The D-value was calculated to reflect the participants' preference level for negative emotion words. The D-value was the mean RT of positive emotion words (incompatible tasks) minus the mean RT of negative emotion words (compatible tasks), which was divided by the standard deviation of the total RT. The difference between the D-values of T1 and T2 was calculated. According to studies of implicit cognition, the difference in the D-value can be understood as the change in the preference level of negative emotion words. The comparison of the difference in D-values between the two

FIGURE 6 | INA of participants controlled by different facial expressions.
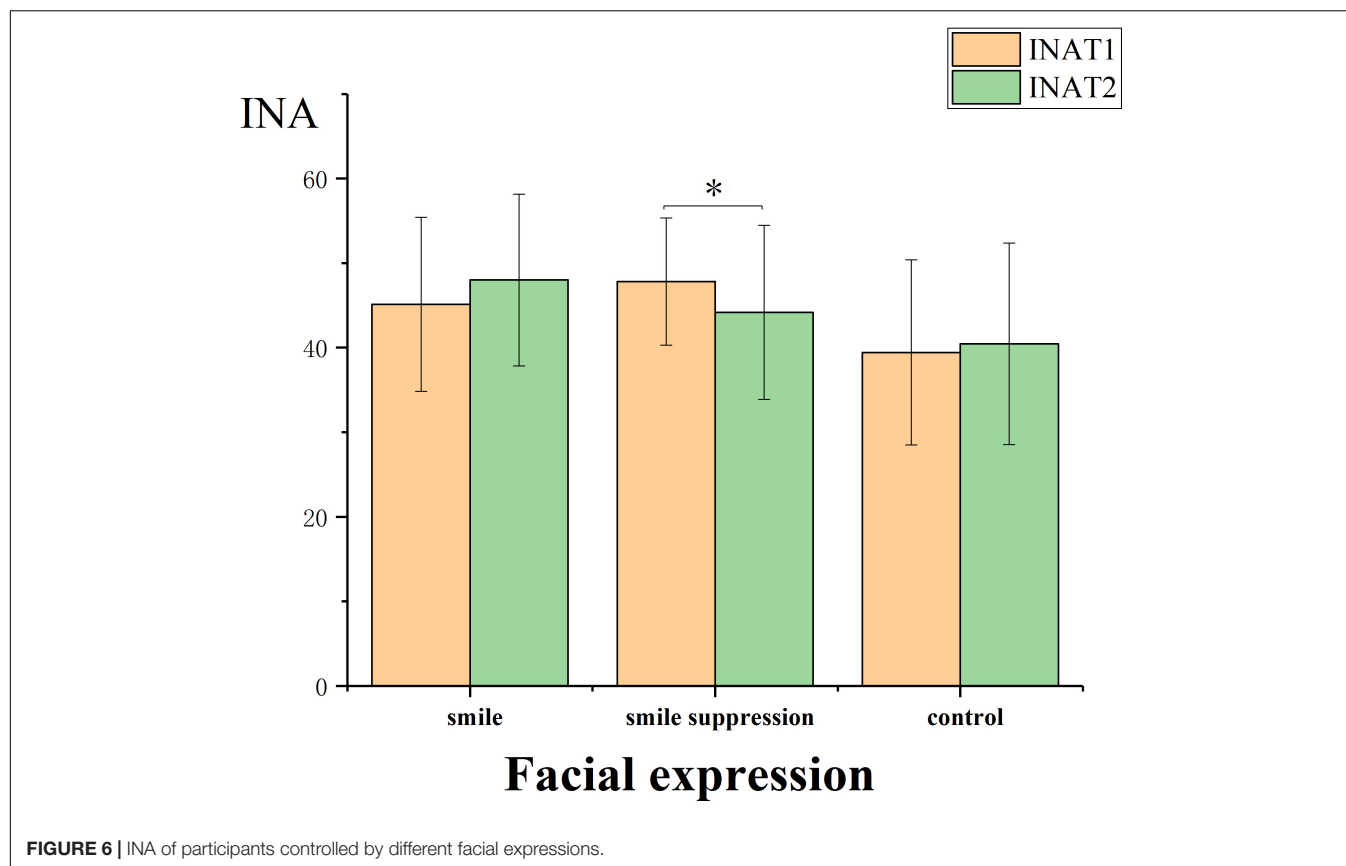
groups showed that the D-value in the smile suppression group (D-value difference = 0.11) was higher than that in the smile group (D-value difference = 0.07) and control group (D-value difference = −0.23). From the perspective of RT, suppressing the expression of a smile could contribute to the positive regulation of negative emotions.

Based on the above analysis of the difference between the score changes of positive emotions and negative emotions and the changes in RT, it is reasonable to deduce that embodied negative emotions initiated by facial expression manipulation technology have a significant regulating effect on the implicit negative emotions initiated by video stimulation.

Experiment 2

In Experiment 1, we found that allowing the participants to suppress the expression of a smile could effectively regulate their implicit negative emotion induced by the video. In other words, when the embodied emotional valence induced by body movement was consistent with the implicit emotion induced by the explicit emotion stimulus, the two compatible emotions had a positive regulating effect on the implicit emotions. To date, research on the relationship between emotion and creative thinking has not reached a consistent conclusion. In particular, the influence of embodied implicit emotion induced by an individual's body movement on creative thinking has

been ignored. The results of Experiment 1 showed that the compatibility of embodied emotions with the valence of emotions evoked by emotional videos could regulate implicit emotions. It is worth exploring whether the regulation of implicit emotions affects the creative thinking process, especially whether there are different influences on fluency, originality, and insight in creative thinking. Based on the results of Experiment 1, Experiment 2 was designed to further explore the impact of the compatibility of embodied emotions (facial emotion manipulation) with the emotions induced by external emotional stimuli (emotional video material) on creative thinking.

Participants

We calculated the sample size that was necessary to achieve 90% power (which is required to detect an adequate effect) *a priori* using G*Power 3.1 (Faul et al., 2007). Based on $f = 0.40$ (Cohen, 1988), this power calculation yielded a recommended sample size of 68 participants. We thus required 80 undergraduate student volunteers from Hubei University to participate in this experiment. The participants were randomly divided into four groups (positive emotion-smile group, positive emotion-smile suppression group, negative emotion-smile group, and negative emotion-smile suppression group), and data from 12 participants who did not maintain their facial expressions as required were excluded. Therefore, the final sample for the study included 68 undergraduates with an average age of 20.01 years ($SD = 1.92$ years). In the overall sample, 27 of

TABLE 1 | Compatible/incompatible reaction time (*ms*) and *D*-value.

		Incompatible task group (Artificial word – positive emotion)		Compatible task group (Artificial word – negative emotion)		Effect		*D*-value		*D*-value Difference
		T1	T2	T1	T2	T1	T2	T1	T2	
Smile group	M	2029.23	1725.21	1836.87	1623.53	192.36	101.68	0.17	0.1	0.07
(*n* = 16)	SD	1078.57	973.14	1218.37	996.4	1136.1	970.2			
Smile suppression group	M	1936.23	1725.73	1870.03	1709.1	66.2	16.63	0.13	0.02	0.11
(*n* = 16)	SD	524.81	637.25	524.33	824.46	517.13	724.9			
Control group	M	1784.01	1508.76	1722.23	1343.97	61.78	164.79	0.11	0.34	−0.23
(*n* = 16)	SD	713.79	553.14	446.54	417.75	586.52	489.39			

the students were males. There were 33 participants assigned to the emotional incompatibility condition and 35 assigned to the emotionally compatible condition; the assignments were random. Participants gave written informed consent prior to the experiment. The study protocol was approved by the Institutional Ethics Committee at Hubei University.

Experimental Design

A between-subject design of 2 (expression control: smile group, smile suppression group) × 2 (video emotion induction: positive, negative) was employed. The dependent variables were the scores for alternative use task (AUT) fluency, originality, and insight in Chinese character riddles.

Experimental Materials

Emotion induction videos

Participants watched short emotion-appropriate videos to induce corresponding negative or positive emotion as in previous studies (Forgas and East, 2008; Forgeard, 2011). The negative emotion-appropriate clips were the same as those in Experiment 1. The positive emotion-appropriate clips (3 min 28 s) were excerpted from movie clips of *Flirting Scholar* (Li et al., 2009). The film tells the interesting story of a Chinese scholar when he was a reading attendant at local officials in ancient times.

Alternative uses task

The AUT proposed by Guilford (1967) was adopted to measure participants' creative thinking. Participants had to produce as many unusual or original uses as possible for common objects, such as paper clips (e.g., "making rings" and "cleaning nails"). The AUT is a reliable indicator of thinking fluency and originality in creative thinking (Runco and Mraz, 1992). Fluency scores were based on the total number of ideas given in the AUT. Originality scores were based on statistically infrequent responses. Scoring was conducted separately by two graduate students majoring in psychology. The answers generated by all participants were collected into a set. Synonyms were identified as the same answer. If the answer appeared to be novel (i.e., 5% or less of the participants in the sample gave the answer), a score of 1 was given, whereas any answer indicated by over 5% of the sample was marked as 0. The interrater agreement [intraclass correlation coefficients (ICCs) = 0.95] was satisfactory. Finally, the fluency and originality scores in solving two problems were averaged for

each participant. In this study, the target task "chopsticks" was selected, and the participants were asked to write as many novel uses of chopsticks on the answer sheets as possible.

Insight in Chinese character riddles

Five pairs of Chinese character riddles were selected as materials from the Chinese Character Riddle Library (Zhen-Zhen et al., 2009), which were used to examine the "prototype inspiration effect" of insight. Qiu et al. believed that the use of Chinese character riddles as experimental materials could effectively explore the cognitive mechanism of insight in creative thinking through prototype inspiration under experimental conditions in Chinese individuals (Zhen-Zhen et al., 2008). Each pair of Chinese character riddles contained a prototype anagram (the prototype riddle was the prototype event in the prototype activation theory) and a target Chinese character riddle (the target riddle was the insight problem in the prototype activation theory). For example, if the prototype Chinese character riddle has a "十个兄弟," then the correct answer is "克," and if the target Chinese character riddle is "九个太阳," then the answer is "旭." The prototype heuristic rate (the correct rate with the prototype heuristic minus the correct rate without the prototype heuristic) was averaged to be 0.58. The scoring was separately conducted by two graduate students majoring in psychology. Each of the five target riddles was scored 1 point for each correct guess and 0 points for a wrong or no guess, and the total score for each correct answer was calculated.

Procedure

In the emotional initiation stage, the participants were required to focus on and watch *My Brothers and Sisters* (Jin et al., 2009) and *Flirting Scholar* (Li et al., 2009) to induce negative and positive emotions, respectively. Then according to the instructions presented on the computer screen, the instructor assisted the participants in completing the facial expression manipulation by using the facial muscle control paradigm to initiated embodied emotions (Wiswede et al., 2009; **Figure 2**). After the standard expression control was maintained, the AUT creativity test phase was performed. The participants were required to follow the main test guideline; that is, they should write as many new practical uses of "chopsticks" on the answer card within 3 min, and the answers should be realistic. During the

entire process, the participants needed to maintain the expression with the chopsticks. They could take a 15-s break after the 3-min timer ended.

After the rest, the insight in Chinese character riddle task was conducted. The participants were required to learn five pairs of prototype riddles and target riddles first. There was no time limit. In this process, the instructor helped the participants understand the prototype riddle. After the participants fully understood the task, they were required to follow the instruction; that is, they should complete the other five pairs of Chinese character riddles quickly and accurately on the answer sheet within 3 min. Participants needed to maintain the expression manipulation with the chopstick during the entire process.

Results and Discussion
Fluency
To analyze the effect of the compatibility of different facial expressions and different emotions on creative thinking, ANOVA was performed using facial expression manipulation and video emotion induction as independent variables and the thinking fluency score as the dependent variable (**Figure 7**). The results showed that the main effects of facial expressions [$F(1, 64) = 0.799$, $p = 0.375$] and explicit emotions [$F(1, 64) = 0.408$, $p = 0.525$] were not significant. However, the interaction between these two factors was significant [$F(1, 64) = 4.512$, $p = 0.038$, $\eta_p^2 = 0.066$]. Simple-effects analysis showed that for the condition in which positive emotion was induced by video, the fluency scores between the smile group and the smile suppression group were significantly different [$F(1, 65) = 4.790$, $p = 0.032$]. The fluency score in the facial smile group ($M = 6.500$) was significantly higher than that in the smile suppression group ($M = 4.588$). However, for the condition in which negative emotion was induced by the video, there was no significant difference in the fluency scores between the facial smile group and the smile suppression group [$F(1, 65) = 0.720$, $p = 0.400$]. This indicated that with positive emotions, the fluency of the group with a constant smile was better than that of the smile suppression group, which suggests that for positive emotions, emotional compatibility is conducive to generating the thinking fluency aspect of creative thinking.

Originality
To analyze the effect of the compatibility of different facial expressions and different emotions on creative thinking, ANOVA was performed using facial expression manipulation and video emotion induction as independent variables and thinking originality score as the dependent variable (**Figure 8**). The results found that the main effects of facial expressions [$F(1, 64) = 0.937$, $p = 0.337$] and explicit emotions [$F(1, 64) = 1.342$, $p = 0.251$] were not significant. However, the interaction between these two factors was significant [$F(1, 64) = 6.699$, $p = 0.012$, $\eta_p^2 = 0.095$]. Furthermore, simple-effects analysis showed that for the condition in which negative emotion was induced by video, the originality scores between the smile group and the smile suppression group were significantly different [$F(1, 65) = 5.990$, $p = 0.017$], and the originality score in the smile group ($M = 1.125$) was significantly lower than that in the smile

suppression group ($M = 2.588$). However, for the condition in which positive emotion was induced by video, there was no significant difference in originality scores between the smile group and the smile suppression group [$F(1, 65) = 1.40$, $p = 0.241$]. This indicated that with negative emotions, the originality of the smile suppression group was better than that of the smile group, which means that with negative emotions, emotional compatibility is conducive to generating the originality aspect of creative thinking.

Insight in Chinese character riddles
To analyze the effect of the compatibility of different facial expressions and different emotions on creative thinking, ANOVA was performed using facial expression manipulation and video emotion induction as independent variables and the insight Chinese character riddle score as the dependent variable (**Figure 9**). The results showed that the main effects of facial expressions [$F(1, 64) = 1.687$, $p = 0.199$] and explicit emotions [$F(1, 64) = 0.323$, $p = 0.572$] were not significant. However, the interaction between these two factors was significant [$F(1, 64) = 8.500$, $p = 0.005$, $\eta_p^2 = 0.117$]. Furthermore, simple-effects analysis showed that for the condition in which positive emotion was induced by the video, there were no significant differences in insight Chinese character riddle scores between the smile group and the smile suppression group [$F(1, 65) = 1.330$, $p = 0.252$]. However, for the condition in which negative emotion was induced by the video, the insight Chinese character riddle score in the smile group ($M = 2.438$) was significantly lower than that in the smile suppression group ($M = 3.647$) [$F(1, 65) = 8.790$, $p = 0.004$]. This indicated that with negative emotions, the insight of the smile suppression group was better than that of the smile group, which means that with negative emotions, emotional compatibility is conducive for generating the insight thinking aspect of creative thinking.

GENERAL DISCUSSION

In this study, we explored the impact of compatibility or incompatibility between emotions induced by facial expressions and emotional videos on creative thinking. In the first part of the study, the participants were measured based on implicit and explicit emotions after watching a negative emotion-inducing video, and then their implicit and explicit emotions were measured again after embodied positive emotions and negative emotions were induced through the manipulation of smiles or suppressed smile facial expressions. The results showed that for the smiling condition, implicit positive emotion decreased, and for the smile suppression condition, implicit negative emotion decreased, which proved that the compatibility of embodied emotion and video-induced implicit emotion has a positive regulating effect on implicit emotion. In the second part of the study, we proved that emotional compatibility promoted the flow of creative thinking, originality, and insight. Specifically, with positive emotions, the AUT fluency score of the emotionally compatible group (i.e., positive emotions/smiling expressions) was significantly higher than that of the emotionally incompatible group (i.e., positive emotions/smile suppression expressions).
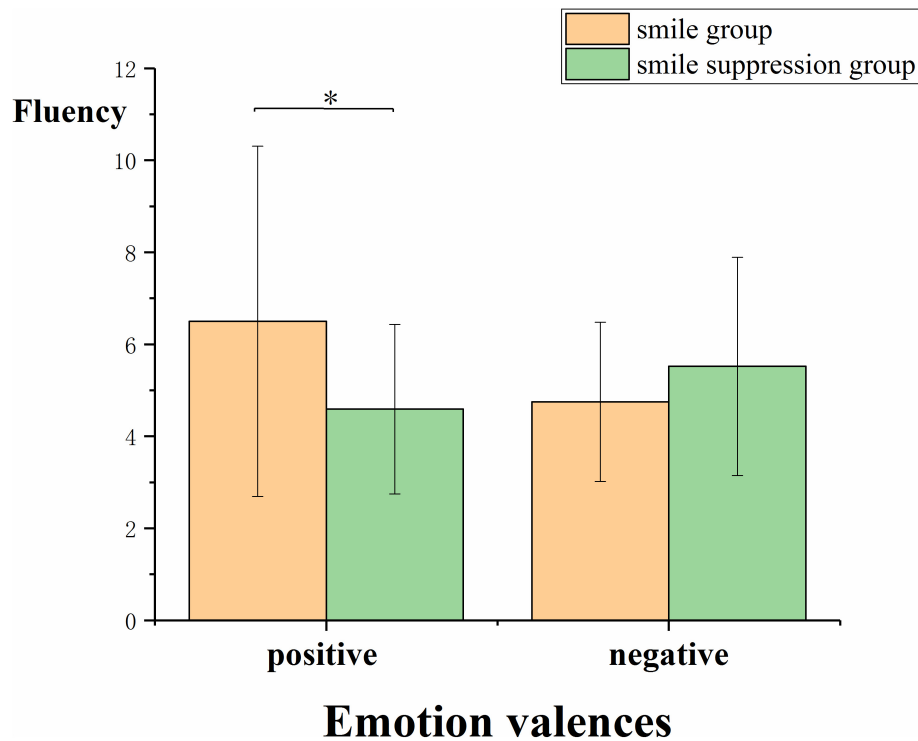
FIGURE 7 | Fluency scores of participants under different emotion-induced conditions and expression manipulation states.
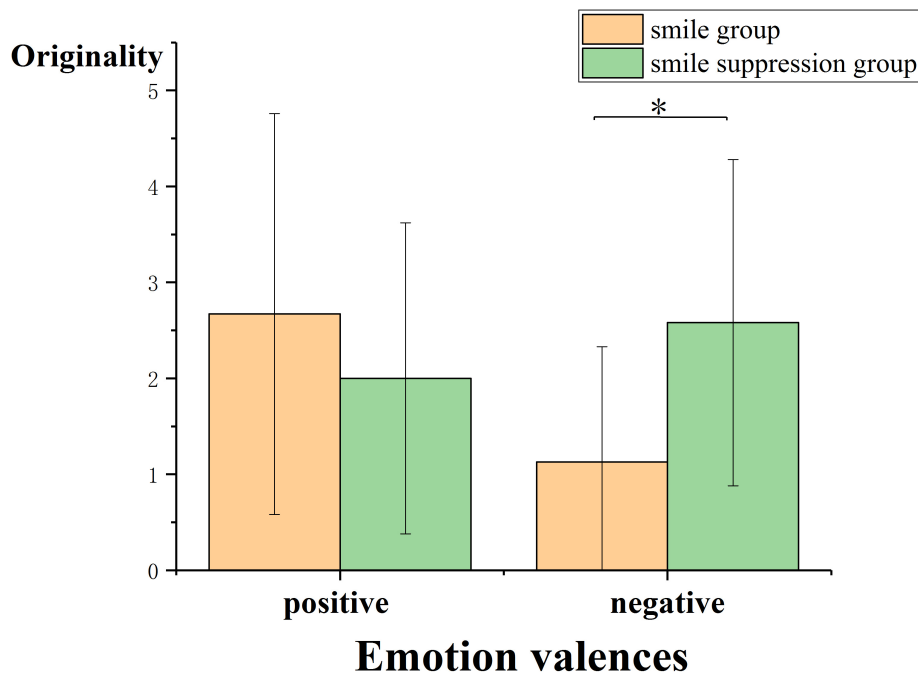
FIGURE 8 | Originality scores of participants under different emotion-induced conditions and expression manipulation states.

However, with negative emotions, the emotionally compatible group (i.e., negative emotions/smile suppression expressions) had a significantly higher originality score and insight in Chinese character riddle score than the emotionally incompatible group (i.e., negative emotion/smile expression). This suggests that the compatibility of emotion while experiencing positive
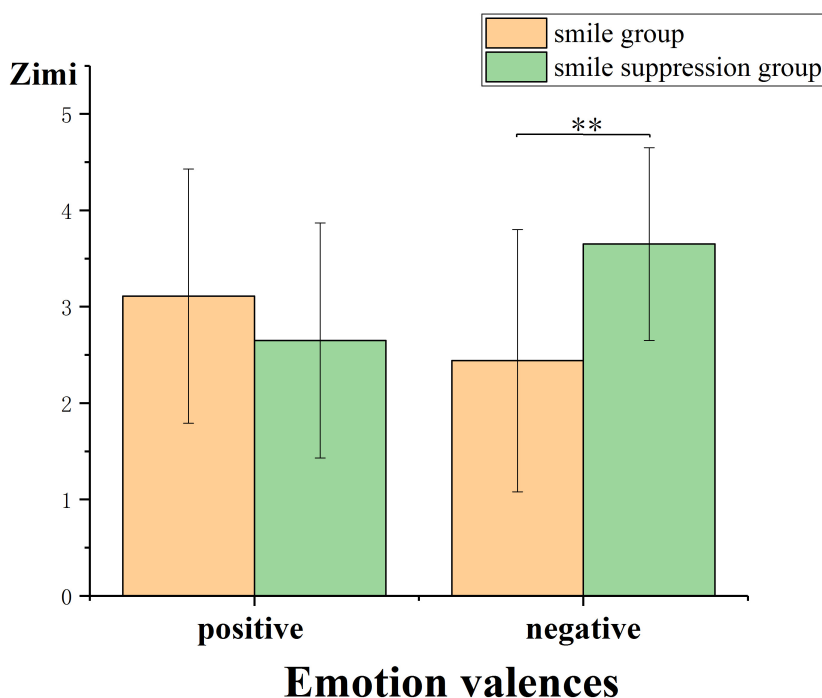
FIGURE 9 | Insight Chinese characters riddle scores of participants under different emotion-induced conditions and expression manipulation states.

emotions enhances the fluency of thinking, whereas the compatibility of emotion while experiencing negative emotions can accelerate the expression of originality and spiritual insight into creative thinking.

Regulating Effect of Embodied Emotions on Emotions

In the first part of the study, embodied positive emotions did not play a positive role in regulating the negative emotions induced by video. Instead, the embodied negative emotions played a positive role in regulating the implicit negative emotions induced by the video, significantly relieving the negative implicit emotions. The experimental results suggested that the compatibility of emotions initiated by the video and embodied emotions initiated by facial expressions helped to regulate implicit emotions. Riskind (1984) and Veenstra et al. (2017) proved that body posture can affect emotional regulation. However, no consensus has been reached on which posture is more conducive to the recovery from negative emotions that have been induced. In Experiment 1, we explored the regulatory effect of embodied emotions induced by manipulating facial expressions on the negative emotions induced by videos, and our results were consistent with the "matching hypothesis" (Riskind, 1982). After negative emotions were experienced, the implicit negative emotion score of the participants who adopted the expression of suppressing a smile was lower than that of the smiling expression group. This indicates that the individuals immersed in negative emotion were not as sad as we imagined. When a person's smiling expression "does not match" the

current emotion (the person has just begun to feel sorrow), the smiling expression had an inhibitory effect on subsequent emotional recovery. The smile and suppressed smile expressions in this study played a special role in guiding and regulating the processing information of the individual. Smiling is a happy expression and more closely matched the participant's emotional state (positive emotion). Suppressing a smile is similar to a sad expression and more closely matched the participant's emotional state (negative emotion), which is more conducive to emotional regulation for emotionally compatible condition.

Fewer cognitive resources are consumed when emotions are better regulated, so creative thinking performance is better when experiencing emotional compatibility than when experiencing emotional incompatibility. The results of our study were inconsistent with those of Veenstra et al. (2017), as they used the self-reporting scale for emotional recovery measurements in their study. The emotions obtained through self-reporting may be distorted owing to social expectations, missing the implicit emotions that reflect the real emotions of the participants. The measurement of embodied implicit emotion should belong to the category of implicit emotions. We believe that indirect measurement was more suitable for measuring implicit emotions than self-reporting (direct measurement) because the autoactivated implicit emotion was only reflected in the explicit emotion scale when it was strong (Quirin et al., 2009). In view of this, in the first part of this study, we made participants evaluate emotionally neutral items and things that rarely have emotions attached to them, such as ancient Chinese hieroglyphs, Russian alphabet artificial words, or English alphabet artificial words. The degree of correlation was used to indirectly measure

the regulation of implicit emotions induced by facial expressions on the emotions induced by video stimulation. Second, any external emotional stimulus not only can induce explicit conscious emotions but can also isolate implicit unconscious emotions behind conscious feelings (Clore and Ketelaar, 1997; Berridge, 1999; Kihlstrom et al., 2000; Winkielman and Berridge, 2004). In other words, in this study, the explicit emotion and the implicit emotion were further separated from the emotions induced by the emotional video. It was found that the embodied emotion induced by facial expression regulated the implicit emotion induced by the emotional stimulus, so it seemed that the measurement of emotion regulation tends to be highly accurate.

The performance of emotional compatibility to promote creative thinking can be explained by the characteristics of implicit emotion regulation. Implicit emotion regulation is effective and fast. Compared with explicit emotion regulation, implicit emotion regulation may be more effective. This is because the process of explicit emotion regulation requires effort and consumes certain self-control resources. When the individual's self-control resources are insufficient, this regulation was not necessarily very effective. However, implicit emotion regulation does not require willpower; it is performed automatically and can be completed without conscious effort or supervision. This process may not consume any or only consume minimal self-control resources, so the process of implicit emotion regulation was more effective and occurs more easily (Koole and Rothermund, 2011; Etkin, 2014; Koole et al., 2015). When emotions are better regulated, fewer cognitive resources are consumed, and creative thinking performance when experiencing emotional compatibility is optimal.

The Influence of Emotional Compatibility on Different Components of Creative Thinking

In the second part of our study (Experiment 2), we observed that emotional compatibility with positive emotions enhanced the fluency of creative thinking compared with that under emotional incompatibility. Emotional compatibility with negative emotions contributed to the originality and insight performance of creative thinking. The dual pathway to creativity model proposes that individuals can generate creative ideas, products, or solutions through either the cognitive flexibility path or the persistence path (De Dreu et al., 2008; Nijstad et al., 2010). Although both smiling expressions (positive emotions) and suppressed smile expressions (negative emotions) can enhance creative thinking, we believed that the cognitive mechanism that affects creative thinking under these two conditions might be different. According to the fluency dimension of creative thinking (see **Figure 6**), a smile under the positive emotional condition will further moderate the increment of negative emotions, and under positive emotional conditions, individuals can improve cognitive flexibility. According to Fredrickson's extension-shaping theory of positive emotions (Fredrickson, 1998), from the perspective of evolution, positive emotions can provide a pleasant and safe situation to expand the attention range of individuals, thereby speeding up the response to new and different stimuli and

ultimately the achievement of cognitive flexibility improvements. When experiencing positive emotions, more attention is paid to internal and subjective data, more cognitive resources are invested, and decisions involve less information about the external environment, resulting in more fluent and original ideas and opinions. Suppressed smile expressions in a negative emotional condition will further moderate the increment of negative emotions as well. Individuals with negative emotions are able to enhance the originality and insight aspects of creative thinking by improving their persistence with respect to thinking about problems (see **Figures 7**, **8**). The mood-input model (Martin and Stoner, 1996) suggested that negative emotions can be regarded as signals of problems in the current environment (Schwarz, 2002), prompting individuals to remain fully alert (Forgas, 2007) and to make additional persistent efforts (George and Zhou, 2002). When experiencing negative emotions, individuals pay more attention to external and objective clues; in this study, the participants needed to suppress the interference of other information (such as riddle expression) when transitioning from the prototype riddle to the target riddle. However, once the direction of a possible breakthrough is determined, people will focus on a specific problem situation with more continuous efforts, ultimately achieving innovation.

The Limitations of the Current Study and Directions for Future Studies

Owing to the complexity of creative thinking and emotions, the current study was only a preliminary exploration of the impact of embodied implicit emotions and explicit emotions on creative thinking. The evidence obtained through behavioral experiments was not sufficient. In future studies, we may use the electroencephalogram (EEG) or event-related potential (ERP) techniques to further explore the changes of neural activity in the brain area when the two emotions are compatible. Although the present study has found that the compatibility of embodied implicit emotions induced by facial expressions and explicit emotions induced by video stimulation has important effects on different components of creative thinking, there are still many factors that may influence creative thinking needed to investigate, such as individual personality, self-efficacy, and goal orientation.

Meanwhile, according to our results, the participants' performance of originality and insight was better under the condition of negative emotion compatibility. In the future, we may need to extend this research to groups with prominent negative emotions – such as depressed individuals. Whether the depressed individuals can really regulate negative emotions through different body postures and further enhance creativity are worth discussing.

CONCLUSION

The study showed that the embodied negative emotion activated by facial expression has a significant regulating effect on the implicit negative emotion activated by video stimulation. The compatibility of emotion activated by facial expression and video

stimulation helps creative thinking, whereas the incompatibility of emotion hinders creative thinking. The compatibility of emotion with a positive emotional condition contributes to the fluency of thinking, whereas the compatibility of emotion with a negative emotional condition contributes to the originality and insight aspects of creative thinking. The influence of such emotional compatibility on creative thinking may be due to the regulating effect of embodied emotions on implicit emotions induced by emotional stimuli.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

ETHICS STATEMENT

Participants gave written informed consent prior to the experiment. The study was approved by the Institutional Ethics Committee at the Hubei University. Written informed consent was obtained from the individual for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

LWu wrote the manuscript and analyzed the data. RH and ZW performed the experiments. JC and LWu conceived and designed the experiments. JS, WY, and LWe revised the manuscript and approved the final version of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01822/full#supplementary-material

REFERENCES

Alter, A. L., and Oppenheimer, D. M. (2006). Predicting short-term stock fluctuations by using processing fluency. *Proc. Natl. Acad. Sci. U.S.A.* 103, 9369–9372. doi: 10.1073/pnas.0601071103

Amabile, T. M. (1983). The social psychology of creativity: a componential conceptualization. *J. Pers. Soc. Psychol.* 45, 357–376. doi: 10.1037/0022-3514.45.2.357

Baas, M., De Dreu, C. K. W., and Nijstad, B. A. (2008). A meta-analysis of 25 years of mood-creativity research: hedonic tone, activation, or regulatory focus? *Psychol. Bull.* 134, 779–806. doi: 10.1037/a0012815

Baas, M., Roskes, M., Sligte, D., Nijstad, B. A., and De Dreu, C. K. W. (2013). Personality and creativity: the dual pathway to creativity model and a research agenda. *Soc. Pers. Psychol. Compass* 7, 732–748. doi: 10.1111/spc3.12062

Bao, J., and Fu, N. (2018). Embodying emotion regulation: the effect of facial expression on implicit affect. *Stud. Psychol. Behav.* 16, 180–187.

Berridge, K. C. (1999). "Pleasure, pain, desire, and dread: hidden core processes of emotion," in *Well-Being: The Foundations of Hedonic Psychology*, eds D. Kahneman, E. Diener, and N. Schwarz (New York, NY: Russell Sage Foundation), 525–557.

Clore, G. L., and Ketelaar, T. (1997). "Minding our emotions: on the role of automatic, unconscious affect," in *Advances in Social Cognition*, Vol. 10, ed. R. S. Wyer (Mahwah, NJ: Lawrence Erlbaum Associates), 105–120.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge.

Craft, A., Cremin, T., Burnard, P., Dragovic, T., and Chappell, K. (2013). Possibility thinking: culminative studies of an evidence-based concept driving creativity? *Education* 3–13, 41, 6–31.

Damian, R. L., and Robins, R. W. (2012). The link between dispositional pride and creative thinking depends on current mood. *J. Res. Pers.* 46, 765–769. doi: 10.1016/j.jrp.2012.07.003

De Dreu, C. K. W., Baas, M., and Nijstad, B. A. (2008). Hedonic tone and activation level in the mood-creativity link: toward a dual pathway to creativity model. *J. Pers. Soc. Psychol.* 94, 739–756. doi: 10.1037/0022-3514.94.5.739

Ding, X., Tang, Y. Y., Deng, Y., Tang, R., and Posner, M. I. (2015). Mood and personality predict improvement in creativity due to meditation training. *Learn. Individ. Dif.* 37, 217–221. doi: 10.1016/j.lindif.2014.11.019

Eastwood, J. D., Frischen, A., Fenske, M. J., and Smilek, D. (2012). The unengaged mind. *Perspect. Psychol. Sci.* 7, 482–495. doi: 10.1177/1745691612456044

Etkin, A. (2014). Emerging insights on implicit emotion regulation. *Neuropsychoanalysis* 13, 42–44. doi: 10.1080/15294145.2011.10773658

Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G*power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/bf03193146

Fernández-Abascal, E. G., and Díaz, M. D. M. (2013). Affective induction and creative thinking. *Creat. Res. J.* 25, 213–221. doi: 10.1080/10400419.2013.783759

Forgas, J. P. (2007). When sad is better than happy: negative affect can improve the quality and effectiveness of persuasive messages and social influence strategies. *J. Exp. Soc. Psychol.* 43, 513–528. doi: 10.1016/j.jesp.2006.05.006

Forgas, J. P., and East, R. (2008). On being happy and gullible: mood effects on skepticism and the detection of deception. *J. Exp. Soc. Psychol.* 44, 1362–1367. doi: 10.1016/j.jesp.2008.04.010

Forgeard, M. J. C. (2011). Happy people thrive on adversity: pre-existing mood moderates the effect of emotion inductions on creative thinking. *Pers. Individ. Differ.* 51, 904–909. doi: 10.1016/j.paid.2011.07.015

Fredrickson, B. L. (1998). What good are positive emotions? *Rev. Gen. Psychol.* 2, 300–319. doi: 10.1037/1089-2680.2.3.300

George, J. M., and Zhou, J. (2002). Understanding when bad moods foster creativity and good ones don't: the role of context and clarity of feelings. *J. Appl. Psychol.* 87, 687–697. doi: 10.1037/0021-9010.87.4.687

Glenberg, A. M., Witt, J. K., and Metcalfe, J. (2013). From the revolution to embodiment: 25 years of cognitive psychology. *Perspect. Psychol. Sci.* 8, 573–585. doi: 10.1177/1745691613498098

Greenwald, A. G., Nosek, B., and Banaji, M. (2003). Understanding and using the implicit association test: I. an improved scoring algorithm. *J. Pers. Soc. Psychol.* 85, 197–216. doi: 10.1037/0022-3514.85.2.197

Guilford, J. P. (1967). *The Nature of Human Intelligence*. New York, NY: McGraw-Hill.

Gyurak, A., Gross, J. J., and Etkin, A. (2011). Explicit and implicit emotion regulation: a dual-process framework. *Cogn. Emot.* 25, 400–412. doi: 10.1080/02699931.2010.544160

Hao, N., Xue, H., Yuan, H., Wang, Q., and Runco, M. A. (2017). Enhancing creativity: proper body posture meets proper emotion. *Acta Psychol.* 173, 32–40. doi: 10.1016/j.actpsy.2016.12.005

Huang, L. (2019). Mind–body dissonance: a catalyst to creativity. *Pers. Soc. Psychol. Bull.* 46, 754–768. doi: 10.1177/0146167219875145

Huang, L., and Galinsky, A. D. (2011). Mind-body dissonance: conflict between the senses expands the mind's horizons. *Soc. Psychol. Pers. Sci.* 2, 351–359. doi: 10.1177/1948550610391677

Isen, A. M., Daubman, K. A., and Nowicki, G. P. (1986). Positive affect facilitates creative problem solving. *J. Pers. Soc. Psychol.* 52, 1122–1131. doi: 10.1037/0022-3514.52.6.1122

Jin, X., Deng, G., Jing, M., and Lin, G. (2009). The evaluation of emotion elicitation effect of videos. *Psychol. Explor.* 29, 83–87.

Kihlstrom, J. F., Mulvaney, S., Tobias, B. A., and Tobis, I. P. (2000). "The emotional unconscious," in *Cognition and Emotion*, eds E. Eich, J. F. Kihlstrom, G. H. Bower, J. P. Forgas, and P. M. Niedenthal (Oxford: Oxford University Press), 30–86.

Koole, S. L., and Rothermund, K. (2011). "I feel better but i don't know why": the psychology of implicit emotion regulation. *Cogn. Emot.* 25, 389–399. doi: 10.1080/02699931.2010.550505

Koole, S. L., Webb, T. L., and Sheeran, P. L. (2015). Implicit emotion regulation: feeling better without knowing why. *Curr. Opin. Psychol.* 3, 6–10. doi: 10.1016/j.copsyc.2014.12.027

Leung, K. Y. A., Liou, S., Qiu, L., Kwan, L. Y.-Y., Chiu, C.-Y., and Yong, J. C. (2014). The role of instrumental emotion regulation in the emotions-creativity link: how worries render neurotic individuals more creative. *Emotion* 14, 845–856.

Li, F., Zhu, Z., and Bai, X. (2009). The duration of happiness and sadness induced by emotional film editing. *Stud. Psychol. Behav.* 7, 32–38.

Martin, L. L., and Stoner, P. (1996). "Mood as input: what we think about how we feel determines how we think," in *Striving and Feeling: Interactions Among Goals, Affect, and Self-Regulation*, eds L. L. Martin and A. Tesser (Mahwah, NJ: Erlbaum), 279–301.

Mastria, S., Agnoli, S., and Corazza, G. E. (2019). How does emotion influence the creativity evaluation of exogenous alternative ideas? *PLoS One* 14:e0219298. doi: 10.1371/journal.pone.0219298

Neal, D. T., and Chartrand, T. L. (2011). Embodied emotion perception: amplifying and dampening facial feedback modulates emotion perception accuracy. *Soc. Psychol. Pers. Sci.* 2, 673–678. doi: 10.1177/1948550611406138

Niedenthal, P. M. (2007). Embodying emotion. *Science* 316, 1002–1005.

Nijstad, B. A., De Dreu, C. K. W., Rietzschel, E., and Baas, M. (2010). The dual pathway to creativity model: creative ideation as a function of flexibility and persistence. *Eur. Rev. Soc. Psychol.* 21, 34–77. doi: 10.1080/10463281003765323

Quirin, M., Kazén, M., and Rohrmann, S. (2009). Implicit but not explicit affectivity predicts circadian and reactive cortisol: using the implicit positive and negative affect test. *J. Pers.* 77, 401–426. doi: 10.1111/j.1467-6494.2008.00552.x

Riskind, J. H. (1982). Feedback effects of physical posture changes on depression and motivation. *Motiv. Emot.* 6, 273–298. doi: 10.1007/bf00992249

Riskind, J. H. (1984). They stoop to conquer: guiding and self-regulatory functions of physical posture after success and failure. *J. Pers. Soc. Psychol.* 47, 479–493. doi: 10.1037/0022-3514.47.3.479

Runco, M. A., and Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creat. Res. J.* 24, 66–75. doi: 10.1080/10400419.2012.652929

Runco, M. A., and Mraz, W. (1992). Scoring divergent thinking tests using total ideational output and a creativity index. *Educ. Psychol. Meas.* 52, 213–221. doi: 10.1177/001316449205200126

Schwarz, N. (2002). "Situated cognition and the wisdom in feelings: cognitive tuning," in *Emotions and Social Behavior. The Wisdom in Feeling: Psychological Processes in Emotional Intelligence*, eds L. F. Barrett and P. Salovey (New York, NY: The Guilford Press), 144–166.

Stanciu, M. M. (2015). Embodied creativity: a critical analysis of an underdeveloped subject. *Procedia Soc. Behav. Sci.* 187, 312–317. doi: 10.1016/j.sbspro.2015.03.058

Strack, F., and Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Pers. Soc. Psychol. Rev.* 8, 220–247. doi: 10.1207/s15327957pspr0803_1

Van Tilburg, W. A. P., and Igou, E. R. (2012). On boredom: lack of challenge and meaning as distinct boredom experiences. *Motiv. Emot.* 36, 181–194. doi: 10.1007/s11031-011-9234-9

Veenstra, L., Schneider, I. K., and Koole, S. L. (2017). Embodied mood regulation: the impact of body posture on mood recovery, negative thoughts, and mood-congruent recall. *Cogn. Emot.* 31, 1361–1376.

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* 54, 1063–1070. doi: 10.1037/0022-3514.54.6.1063

Wiers, R. W., and Stacy, A. W. (2006). Implicit cognition and addiction. *Curr. Dir. Psychol. Sci.* 15, 292–296.

Winkielman, P., and Berridge, K. C. (2004). Unconscious emotion. *Curr. Dir. Psychol. Sci.* 13, 120–123.

Wiswede, D., Münte, T. F., Krämer, U. M., and Rüsseler, J. (2009). Embodied emotion modulates neural signature of performance monitoring. *PLoS One* 4:e5754. doi: 10.1371/journal.pone.0005754

Zhen-Zhen, W., Jiang, Q., and Qing-Lin, Z. (2008). Exploring the mechanism for prototype elicitation effect in insight. *Psychol. Dev. Educ.* 24, 31–35.

Zhen-Zhen, W., Jiang, Q., and Qing-Lin, Z. (2009). Exploring a new experimental paradigm in the brain research of insight. *J. Psychol. Sci.* 32, 122–125.

How Does the Valence of Wording Affect Features of a Scale? The Method Effects in the Undergraduate Learning Burnout Scale

Biao Zeng, Hongbo Wen* and Junjie Zhang

Collaborative Innovation Center of Assessment Toward Basic Education Quality, Beijing Normal University, Beijing, China

A combination of both positively and negatively worded items is often employed in a survey to reduce participants' acquiescence bias, but such a combination may hurt the validity of the survey. The current study investigated the effect of valence of wording on participants' (N = 1132) responses to four versions of the Undergraduate Learning Burnout (ULB) scale. The results showed that the valence of wording affected a number of features of the scale. The internal consistency of both the original and the original-reverse versions (consisted of both positively and negatively worded items) was lower than that of the positive-only and the negative-only versions. The original and the original-reverse versions also had more factors than the positive-only and the negative-only versions. The original and the original-reverse versions showed method effects from both the positively and the negatively worded items, and those from the negatively worded items were stronger than those from the positively worded items. The method effects were predicted by participants' subjective well-being and future academic career plans. Together, this study suggests that using a combination of positively and negatively worded items can lead to a predictable response style and significant method effects, which reduce the scale's internal consistency and change the factor structure of the scale.

Keywords: learning burnout, valence of wording, method effect, parallel analysis, multitrait-multimethod model

INTRODUCTION

Likert scales are often used in developing surveys in the fields of psychology and education. These scales are consisted of a series of statements related to the target traits, evaluating the respondents' attitudes, opinions, evaluations, and intentions regarding a specific object or event. However, people often show response bias or response set when responding to these scales. Response bias refers to a systematic tendency to respond to a questionnaire on some basis other than the specific item content (i.e., what the items were designed to measure; Cronbach, 1946; Paulhus, 1991). Often times, when responding to a statement on a Likert scale, people tend to take into account only one side of the statement while ignoring the other side. For example, one category of response bias is acquiescence bias, which refers to people's tendency to select positively worded items rather than the negatively worded ones (Watson, 1992). In order to reduce the influence of response bias, researchers can use a combination of positively and negatively worded items in developing

questionnaires. The negatively worded items can introduce cognitive "bumps" to the participants, which increases the chance for them to complete the survey more carefully (Podsakoff et al., 2003) and in turn increases both the scale's accuracy in measuring the target characteristics (Anderson and Gerbing, 1988) and its measurement efficacy (Worcester and Burns, 1975; Chang, 1995).

An important assumption of using such a combination in formulating a scale is that both the positively and the negatively worded items should measure the same constructs (Marsh, 1996). However, research has shown that correlations between items within a scale that has both positively and negatively worded items are weaker than those within a scale that has only positively worded items (DiStefano and Motl, 2006; Carlson et al., 2011) and that using a combination of positively and negatively worded items reduces the internal consistency of a scale (Lee et al., 2008).

Using such a combination may also change the factor structure of the scale by bringing in unrelated factors to the target traits that the scale intends to measure (Corwyn, 2000; Vautier and Pohl, 2009; John et al., 2019), thereby violating the assumed unidimensional structure of the scale (Cordery and Sevastos, 1993; Jost and Thompson, 2000; Hevey et al., 2010). For example, Marsh et al. (2010) conducted confirmatory factor analysis (CFA) on the Rosenberg Self-Esteem Scale (RSES) and found that this unidimensional construct ended up having two additional factors, corresponding to the positively and negatively worded items, respectively, due to its mixed valence of wording. Some studies showed that in addition to the comprehensive self-esteem factor, either the positively worded items (Wang et al., 2001) or the negatively worded items (DiStefano and Motl, 2006) could bring a method factor. Other studies suggest that both the positively and the negatively worded items can induce method factors (Lindwall et al., 2012).

Such a change in the internal consistency and factor structure of a scale that has a mixed valence of wording can be attributed to the method effects. The method effects refer to tendencies to answer questions in survey-based criteria unrelated to the content being measured, which causes irrelevant systematic variance (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999; Lindwall et al., 2012). Furthermore, both positively and negatively worded items can introduce significant method effects (Bolin and Dodder, 1990; Wang et al., 2001). Some studies demonstrate that the negatively worded items cause stronger method effects compared to positively worded items (Marsh, 1996; Quilty et al., 2006; DiStefano and Motl, 2009), whereas others show that it is the positively worded items that lead to the stronger method effects (Farh and Cheng, 1997; Lindwall et al., 2012). The differential direction in which valence of wording affects the strength of the method effects can be attributed to cultural differences. For example, the phenomenon that the positively worded items induce stronger method effects is particularly prevalent in China where people value modesty. As confirmed by Farh and Cheng (1997), Chinese people consistently underestimated their performance when responding to the RSES, resulting in stronger method effects being associated more with the positively worded items.

In order to explore the method effects on a scale's validity, the multitrait-multimethod (MTMM) model in CFA is often employed to evaluate the discriminant and convergent validities (Campbell and Fiske, 1959). The MTMM model mainly includes two subtypes of specific models, the Correlated Trait-Correlated Method (CT-CM) and the Correlated Trait-Correlated Uniqueness (CT-CU). The CT-CM model treats method effects as a latent variable and separates the method variance from the characteristic variance. It is therefore able to measure the method effects directly from the method variance (Widaman, 1985). The CT-CU model, on the other hand, is proposed to solve the problem of low identification rate in the CT-CM model (Marsh, 1989). The CT-CU model classifies the method effects as residuals; however, this model does not assume independent method factors, which makes it unable to estimate the strength of the method effects directly or to examine the relationship between the method factors and external variables. For this reason, the CT-CM model is often used preferentially over other methods in estimating the method effects (Tomás and Oliver, 1999; Conway et al., 2004) unless it does not converge in practical applications (Lance et al., 2002; Wang, 2014). However, a drawback of the CT-CM model is that it allows for correlations among different method factors, which is not inevitable in actuality (Wang, 2014). In this case, the use of a more restricted Correlation Trait-Uncorrelated Method (CT-UM) model is more suitable (Widaman, 1985). Furthermore, many studies have found that the CT-UM model can better fit the data than other traditional models (Reise et al., 2007; Cao and Gu, 2010; Martel et al., 2010). Therefore, we opted for the CT-UM model to examine the method effects in this study.

Using a new research paradigm that can determine whether the latent factors of method effects are related to other constructs or variables, DiStefano and Motl (2009) confirm that the construct or variable of interest are related to the method effects, suggesting that the method effects are not simply a systematic error, but rather a stable style of responses related to the characteristics of the participants. In Quilty et al. (2006), participants who were emotionally stable were more likely to disagree with the negatively worded items, and participants with high avoidance motivation were more likely to agree with these items. Lindwall et al. (2012) found that participants with higher life satisfaction were more likely to agree with the positively worded items, whereas those with higher levels of depression were more likely to disagree with these items. In addition, participants' current emotional valence can also predict the method effects. For example, in Brosan et al. (2011), participants with higher levels of anxiety experienced bias in information processing and were more susceptible to agree with the negatively worded items, resulting in a biased pattern in responses.

There are a number of limitations to previous research that discuss the effect of using a combination of positively and negatively worded items. First, previous research on the effects of valence of wording focuses almost exclusively on a single valence, yet the effect of using different combinations of both positively and negatively worded items remains underexplored. Second, there are mixed results of whether the valence of wording introduces significant method effects and of which

valence introduces stronger method effects. Third, most research that investigated the method effects focused on the field of personality traits, but rarely tackled issues related to learning burnout. Learning burnout refers to students' negative attitudes toward learning and dismissive behavior due to a lack of interest in learning or chronic tress related to learning (Lian et al., 2005). Learning burnout is an important construct to study due to its prevalence and adverse consequences, such as reduced enthusiasm and commitment to learning (Zhang et al., 2009), poor academic performance (Gao, 2013; Wang, 2019), lowered self-esteem (Shi and Tan, 2008), and an overall worse subjective well-being (Shan et al., 2010).

The goal of the current study is to investigate the effect of different valence of wording on the outcomes of the Undergraduate Learning Burnout (ULB) scale so as to estimate the parameters of the scale accurately and provide references for measuring the learning burnout in college students. We mainly focused on how valence of wording affected the internal consistency and factor structure of the ULB, whether there would be significant method effects, and which statements would be associated with stronger method effects. We also examined if students' subjective well-being and future academic career plans would predict the method effects. We hypothesized that a combination of positively and negatively worded items would reduce the internal consistency, produce additional factors that are irrelevant to the variables of interest, and introduce significant method effects, with the negatively worded items inducing stronger method effects than the positively worded ones do. We also predicted that such method effects could be predicted by students' subjective well-being and future academic career plans.

MATERIALS AND METHODS

Participants and Materials

A total of 1132 students (368 males and 764 females) participated in this study. Among all participants, 414 were freshmen, 243 were sophomores, 66 were juniors, 170 were seniors, and 239 were graduate students.

The ULB scale used in this study was adopted by Lian et al. (2005) from the Marlach Burnout Inventory (Maslach and Jackson, 1981). The ULB scale has a high internal consistency as indicated by its overall α coefficient of 0.87. Lian et al. (2005) original ULB scale consisted of 20 items on five-point Likert scales: "1 = strongly disagree," "2 = somewhat disagree," "3 = uncertain," "4 = somewhat agree," and "5 = strongly disagree," where higher scores indicated greater learning burnout. Importantly, in this study, we eliminated the "uncertain" option because research has shown that such an option may be chosen due to a number of factors irrelevant to the target traits of the participants (such as participants not understanding the questions or not having a clear viewpoint), inducing measurement error and threatening the scale's reliability and validity (Raaijmakers et al., 2000; Kulas et al., 2008).

In addition to using the original ULB scale developed by Lian et al. (2005), we have also created three new versions of the scale by changing the valence of wording. The four versions

of the ULB scale are therefore included: the original version (12 positively worded items and 8 negatively worded items), the original-reverse version (where we changed all positively worded items in the original version to negatively worded and vice versa), the positive version (20 positively worded items), and the negative version (20 negatively worded items). Additionally, a demographic survey and a two-item questionnaire querying participants' subjective well-being and future academic career plans (whether they expected to pursue a master's or doctoral degree) on four-point Likert scales similar to the ULB scale were also administered in all four conditions.

Procedure

Participants completed the questionnaires online via a platform similar to the Amazon Mechanical Turk. Participants were allowed to choose one of the four versions of the ULB: 316 participants completed the original version, 267 did the original-reverse version, 288 did the positive version, and 261 did the negative version.

Data Analysis

Data from 36 participants were excluded for outlying scores, which were defined as being more than 2.5 standard deviations away from the mean of total response scores (Liu, 2019). The final sample consisted of 1,096 participants (306 in original condition, 258 in original-reverse condition, 277 in positive condition, and 255 in negative condition).

We conducted a parallel analysis of scale dimensions on each of the four versions by comparing actual data points with 50 simulated random data points. If the variation explained by the factors in the actual data was even smaller than that in the simulated random data, then the factors were discarded (Reise et al., 2000). After determining the number of scale dimensions, we conducted exploratory factor analysis (EFA) on each of the four versions by building two models—which contained one factor and two factors, respectively—in Mplus 8.3. Because the scoring scales used in this study were four-point Likert scales, we have chosen the variation adjusted weighted least squares (WLSMV) estimation method to analyze our data. Research has shown that when the score range is no wider than four points, it is not appropriate to conduct analyses using the maximum likelihood estimation method because this method cannot give an accurate estimation of the parameter or the standard error (Finney and Distefano, 2006). In contrast, the WLSMV is designed specially to deal with category changes, and it performs better than other estimation methods when dealing with categorical data (Beauducel and Herzberg, 2006; Brown, 2006; Finney and Distefano, 2006). In addition, the WLSMV method can also obtain an accurate estimation of the parameter in both skewed and small sample data (Flora and Curran, 2004; Beauducel and Herzberg, 2006). The final EFA results were obtained after factor rotation via the GEOMIN oblique rotation method, which was to account for correlations and cross-loadings among factors.

CFA was also performed in Mplus 8.3 using the variation adjusted weighted least squares estimation method. A total of

five models were established by adopting a method of the fixed variance of 1 (**Figure 1**). Two models, Model 1 and Model 2, were set up based on the results from the EFA. Three additional models were also set up for the original version and the original-reverse version, which might have method effects due to containing both positively and negatively worded items. Among these models, Model 1 contained only one substantive general learning burnout (GLB) factor. Model 2 contained two substantive factors of learning burnout, corresponding to positively and negatively worded items, respectively. Models 3, 4, and 5 adopted the CT-UM model to analyze the method effects. Model 3 and Model 4 both included a substantive learning burnout factor and a method factor. The method factor in Model 3 was obtained from all the positively worded items, and the method factor in Model 4 was obtained from all the negatively worded items. Model 5 contained one learning burnout factor and two method factors that were obtained from the positively and the negatively worded items, respectively.

Finally, a path analysis of the method factors was conducted in Mplus 8.3, using subjective well-being and academic career plans as the predictive variables and the latent method factor of the scale as the outcome variable, to explore whether the method effects could be predicted participants' subjective well-being and academic career plans.

RESULTS

Descriptive Statistics

Table 1 presents the mean response scores on each item in each of the four versions of the ULB scale. Scores on each item of the four versions were approximately normally distributed, as both the absolute values of the kurtosis and the skewness of all items were within 1. However, results from a Henze-Zirkler multivariate normality test suggested that none of the scores on all four versions showed normal distribution ($p < 0.001$;
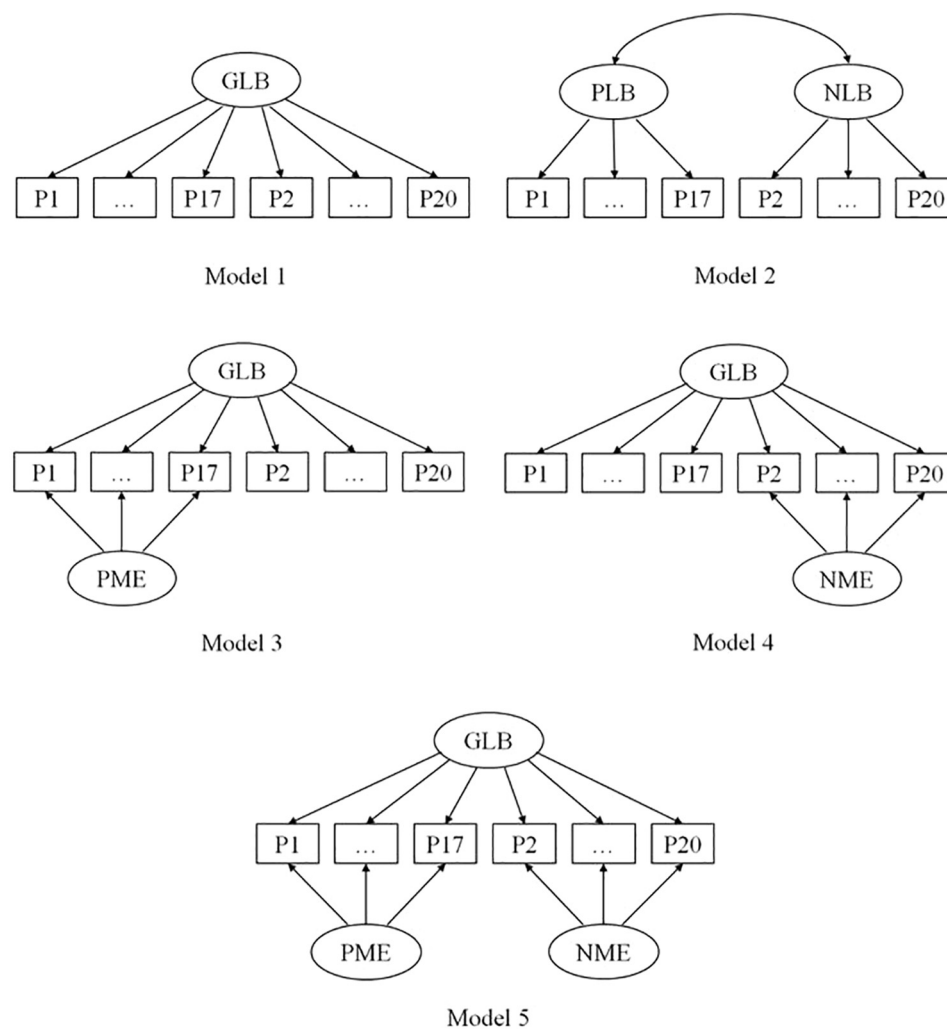
FIGURE 1 | The five factor structure models of the original version of the ULB scale. GLB, general learning burnout factor; PLB, positive learning burnout factor; NLB, negative learning burnout factor; PME, positive method effect; NME, negative method effect.

TABLE 1 | Descriptive statistics for the four versions of the ULB scale.

	Original		Original-Reverse		Positive		Negative	
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Item 1	2.37	0.73	2.24	0.76	2.30	0.59	2.13	0.74
Item 2	1.74	0.78	2.02	0.74	2.00	0.63	1.58	0.64
Item 3	2.45	0.70	2.28	0.69	2.49	0.61	2.16	0.72
Item 4	2.32	0.85	2.45	0.75	2.53	0.66	2.19	0.81
Item 5	2.51	0.77	2.41	0.69	2.47	0.61	2.43	0.78
Item 6	2.34	0.73	2.36	0.72	2.23	0.68	2.24	0.80
Item 7	2.60	0.79	2.83	0.71	2.71	0.59	2.48	0.77
Item 8	2.57	0.74	2.62	0.69	2.55	0.64	2.36	0.80
Item 9	2.12	0.72	2.51	0.66	2.57	0.62	2.07	0.72
Item 10	2.33	0.79	2.40	0.72	2.45	0.71	2.30	0.78
Item 11	2.25	0.67	2.04	0.70	2.28	0.60	1.92	0.66
Item 12	2.18	0.81	2.41	0.78	2.34	0.72	2.52	0.81
Item 13	2.25	0.78	2.10	0.81	2.18	0.64	1.92	0.74
Item 14	2.78	0.68	2.30	0.62	2.28	0.58	2.58	0.81
Item 15	2.25	0.74	1.77	0.77	2.28	0.65	1.77	0.69
Item 16	2.29	0.82	2.38	0.71	2.47	0.67	2.16	0.78
Item 17	2.41	0.66	2.36	0.64	2.47	0.62	2.29	0.72
Item 18	2.51	0.73	2.24	0.65	2.39	0.60	2.36	0.79
Item 19	2.37	0.78	2.24	0.66	2.34	0.61	2.27	0.78
Item 20	2.59	0.78	2.74	0.76	2.60	0.66	2.45	0.84

Original, original version of the ULB scale; Original-Reverse, original-reverse version of the ULB scale; Positive, positive version of the ULB scale; Negative, negative version of the ULB scale.

Korkmaz et al., 2014). The mean of score across all items in the positive version ($M = 2.40$, $SD = 0.63$) was the highest, followed by the original version ($M = 2.35$, $SD = 0.75$) and the original-reverse version ($M = 2.34$, $SD = 0.71$), and the negative version ($M = 2.21$, $SD = 0.76$) had the lowest scores. A one-way ANOVA showed that the difference in mean scores by version was significant, $F(3,1092) = 12.22$, $p < 0.001$. A Bonferroni *post hoc* test revealed that the mean score on the negative version was significantly lower than that on the original version ($M_{\text{diff}} = -3.03$, $p < 0.001$), the original-reverse version ($M_{\text{diff}} = -2.51$, $p < 0.01$), and the positive version ($M_{\text{diff}} = -3.76$, $p < 0.01$), whereas the mean scores on the latter three versions did not differ from each other.

Internal Consistency Analysis

The internal consistency of both the original (Cronbach's alpha = 0.79) and the original-reserve (Cronbach's alpha = 0.81) versions was lower than that of both the positive (Cronbach's alpha = 0.91) and the negative versions (Cronbach's alpha = 0.91).

Exploratory Factor Analysis

Results from the Kaiser–Meyer–Olkin Sampling Adequacy Measure and Bartlett's spherical test have confirmed that our data were appropriate for factor analysis. Results from a parallel analysis showed that both the original and the original-reverse versions had more factors than the positive and negative versions (**Figure 2**). Specifically, the original and original-reverse versions had two factors because two eigenvalues were much higher than the average eigenvalues of the random matrix, whereas

the positive and the negative versions each had only one factor because only one eigenvalue was much higher than the average eigenvalue of the random matrix for each version. Although in the negative version, the eigenvalue of the second factor was also above the average eigenvalue curve of the random matrix, the difference was neglectable as it might be due to sampling or other random errors.

We then created two models, one with one factor (Model 1) and the other with two factors (Model 2), based on the results from the parallel analysis and conducted EFA using these models on each version of the ULB scale. **Tables 2**, **3** present the factor loading of each of the four versions that we obtained. Under Model 1, the factor loading of the positive ($M = 0.66$) and the negative ($M = 0.63$) versions was higher than that of the original ($M = 0.35$) and the original-reverse versions ($M = 0.45$). Furthermore, the factor loading of most items in the positive and the negative versions was above 0.50, indicating that these two versions were in line with the one-factor structure; on the other hand, the factor loading of many items in the original and the original-reverse versions was lower than 0.30, with that of some even being negative values, suggesting that these versions might not fit into the one-factor structure. Under Model 2, the correlation between the two factors was much weaker in the original ($r = -0.08$, $p > 0.05$) and the original-reverse versions ($r = 0.13$, $p > 0.05$) than in the positive ($r = 0.64$, $p < 0.05$) and the negative versions ($r = 0.71$, $p < 0.05$). In addition, the positive and the negative versions showed cross-factor loading in many items, indicating that these two versions were well-suited for the one-factor structure. In contrast, most items in the original and the original-reverse versions had a higher loading on one factor and a lower loading on the other one, with only a few items showing cross-factor loading, suggesting that these two versions were better suited for the two-factor structure rather than the one-factor structure.

Confirmatory Factor Analysis

We conduct CFA using both Model 1 and Model 2 on the EFA results from each version of the ULB (**Table 4**). The comparative fit index (CFI) value for the ideal model fit should be no less than 0.90 (Hu and Bentler, 1999). The root mean square error of approximation (RMSEA) value should be no greater than 0.08 (Browne and Cudeck, 1992), with the highest acceptable range being no greater than 0.10 (Steiger, 1990). Under Model 1, the CFI values of the original and the original-reverse versions were 0.27 and 0.64, and the RMSEA values were 0.19 and 0.14, respectively, all of which indicated that these versions did not fit into the one-factor model and were therefore not composed of a unidimensional structure. On the other hand, the CFI values of the positive and negative versions were 0.93 and 0.93, and the RMSEA values were 0.09 and 0.10, respectively. Although the RMSEA values of these two version scales still did not reach the ideal level, they were in the acceptable range, suggesting that these two versions fitted in the one-factor model as they had only one learning burnout factor. Under Model 2, the positive and negative versions, the CFI values were 0.95 and 0.96, and the RMSEA values were 0.08 and 0.07, respectively. However, as these two versions fitted in the one-factor model

FIGURE 2 | The solid lines represent scree plot obtained from the actual data. The dashed lines are scree plot averaged across 50 data points in a simulated dataset generated from a population, where all variables are uncorrelated, with the same number of participants and items as the original dataset.
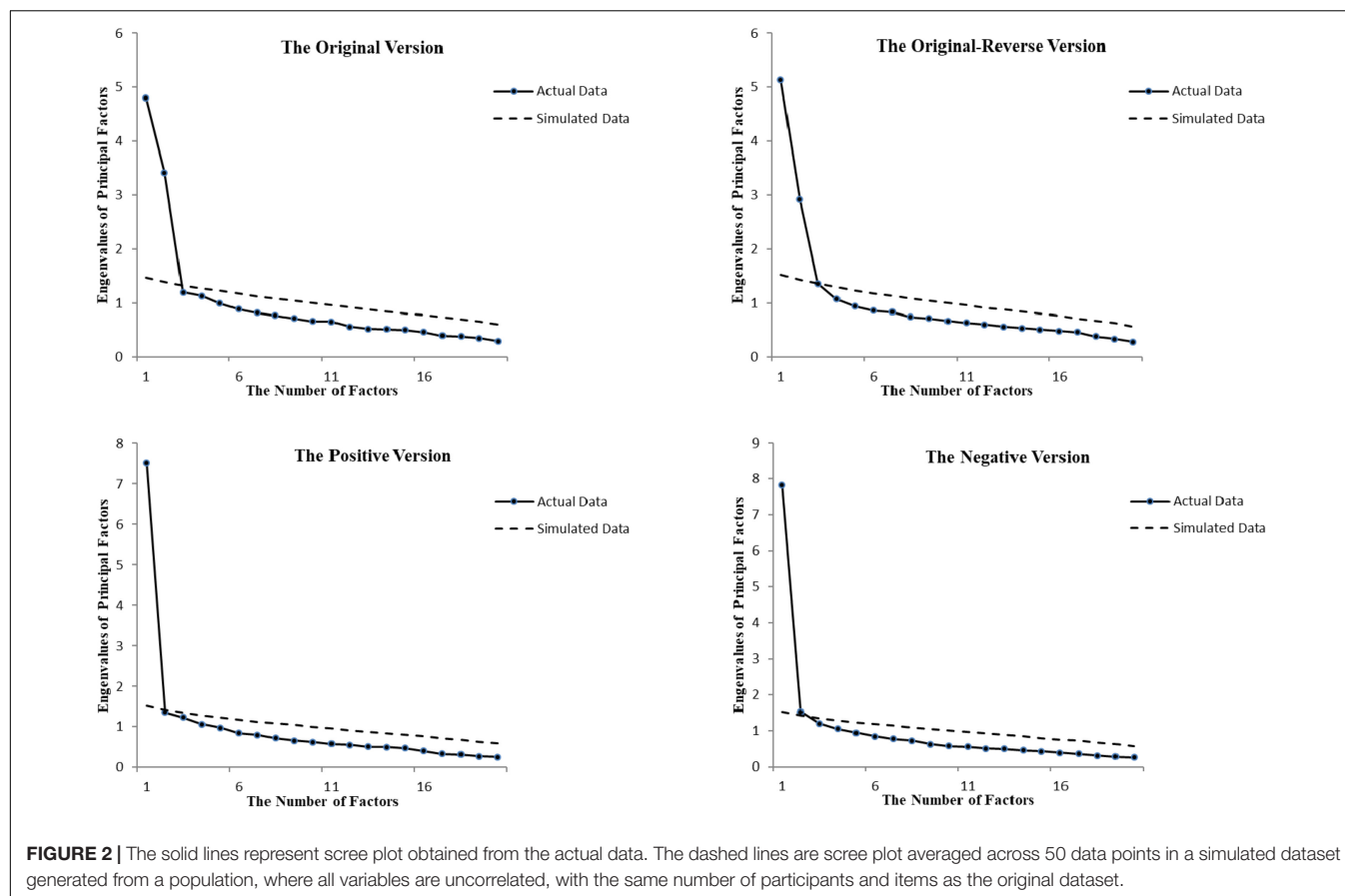
TABLE 2 | Standardized factor loadings for the four versions of the ULB scale in Model 1.

	Original	Original-Reverse	Positive	Negative
Item 1	0.05	0.28*	0.62*	0.63*
Item 2	0.46*	0.54*	0.50*	0.47*
Item 3	−0.14*	0.16*	0.68*	0.71*
Item 4	0.66*	0.74*	0.64*	0.67*
Item 5	0.77*	0.83*	0.81*	0.82*
Item 6	−0.12*	−0.01*	0.57*	0.63*
Item 7	0.50*	0.52	0.67*	0.61*
Item 8	−0.14*	−0.11	0.53*	0.63*
Item 9	0.68*	0.71*	0.63*	0.77*
Item 10	0.61*	0.66*	0.70*	0.68*
Item 11	−0.11	0.20*	0.69*	0.69*
Item 12	0.54*	0.52*	0.54*	−0.01
Item 13	0.01	0.17*	0.65*	0.65*
Item 14	0.60*	0.76*	0.79*	0.65*
Item 15	−0.13*	0.16*	0.58*	0.65*
Item 16	0.67	0.67*	0.71*	0.65*
Item 17	−0.01	0.40*	0.82*	0.65*
Item 18	0.75*	0.74*	0.81*	0.65*
Item 19	0.71*	0.73*	0.69*	0.65*
Item 20	0.69*	0.41*	0.58*	0.66*

*$p < 0.05$.

ideally, dividing their factor structure into two dimensions would not have much practical significance. As for the original and original-reverse versions, the CFI values were 0.86 and 0.84, and the RMSEA values were 0.08 and 0.09, respectively. The level of model fitting improved greatly but still has not reached the ideal fit. These results suggest that these two versions are not a simple combination of two learning burnout factors, but rather, it is possible that multi-dimensional structure of the original and the original-reverse versions is a result of having both one learning burnout factor plus one or two other factors induced by method effects.

To test the aforementioned idea, we created Models 3, 4, and 5 by adding one positive, one negative, or both method factors to Model 1. The CFA results of the original and the original-reverse versions based on these three models are shown in **Table 5**. For the original version, the level of fitting in Model 3 was slightly worse than that in Model 2, with a CFI value of 0.86 and an RMSEA value of 0.08. The level of fitting of the original-reverse version in Model 3 was better than that in Model 2, with a CFI value of 0.85 and an RMSEA value of 0.09. The level of fitting in Model 4 for both versions was better than that in both Model 2 and Model 3, with the CFI values being 0.89 for the original version and 0.89 for the original-reverse version, and the RMSEA values being 0.08 and 0.08, which were close to the ideal fit. Model 5 was the most ideal among all models created. The CFI values were 0.90 and 0.92, and RMSEA values were 0.08 and 0.07, for the two versions, respectively. These results suggest versions of the ULB that contained both positively and

TABLE 3 | Standardized factor loadings for the four versions of the ULB scale in Model 2.

	Original		Original-Reverse		Positive		Negative	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
Item 1	0.17	**0.67***	0.19	**0.50***	0.15	**0.55***	0.00	**0.71***
Item 2	**0.48***	0.12	**0.53***	0.06	**0.47***	0.07	0.00	**0.54***
Item 3	−0.03	**0.67***	−0.01	**0.68***	**0.58***	0.16	**0.50***	0.27
Item 4	**0.66***	0.02	**0.74***	0.02	**0.65***	0.03	**0.76***	−0.07
Item 5	**0.78***	0.04	**0.83***	0.01	*0.56**	*0.34*	**0.68***	0.19
Item 6	−0.03	**0.53***	−0.10	**0.40***	**0.43***	0.20	**0.73***	−0.08
Item 7	**0.47***	−0.24	**0.58***	−0.27	**0.74***	−0.03	**0.88***	−0.29
Item 8	−0.05	**0.62***	−0.21	**0.39***	**0.42***	0.16	**0.67***	−0.02
Item 9	**0.71***	0.12	**0.72***	−0.07	**0.76***	−0.10	**0.71***	0.10
Item 10	**0.63***	0.12	**0.67***	−0.06	−0.10	**0.89***	0.04	**0.73***
Item 11	0.00	**0.68***	0.03	**0.69***	*0.32*	*0.46**	0.27	**0.50***
Item 12	**0.53***	−0.04	**0.55***	−0.12	**0.32***	0.28	0.23	**−0.27***
Item 13	0.10	**0.57***	0.04	**0.55***	**0.44***	0.28	*0.34*	*0.37**
Item 14	**0.59***	−0.05	**0.76***	0.01	*0.51**	*0.37*	**0.37***	0.02
Item 15	−0.01	**0.66***	−0.03	**0.77***	0.14	**0.51***	0.14	**0.43**
Item 16	**0.67***	0.00	**0.68***	−0.05	0.07	**0.72***	−0.04	**0.75***
Item 17	0.11	**0.69***	*0.31*	**0.54***	**0.90***	−0.03	**0.74***	0.08
Item 18	**0.75***	−0.01	**0.74***	0.01	**0.72***	0.15	**0.71***	0.13
Item 19	**0.70***	−0.05	**0.71***	0.07	0.00	**0.78***	0.04	**0.81***
Item 20	**0.67***	−0.14	**0.42***	−0.04	*0.30*	**0.35***	**0.61***	0.09

The higher factor loadings for each item within one version of the ULB were expressed in bold, and the factor with cross-loadings (both factor loadings being greater than 0.30) were italicized.

negatively worded items had one learning burnout factor and two additional method factors for the positively and negatively worded items, respectively.

In order to further explore the method effects in the original and original-reverse versions, we analyzed the loading of the method factors in Model 5 (**Table 6**). The results showed that the loading of the negative method factor (original: $M = 0.62$; original-reverse: $M = 0.56$) was higher than that of the positive one (original: $M = 0.01$; original-reverse: $M = 0.03$) for both versions, with the loading of all negatively worded items reaching significance ($p < 0.05$) except for one insignificant case. As for the positively worded items, many of those items in the original version bared a factor loading of a negative value,

TABLE 4 | Model fit indices for in the four versions of the ULB scale in Model 1 and Model 2.

	χ^2	df	CFI	RMSEA [90% CI]
Model 1				
Original	1954.91	170	0.27	0.19 [0.18, 0.19]
Original-Reverse	1030.15	170	0.64	0.14 [0.13, 0.15]
Positive	580.44	170	0.93	0.09 [0.09, 0.10]
Negative	559.48	170	0.93	0.10 [0.09, 0.10]
Model 2				
Original	504.41	169	0.86	0.08 [0.07, 0.09]
Original-Reverse	545.67	169	0.84	0.09 [0.08, 0.10]
Positive	467.57	151	0.95	0.09 [0.08, 0.10]
Negative	364.53	151	0.96	0.07 [0.07, 0.08]

TABLE 5 | Model fit indices for the original and the original-reverse versions of the ULB scale in Models 3, 4, and 5.

	χ^2	df	CFI	RMSEA [90% CI]
Original				
Model 3	507.31	162	0.86	0.08 [0.08, 0.09]
Model 4	435.96	158	0.89	0.08 [0.07, 0.08]
Model 5	406.50	150	0.90	0.08 [0.07, 0.08]
Original-Reverse				
Model 3	510.95	158	0.85	0.09 [0.08, 0.10]
Model 4	434.67	162	0.89	0.08 [0.07, 0.09]
Model 5	349.83	150	0.92	0.07 [0.06, 0.08]

and only one item's factor loading was significant ($p < 0.05$). These results suggest that versions of the ULB scales that contained both positively and negatively worded items had more method effects from the negatively worded items than from the positively worded items.

Correlates of Method Effects

We conducted path analysis to explore whether some other variables, such as the subjective well-being and future academic career plans, can predict the method effects. We took these two variables as the predictor variables and the two potential method factors of the original version of the ULB (as outlined in Model 5) as the outcome variables. As presented in **Table 7**, we found a strong negative correlation between future academic career plans and the learning burnout factor, $r = −0.62$, $p < 0.001$,

TABLE 6 | Standardized factor loadings for each item of the original and the original-reverse versions of the ULB scale in Model 5.

	Original		Original-Reverse	
	PME	NME	PME	NME
Item 2		0.47*	0.28*	
Item 4		0.66*	0.45*	
Item 5		0.77*	0.35*	
Item 7		0.49*	0.54*	
Item 9		0.70*	0.47*	
Item 10		0.63*	0.00	
Item 12		0.53*	0.26*	
Item 14		0.60*	0.40*	
Item 16		0.67*	0.06	
Item 18		0.75*	0.30*	
Item 19		0.70*	-0.08	
Item 20		0.68*	0.52*	
Item 1	−0.22			0.47*
Item 3	0.08			0.69*
Item 6	−0.15			0.40
Item 8	−0.23			0.40*
Item 11	0.23			0.69*
Item 13	−0.07			0.55*
Item 15	0.55*			0.75*
Item 17	−0.11			0.57*

PME, method effect induced by positively worded items; NME, method effect induced by negatively worded items. *p < 0.05.

TABLE 7 | Correlations between latent factors of the original ULB scale and their two potential predictors (subjective well-being and future academic career plans).

	GLB	PME	NME
Subjective well-being	0.28	−0.86***	0.50***
Future academic career plans	−0.62**	0.65***	−0.78***

**p < 0.01, ***p < 0.001.

but there was no significant correlation between subjective well-being and learning burnout factor, $r = 0.28$, $p > 0.05$. However, there was a strong negative correlation between subjective well-being and the method factor induced by the positively worded items, $r = -0.86$, $p < 0.001$, and a strong positive correlation between subjective well-being and the method factor induced by the negatively worded items, $r = 0.50$, $p < 0.001$, suggesting that participants with higher subjective well-being were more likely to agree with the negatively worded items and to disagree with the positively worded items. On the contrary, there was a strong positive correlation between future academic career plans and the method factor induced by the positively worded items, $r = 0.65$, $p < 0.001$, and a strong negative correlation between future academic career plans and the method factor induced by the negatively worded items $r = -0.78$, $p < 0.001$, suggesting that participants with higher future academic career plans were more likely to agree with the positively worded items and to disagree with the negatively worded items. In addition, both subjective well-being and future academic career plans were negatively

correlated with learning burnout scores, $r = -0.25$, $p < 0.01$; $r = -0.21$, $p < 0.01$, suggesting that the higher one's subjective well-being or future academic career plans are, the less learning burnout they have.

DISCUSSION

In this study, we explored the impact of the valence of wording on the scale measurement through four versions of the ULB scale with different combinations of positively and negatively worded items. Using new statistical methods and modeling, we confirmed that a combination of both positively and negatively worded items undermines the internal consistency and changed the factor structure of the scale by introducing additional method factors. More specifically, we showed that both positively and negatively worded items can induce method effects in scales that contain both positively and negatively worded items and that such method effects can be predicted by participants' subjective well-being and future academic career plans.

Our results confirmed that mixed valence of wording (or using a combination of positively and negatively worded items) reduces the scale's internal consistency as a result of the method effects, which was consistent with the results in previous literature (Lee et al., 2008). Both versions with only unidirectional wording had greater internal consistency than both with a combination of positive and negative wording. Because the contents of the four versions were the same, differences in internal consistency can only be explained by the effect of the wording valance. We therefore suggest that the contents measured by different valences of wording are actually not completely consistent as the combination of different valences of wording introduces irrelevant variations to the scale (DiStefano and Motl, 2006; Lee et al., 2008; Carlson et al., 2011). Specifically speaking, the original and original-reverse version scale included both positively and negatively worded items, negatively worded items were more complicated than positively worded items, and the information processing methods required by the equivalent negatively and positively worded items were not inconsistent (Mayo et al., 2004), which made participants' responses to positively and negatively worded items different. As a result, the correlation between the scores of different items in the scale decreased, eventually causing a decrease in the internal consistency of the scales.

Our parallel analysis and EFA showed that a scale with a unidirectional valence of wording has a unidimensional structure, whereas one with a mixed valence of wording changes such a structure by bringing additional factors. These findings are consistent with previous literature, where two factors are present even if only one construct was measured when a scale employs mixed valence of wording (Conroy, 2004; DiStefano and Motl, 2009). A commonly agreed explanation for such a structural change is the method effects induced by mixed valences of wording. However, there is still controversy regarding which factors are responsible for the multidimensional structures and whether the positively and negatively worded items measure the same traits. Some researchers argue that positively and negatively worded items measure different traits (Joireman et al., 2008;

Boduszek et al., 2013), while others suggest that those items measure the same trait and that the multidimensional structures are only a result of method effects induced by different valences of wording. For example, Wang et al. (2010) and Cai (2017) suggest that the multidimensional structure of the RSES is composed of a self-esteem factor and a method factor. The results of the current study support the latter argument. Our data fit in models that contained method factor(s) better than in the model that did not; more importantly, the model that contained a learning burnout factor and two method factors (for positively and negatively worded items, respectively) fitted the data best, suggesting that the two-factor structure of the ULB scale is composed of a learning burnout factor and two method factors.

However, our findings that the ULB scale contained only one substantive learning burnout factor is inconsistent with the three-factor structure that was originally assumed. It is likely due to the use of different criteria in determining the number of factors in the EFA. Based on Kaiser's rule, which supports the three-factor structure, whether a factor will be retained is determined by whether the eigenvalues of the actual data are greater than one (Henson and Robert, 2006). Although Kaiser's rule was one of the most commonly used criteria, it tends to retain too many factors (Zwick and Velicer, 1986). In contrast, the method of parallel analysis use in the current study is considered the most precise method for retaining factors and is better than many other methods, including Kaiser's rule (Bartlett, 1950; Humphreys and Montanelli, 1975; Zwick and Velicer, 1986; Silverstein, 1987). The parallel analysis determines whether a factor is retained based on comparisons with the real data by comparing the eigenvalue against the average eigenvalue of the random matrix. Our findings indicate that the traditional Kaiser's rule may overestimate the number of factors, whereas using parallel analysis can avoid this problem and give a more accurate estimate of how many factors a scale has.

A further analysis of loading on the method factors in the model that contains a learning burnout factor and two method factors (Model 5) for the two mixed-valence versions shows that the method effect induced by the negatively worded items is stronger than that by the positively worded items. These findings are consistent with many previous studies (Marsh, 1996; Quilty et al., 2006; DiStefano and Motl, 2009). We propose three possible explanations for this phenomenon. The first is related to the negatively worded items themselves: The negative expression of a statement usually requires more complex cognitive processing and therefore pose difficulties in semantic understanding, which, in turn, affects the participants' answers and leads to a greater measurement bias (Podsakoff et al., 2003). This explanation is supported by the finding that the mean score of the negative version was lower compared to the other three versions, and such a tendency was present in only the negative version but not in the positive version (as scores on the positive version did not differ from those of the two combination versions), suggesting that participants' responses are more likely to be affected by negatively worded items than by positively worded items. The second explanation is related to the age and traits of the participants: previous research has shown that young people are more susceptible to negative

information than older people (Carstensen and Mikels, 2005). Participants in the current study are all college students; they are therefore more likely to agree to the negatively worded items, which might in turn induce a stronger negative method effect. The third explanation is related to cultural differences. Culture plays a role in the strength of method effects. For example, Chinese people, living in a collectivist culture and valuing modesty, usually report a lower level of self-esteem in an attempt to maintain social approval, resulting in a stronger method effect in positively worded items (Farh and Cheng, 1997). However, this phenomenon is domain-specific, and it is more likely to present when assessing positive traits (such as self-esteem) or self-evaluating one's own performance (Schmitt and Allik, 2005). Namely, not all traits will show this tendency, with learning burnout being one of the cases. Due to China's strong competitive culture and huge employment pressure, Chinese students generally have a higher level of learning burnout than students in other countries (Jacobs and Dodd, 2003; Sun, 2019). This is the opposite of the case of self-esteem, and therefore, there is a stronger method effect in the negatively worded items.

The tendency for participants with high subjective well-being to agree with the negatively worded items and disagree with the positively worded items and the reversed tendency for participants with high academic expectations confirms that the method effects represent a stable response style and can be predicted by positive emotional traits (DiStefano and Motl, 2009). Yet, our findings on the directions of the correlations are somewhat surprising. Previous studies have shown that people with positive emotions are more likely to agree with positively worded items, while those with negative emotions are more likely to recognize negatively worded items (Podsakoff et al., 2003; Lindwall et al., 2012). In the current study, however, only the academic expectation measure is consistent with this tendency. The seemingly counterintuitive findings regarding subjective well-being, which is also a positive emotion, may be explained by the desensitization theory. Desensitization refers to the phenomenon in which individuals' cognitive, emotional, physiological, and behavioral responses to a stimulus are reduced or eliminated over time due to repeated exposure (Funk, 2005). Individuals with higher subjective well-being are more likely to experience or pay attention to positive stimuli in general, which makes their cognitive and emotional responses to such stimuli gradually decline over time; yet, stimuli containing negative words are relatively rare or less attended to, leading to higher sensitivity to these stimuli. This phenomenon therefore still supports the nature of method effects as a stable response style of participants.

We should take measures to reduce the method effect on the measurement as much as possible. This can be achieved through two methods. The first is program control, or changing the valence of wording. More specifically, we can convert unipolar items into bipolar items. In other words, we can change a unidirectional item into one that contains a pair of antonyms (Schweizer and Schreiner, 2010). For instance, the statement that "mastery of professional knowledge is easy for me" can be changed to "I feel mastery of professional knowledge" with

two options—"easy" and "difficult"—for participants to choose from. The second method is statistical control. As demonstrated by the current study, we can use MTMM models (such as the CT-UM model) to test for method effects. Such models are especially helpful when the target scale contains both positively and negatively worded items. If the target scale has significant method effects, then the method factors should be included in the model when performing CFA so that the parameters of the scale can be corrected, and the actual parameters can be accurately estimated.

One limitation of the current study is the lack of random assignment—participants were free to choose to answer one of the four versions of the ULB scale. However, note that the participants were blind to the conditions that they were asked to choose from (they were given four numbers to choose from without any other labeling), nor did they know how the four versions differed until they were debriefed after completing the study. Another limitation is that the sample composition was not well-balanced as our sample consisted of predominantly females (65%), which may hurt the generalizability of our results to the entire population. Future research should conduct large-scale sampling and exercise more control over the assignment of conditions to obtain even more convincing and reliable results. In addition, even after the negatively worded items in the original scale were converted to positively worded items (for the original-reverse version), some of these items still had a significant factor loading on method effects, which may be due to some random factors unrelated to the measurement. This study was not designed to separate these unrelated factors; future research can analyze the potential factors involved in this phenomenon. Finally, the current study employs an anchor test as it used the four different versions of the ULB scale; another future research direction is to conduct an anchor test analysis to further investigate method effects.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

BZ contributed to developing the research design, collecting data, conducting data analyses, and writing up the research in Chinese and revising the translated manuscript. HW, similarly, contributed to guiding the research design and revising the manuscript both before and after it was translated into English. JZ contributed to revising and improving the write-up. All authors contributed to the article and approved the submitted version.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Anderson, J. C., and Gerbing, D. W. (1988). Structural equation modeling in practice: a review and recommended two-step approach. *Psychol. Bull.* 103, 411–423. doi: 10.1037/0033-2909.103.3.411

Bartlett, M. S. (1950). Tests of significance in factor analysis. *Br. J. Psychol.* 3, 77–85. doi: 10.1111/j.2044-8317.1950.tb00285.x

Beauducel, A., and Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Struct. Equ. Model.* 13, 186–203. doi: 10.1207/s15328007sem1302_2

Boduszek, D., Hyland, P., Dhingra, K., and Mallett, J. (2013). The factor structure and composite reliability of the rosenberg self-esteem scale among

ex-prisoners. *Pers. Individ. Differ.* 55, 877–881. doi: 10.1016/j.paid.2013.07.014

Bolin, B. L., and Dodder, R. A. (1990). The affect balance scale in an american college population. *J. Soc. Psychol.* 130, 839–840. doi: 10.1080/00224545.1990.9924639

Brosan, L., Hoppitt, L., Shelfer, L., Sillence, A., and Mackintosh, B. (2011). Cognitive bias modification for attention and interpretation reduces trait and state anxiety in anxious patients referred to an out-patient service: results from a pilot study. *J. Behav. Ther. Exper. Psychiatr.* 42, 258–264. doi: 10.1016/j.jbtep.2010.12.006

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford Press.

Browne, M. W., and Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociol. Methods Res.* 21, 230–258. doi: 10.1177/0049124192021002005

Cai, M. (2017). *An Examination of the Wording Effect in Rosenberg Self-Esteem Scale*. M. A. thesis, Shandong Normal University, Shandong.

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56, 81–105. doi: 10.1037/h0046016

Cao, Y., and Gu, Q. (2010). The development of the PISA's chinese reading test and dimensionality assessment. *Exam. Res.* 6, 80–92.

Carlson, M., Wilcox, R., Chou, C.-P., Chang, M., Yang, F., Blanchard, J., et al. (2011). Psychometric properties of reverse-scored items on the CES-D in a sample of ethnically diverse older adults. *Psychol. Assess.* 23, 558–562. doi: 10.1037/a0022484

Carstensen, L. L., and Mikels, J. A. (2005). At the intersection of emotion and cognition: aging and the positivity effect. *Curr. Direct. Psychol. Sci.* 14, 117–121. doi: 10.1111/j.0963-7214.2005.00348.x

Chang, L. (1995). Connotatively consistent and reversed connotatively inconsistent items are not fully equivalent: generalizability study. *Educ. Psychol. Measur.* 55, 991–997. doi: 10.1177/0013164495055006007

Conroy, D. E. (2004). The unique psychological meanings of multidimensional fears of failing. *J. Sport Exerc. Psychol.* 26, 484–491. doi: 10.1123/jsep.26.3.484

Conway, J. M., Lievens, F., Scullen, S. E., and Lance, C. E. (2004). Bias in the correlated uniqueness model for MTMM data. *Struct. Equ. Model.* 11, 535–559. doi: 10.1207/s15328007sem1104_3

Cordery, J. L., and Sevastos, P. P. (1993). Responses to the original and revised Job diagnostic survey: is education a factor in responses to negatively worded items? *J. Appl. Psychol.* 78, 141–143. doi: 10.1037/0021-9010.78.1.141

Corwyn, R. F. (2000). The factor structure of global self-esteem among adolescents and adults. *J. Res. Pers.* 34, 357–379. doi: 10.1006/jrpe.2000.2291

Cronbach, L. J. (1946). Response sets and test validity. *Educ. Psychol. Measur.* 6, 475–494. doi: 10.1177/001316444600600405

DiStefano, C., and Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Struct. Equ. Model.* 13, 440–464. doi: 10.1207/s15328007sem1303_6

DiStefano, C., and Motl, R. W. (2009). Personality correlates of method effects due to negatively worded items on the rosenberg self-esteem scale. *Pers. Individ. Differ.* 46, 309–313. doi: 10.1016/j.paid.2008.10.020

Farh, J.-L., and Cheng, B.-S. (1997). Modesty bias in self-rating in Taiwan: impact of item wording, modesty value, and self-esteem. *Chin. J. Psychol.* 39, 103–118.

Finney, S. J., and Distefano, C. (2006). "Nonnormal and categorical data in structural equation modeling," in *Structural Equation Modeling: A Second Course*, eds G. R. Hancock and R. O. Mueller (Greenwich, CT: IAP), 269–314.

Flora, D. B., and Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychol. Methods* 9, 466–491. doi: 10.1037/1082-989x.9.4.466

Funk, J. B. (2005). Children's exposure to violent video games and desensitization to violence. *Child Adolesc. Psychiatr. Clin. N. Am.* 14, 387–404. doi: 10.1016/j.chc.2005.02.009

Gao, B. (2013). The types of college students' Learning Burnout and their effect on learning. *Chin. J. Spec. Educa.* 84–89.

Henson, D. R., and Robert, J. K. (2006). Use of exploratory factor analysis in published research common errors and some comment on improved practice. *Educ. Psychol. Measur.* 66, 393–416. doi: 10.1177/0013164405282485

Hevey, D., Pertl, M., Thomas, K., Maher, L., Craig, A., and Ni Chuinneagain, S. (2010). Consideration of future consequences scale: confirmatory factor analysis. *Pers. Individ. Differ.* 48, 654–657. doi: 10.1016/j.paid.2010.01.006

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Humphreys, L. G., and Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multiv. Behav. Res.* 10, 193–205. doi: 10.1207/s15327906mbr1002_5

Jacobs, S. R., and Dodd, D. K. (2003). Student burnout as a function of personality, social support, and workload. *J. Coll. Stud. Dev.* 44, 291–303. doi: 10.1353/csd.2003.0028

John, T., Kulas, O., Rachael, K., and Lindsey, K. (2019). Confound it!: social desirability and the "reverse-scoring" method effect. *Eur. J. Psychol. Assess.* 35, 855–867. doi: 10.1027/1015-5759/a000459

Joireman, J., Balliet, D., Sprott, D., Spangenberg, E., and Schultz, J. (2008). Consideration of future consequences, ego-depletion, and self-control: support for distinguishing between CFC-Immediate and CFC-Future sub-scales. *Pers. Individ. Differ.* 45, 15–21. doi: 10.1016/j.paid.2008.02.011

Jost, J. T., and Thompson, E. P. (2000). Group-based dominance and opposition to equality as independent predictors of self-esteem, ethnocentrism, and social policy attitudes among African Americans and European Americans. *J. Exper. Soc. Psychol.* 36, 209–232. doi: 10.1006/jesp.1999.1403

Korkmaz, S., Goksuluk, D., and Zararsiz, G. (2014). MVN: an R package for assessing multivariate normality. *R J.* 6, 151–162. doi: 10.32614/rj-2014-031

Kulas, J. T., Stachowski, A. A., and Haynes, B. A. (2008). Middle response functioning in Likert-responses to personality items. *J. Bus. Psychol.* 22, 251–259. doi: 10.1007/s10869-008-9064-2

Lance, C. E., Noble, C. L., and Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychol. Methods* 3, 186–207. doi: 10.1037//1082-989X.7.2.228

Lee, P.-H., Chang, L.-I., and Ravens-Sieberer, U. (2008). Psychometric evaluation of the Taiwanese version of the Kiddo-KINDL generic children's health-related quality of life instrument. *Q. Life Res.* 17, 603–611. doi: 10.1007/s11136-008-9328-3

Lian, R., Yang, L. X., and Wu, L. H. (2005). Relationship between professional commitment and learning burnout of undergraduates and scales developing. *Acta Psychol. Sin.* 37, 632–636. doi: 10.1111/j.1744-7909.2005.00136.x

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., et al. (2012). Method effects: the problem with negatively versus positively keyed items. *J. Pers. Assess.* 94, 196–204. doi: 10.1080/00223891.2011.645936

Liu, H. (2019). *Advanced Psychological Statistics*. Beijing: Renmin University of China Press.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: many problems and a few solutions. *Appl. Psychol. Measur.* 13, 335–361. doi: 10.1177/014662168901300402

Marsh, H. W. (1996). Positive and negative global self-esteem: a substantively meaningful distinction or artifactors? *J. Pers. Soc. Psychol.* 70, 810–819. doi: 10.1037/0022-3514.70.4.810

Marsh, H. W., Scalas, L. F., and Nagengast, B. (2010). Longitudinal tests of competing factor structures for the rosenberg self-esteem scale: traits, ephemeral artifacts, and stable response styles. *Psychol. Assess.* 22, 366–381. doi: 10.1037/a0019225

Martel, M. M., Gremillion, M., Roberts, B., Von Eye, A., and Nigg, J. T. (2010). The structure of childhood disruptive behaviors. *Psychol. Assess.* 22, 816–826. doi: 10.1037/a0020975

Maslach, C., and Jackson, S. E. (1981). The measurement of experienced burnout. *J. Organ. Behav.* 2, 99–113. doi: 10.1002/job.4030020205

Mayo, R., Schul, Y., and Burnstein, E. (2004). "I am not guilty" vs "I am innocent": successful negation may depend on the schema used for its encoding. *J. Exper. Soc. Psychol.* 40, 433–449. doi: 10.1016/j.jesp.2003.07.008

Mutheìn, L. K., and Mutheìn, B. O. (1998). *Mplus User's Guide*, 8th Edn, Los Angeles, CA: Mutheìn & Mutheìn.

Paulhus, D. L. (1991). "Measurement and control of response bias," in *Measures of Personality and Social Psychological Attitudes*, Vol. 1, eds J. P. Robinson, P. R. Shaver, and L. S. Wrightsman (San Diego, CA: Academic Press), 17–59. doi: 10.1016/b978-0-12-590241-0.50006-x

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *J. Appl. Psychol.* 88, 879–903. doi: 10.1037/0021-9010.88.5.879

Quilty, L. C., Oakman, J. M., and Risko, E. (2006). Correlates of the rosenberg self-esteem scale method effects. *Struct. Equ. Model.* 13, 99–117. doi: 10.1207/s15328007sem1301_5

Raaijmakers, Q. A. W., van Hoof, A., Hart, H., Verbogt, T. F. M. A., and Vollebergh, W. A. M. (2000). Adolescents' midpoint responses on Likert-type scale items: neutral or missing values? *Intern. J. Public Opin. Res.* 12, 208–216. doi: 10.1093/ijpor/12.2.209

Reise, S. P., Morizot, J., and Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Q. Life Res.* 16, 19–31. doi: 10.1007/s11136-007-9183-7

Reise, S. P., Waller, N. G., and Comrey, A. L. (2000). Factor analysis and scale revision. *Psychol. Assess.* 12, 287–297. doi: 10.1037/1040-3590.12.3.287

Schmitt, D. P., and Allik, J. (2005). Simultaneous administration of the rosenberg self-esteem scale in 53 nations: exploring the universal and culture-specific

features of global self-esteem. *J. Pers. Soc. Psychol.* 89, 623–642. doi: 10.1037/0022-3514.89.4.623

Schweizer, K., and Schreiner, M. (2010). Avoiding the effect of item wording by means of bipolar instead of unipolar items: an application to social optimism. *Eur. J. Pers.* 24, 137–150. doi: 10.1002/per.748

Shan, C., Wang, Y., and Liu, X. (2010). Research on the relationship between learning burnout and subjective well-being of college students. *Chin. J. Health Psychol.* 18, 951–954.

Shi, J., and Tan, Y. (2008). A comparatiue study on the relationships of primary and middle school students logical reasoning ability, metacognition, and the level of attention academic achievement among. *Psychol. Sci.* 31, 736–737.

Silverstein, A. B. (1987). Note on the parallel analysis criterion for determining the number of common factors or principal components. *Psychol. Rep.* 61, 351–354. doi: 10.2466/pr0.1987.61.2.351

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multiv. Behav. Res.* 25, 173–180. doi: 10.1207/s15327906mbr2502_4

Sun, X. (2019). *A Survey on the Current Situation and Causes of Academic Burnout of Postgraduates.* Master's thesis, Zhejiang University, Hangzhou.

Tomás, J. M., and Oliver, A. (1999). Rosenberg's self-esteem scale: two factors or method effects. *Struct. Equ. Model.* 6, 84–98. doi: 10.1080/10705519909540120

Vautier, S., and Pohl, S. (2009). Do balanced scales assess bipolar constructs? The case of the STAI scales. *Psychol. Assess.* 21, 187–193. doi: 10.1037/a0015312

Wang, J., Siegal, H. A., Falck, R. S., and Carlson, R. G. (2001). Factorial structure of Rosenberg's self-esteem scale among crack-cocaine drug users. *Struct. Equ. Model.* 8, 275–286. doi: 10.1207/S15328007SEM0802_6

Wang, M., Cai, B., Wu, Y., and Dai, X. (2010). The factor structure of Chinese Rosenberg Self-esteem Scale affected by item statement method. *Psychol. Explor.* 30, 64–69.

Wang, M. C. (2014). *Latent Variable Modeling and Mplus Application (Foundation).* Chongqing: Chongqing University Press.

Wang, Y. (2019). The relationship among non-English majors' autonomous learning ability, learning burnout and english grades: a structural equation modeling study. *Chin. J. Appl. Linguist.* 42, 79–91. doi: 10.1515/CJAL-2019-0005

Watson, D. (1992). Correcting for acquiescent response bias in the absence of a balanced scale: an application to class consciousness. *Sociol. Methods Res.* 21, 52–88. doi: 10.1177/0049124192021001003

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Appl. Psychol. Measur.* 9, 1–26. doi: 10.1177/014662168500900101

Worcester, R. M., and Burns, T. R. (1975). Statistical examination of relative precision of verbal scales. *J. Market Res. Soc.* 17, 181–197.

Zhang, Z., Chen, Z., Qiao, F., and Ni, A. (2009). The relationship between undergraduates' coping efficacy,Learning burnout and learning engagement. *Chin. J. Health Psychol.* 17, 28–30.

Zwick, W. R., and Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychol. Bull.* 99, 432–442. doi: 10.1037/0033-2909.99.3.432

Lost in Learning: Hypertext Navigational Efficiency Measures Are Valid for Predicting Learning in Virtual Reality Educational Games

Chris Ferguson* and Herre van Oostendorp

Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands

The lostness measure, an implicit and unobtrusive measure originally designed for assessing the usability of hypertext systems, could be useful in Virtual Reality (VR) games where players need to find information to complete a task. VR locomotion systems with node-based movement mimic actions for exploration and browsing found in hypertext systems. For that reason, hypertext usability measures, such as "lostness" can be used to identify how disoriented a player is when completing tasks in an educational game by examining steps made by the player. An evaluation of two different lostness measures, global and local lostness, based on two different types of tasks, is described in a VR educational game using 13 college students between 14 and 18 years old in a first study and extended using 12 extra participants in a second study. Multiple Linear Regression analyses showed, in both studies, that local lostness, and not global lostness, had a significant effect on a post-game knowledge test. Therefore, we argued that local lostness was able to predict how well-participants would perform on a post-game knowledge test indicating how well they learned from the game. In-game experience aspects (engagement, cognitive interest, and presence) were also evaluated and, interestingly, it was also found that participants learned less when they felt more present in the game. We believe these two measures relate to cognitive overload, which is known to have an adverse effect on learning. Further research should investigate the lostness measure for use in an online adaptive game system and design the game system in such a way that the risk of cognitive overload is minimized when learning, resulting in higher retention of information.

Keywords: VR game, game analytics, lostness measures, predictive validity, learning

1. INTRODUCTION

Within education, serious games are often used as learning tools. A key aim of many of these educational serious games is information acquisition, without explicit teaching, where items must be discovered to proceed further in the game. In particular, narrative-centered discovery games have considerable learning potential. In these games, players are transported to another place and time period and must explore the environment, completing tasks to discover items that reveal the in-game story (Malone and Lepper, 1987; Lester et al., 2014). This stimulates players to construct appropriate mental models (Wasserman and Banks, 2017; Furlough and Gillan, 2018).

When it comes to these types of educational games, in-game analytics is a major focus as it is beneficial to know how a player is performing during play. For example, these analytics can help to characterize players when carrying out such information-searching tasks (e.g., as chaotic or goal-oriented) and can be instrumental in improving the game (e.g., where confusion arises). Most importantly, the analytics can be used to dynamically improve (adapt) the game based on the in-game performance of the player, which may lead to better learning (Van Oostendorp et al., 2014).

There are many different in-game analytics and analytic systems (see e.g., Drachen and Schubert, 2013; Westera et al., 2014; Loh et al., 2015). However, game analytics are often based on shallow interaction data (Perez-Colado et al., 2018) and their success in predicting learning can be limited or unknown. Often, data mining algorithms are used rather than a full analysis of the task that needs to be performed, along with the cognitive processes and cognitive problems that players encounter whilst carrying out these tasks. Consequently, this leads to a multitude of data being collected to develop analytics that are not aligned with the mental operations a player needs to carry out in building an adequate mental model to effectively complete a task (Furlough and Gillan, 2018). In contrast, we aim to sample characteristic difficulties in executing and orchestrating specific information processes needed to solve an (information search) task (Elshout, 1976). With this in mind, in this study game analytics will be developed based on a cognitive analysis of the search task.

Due to the perceived high engagement and assumed higher knowledge transfer of VR, there has been a transition from traditional PCs to VR serious games on different domains. These include evacuation training and hazard awareness (Feng et al., 2018). This is of particular interest as these two domains make use of spatial activities, particularly navigation, in the case of evacuation training. In a parallel study (Ferguson et al., 2020b) a VR educational game condition was compared to a traditional, non-VR desktop condition. The research question was whether or not VR had a positive effect on the retention of educational content and navigational efficiency. We define navigational efficiency as how efficiently a user completes a goal (i.e., finding information), in terms of taking the shortest path by visiting the minimum number of locations without revisiting the same locations multiple times (Smith, 1996). In this study, VR was found to have a significant positive benefit on spatial activities, in the form of increased retention of spatial knowledge, and on navigational efficiency as participants were better able to efficiently navigate the virtual environment. As these VR games for learning become more prevalent, analytics should be investigated for VR educational games that may be more appropriate than analytics already available and therefore we will examine whether we can use navigational efficiency as an implicit indicator of learning performance. Because we have to focus on the performance of players in the VR condition, we do not want to confound the VR data from that parallel study with non-VR data in correlational analyses, and we restrict ourselves to only using the VR data. Therefore, this paper will focus on only the VR participants from that parallel study (Ferguson

et al., 2020b) to investigate if possible implicit (unobtrusive) navigational efficiency measures are valid for predicting learning in certain types of VR educational games.

We present a study focused on the implicit and unobtrusive measuring of the navigation patterns of players during playing a Virtual Reality (VR) educational game. We assume that all interactions a player makes in a game are related to the learning process, and, thus, can form a useful and valid indicator of the learning process and, consequently, predict the final learning results. Our assumption is, therefore, that game analytics relevant to learning should be based on information processing characteristics indicative of learning, and thus correlate with learning. The predictive validity of these so-called cognition-based game analytics concerning learning will be examined in this study, at the same time considering other potentially relevant factors, such as game experience aspects, that is, how a user feels whilst playing the game.

2. THEORETICAL BACKGROUND

2.1. Navigation and the Issue of Disorientation

As mentioned above, many educational games require navigating around a virtual environment to complete tasks to acquire information as part of the learning process. Unfortunately, navigating through such games can be challenging and lead to disorientation, where players lose their sense of location and direction (Conklin, 1987; Head et al., 2000). This happens when navigation is too much of a cognitive burden and leads to cognitive overload (Gwizdka and Spence, 2007), i.e., an excessive amount of load placed on a person's working memory when carrying out a task (Chen, 2016).

Disorientation and cognitive overload when navigating large, complex systems is a well-known problem within hypertext systems like websites (Conklin, 1987; Edwards and Hardman, 1989; Gwizdka and Spence, 2007). Edwards and Hardman (1989) coined the term "Lost in Hyperspace" for this phenomenon. It is a key issue as cognitive overload leads to inhibited learning (Sweller et al., 2011) when a user simultaneously has to spend limited cognitive resources to both navigation and comprehension (Sharples, 1999; Bolter, 2000). Navigation is part of the learning process and navigation problems leading to disorientation have a negative effect on learning (Sweller et al., 2011). Disorientation can be measured by navigational efficiency, where a player who is inefficient during navigation (low navigational efficiency) can be classified as disoriented and an efficient player (with high navigational efficiency) can be classified as not disoriented at all. Therefore, to provide analytics for learning, a metric must be identified that can measure the navigational efficiency of a player also within a VR game. We argue that navigation in VR systems with node-based movements can be compared with navigation as occurring in hypertext systems, and, for that reason, a navigation efficiency measure with validated success in hypertext environments can also be useful in a VR environment

Node-based movement systems with fast movement speeds are used in VR due to its effectiveness in reducing the likelihood

of motion sickness (Habgood et al., 2018). This system of movement has the additional effect of restricting players to only visiting set locations, allowing developers to guide players, assisting them with navigation, rather than them being able to roam freely around the environment as in traditional first-person games (see **Figure 1**). When using this kind of movement systems, the locations a player can visit are stored in a spatial node graph, a link-node model where nodes are shown as being linked together (see **Figure 2**). A player can move directly to another node only if the target node and the current node are linked. Otherwise, they must navigate through other nodes, following a path to the target node similar to navigating a hypertext system.

User navigation behavior in the context of web information search tasks consists of sequentially moving through a set of pages, deciding at each step (each hyperlink) where to go. It can be divided into two main stages of cognitive processing: attention cycle and action-selection cycle (Kitajima et al., 2000; Karanam et al., 2016; Van Oostendorp et al., 2017). The attention cycle is further divided into two stages: parsing the web page in high-level schematic regions and focusing on one of these regions. The action-selection cycle is also divided into two stages: comprehension of screen objects, e.g., hypertext links, and selecting the most appropriate screen object after a (semantic) evaluation. The hyperlink with the highest similarity to the goal is selected. This evaluation and selection process is repeated for every new screen page until the user reaches the target page.

Because of the basic similarities between a spatial node graph and a web graph with its node-link structure and their use, as well as the fact that both systems feature rich information spaces and encourage exploration, it can be argued that, when carrying out information-searching tasks, actions and their structure in hypertext systems are very similar to actions carried out in VR educational games that use a node-based locomotion system. In **Table 1**, we summarized the set of typical actions and processes in hypertext and VR systems. As indicated in the table, many actions and assumed cognitive processes are similar, just like the characteristic difficulties these processes might evoke. These involve but are not limited to: focusing on the wrong location, an overload of working memory, failures in the activation of information from long-term memory (Karanam and Van Oostendorp, 2020). In addition to this, similar to hypertext systems, in VR educational games, users have to navigate large and complex information spaces, making many decisions, to carry out and solve information-searching tasks.

For both node-based VR educational games and hypertext, navigation is a key issue in a learning process where information acquisition is key. Users suffering from disorientation and cognitive overload will have trouble completing information-searching tasks, leading to impaired learning performance. Therefore, based on the above analysis of the cognitive task, measures successfully designed for assessing the usability of hypertext systems (hypertext usability measures) can perhaps be used in these types of VR games using node-based movement and mimic exploration and browsing found in hypertext systems. An attractive feature of the availability of such measures is that they could result in possible real-time monitoring values for subsequently adapting a game to a player's skill level, leading

to optimal learning (Shute et al., 2017). It is not our purpose to validate the task analysis above, but to provide evidence for the relevance of the navigation efficiency measures as cognition-based real-time game analytics.

2.2. Lostness

We pose that lostness (Smith, 1996), an unobtrusive navigational efficiency measure shown to be successful in predicting success in information-seeking tasks in hypertext, can also be used in VR games that use a node-link approach, due to the similarity in process and structure. This measure has already been shown to be effective in VR for generating navigational support as well as having a strong correlation with spatial ability (Van Oostendorp and Karanam, 2013), which is strongly linked with navigational efficiency.

The lostness measure (Smith, 1996) was created as a method for measuring hypertext usability in terms of task performance based on the belief of the author that "measures based on time and errors seem inappropriate for hypertext systems which, by their nature, encourage exploration, and browsing." Instead, a better approach is to assess task performance in terms of the efficiency in how users find information and the degree in which they become lost in the information search whilst looking for this information (Gwizdka and Spence, 2007).

For information-searching tasks, lostness is defined by the total number of information items inspected compared with the minimum number of items that need to be inspected to locate the requested information. This leads to the following formula:

$$L = \sqrt{\left(\frac{N}{S} - 1\right)^2 + \left(\frac{R}{N} - 1\right)^2}, \tag{1}$$

where R is the *minimum* number of nodes needed to be visited, S is the *total* number of nodes a player has visited, and N is the number of *unique* nodes a player has visited. This will return a value between 0 and $\sqrt{2}$. 0 indicates that the task has been completed perfectly and the player was not disoriented (high navigational efficiency) and $\sqrt{2}$ indicates that a user was completely disoriented (low navigational efficiency) whilst completing the task. The S and N-values must be logged as the player is carrying out the task whilst the R-value must be manually specified or calculated using a path-finding algorithm. Two components are important in this formula: $(\frac{N}{S} - 1)$ indicating the degree of repetition to already visited nodes, and $(\frac{R}{N} - 1)$ indicating the degree of detours during navigation. This is a highly versatile measure that can be used in any activity that can be defined in a minimum number of steps (R) and the path that a user/player/person taken can be measured to provide S and N.

2.2.1. Global vs. Local Lostness

In a hypertext environment, one can distinguish information-searching tasks as either *gathering* tasks or *fact-finding* tasks. Gathering tasks are defined as information searching tasks where target information spread out over different areas in the virtual space must be located and combined, whereas, fact-finding tasks are defined as tasks where required information can be located in a single, specific place (Puerta Melguizo et al., 2012).

FIGURE 1 | The node-based movement system within *"the Chantry"*. Players can only move to locations shown by a footsteps icon and pick up items when shown a hand icon. Some objects have a label underneath, or on the back, representing a task.

Educational games teach content using in-game learning activities, tasks and objectives, which are based on real-life examples, and act as practice for real-life tasks (Lester et al., 2018). In a game, a player can be given a task to complete, such as finding three items that are related to a certain topic. Each of these items can be regarded as an objective. To complete a task, a player must achieve all of the objectives. As such, we regard tasks and objectives as representing these in-game learning activities, in which tasks have a larger scope, where "global" gathering activities are needed, as they consist of finding information that needs to be gathered together, while finding the separate objectives represent "local" fact-finding activities, which have a smaller scope.

Based on these two types of tasks, two corresponding versions of the lostness measure were defined: *global lostness*, which focuses on tasks, and *local lostness*, which focuses on objectives.

This allows information-searching tasks to be measured from both a "global" task-based perspective and a more "local" objective-based perspective.

Global lostness: Global lostness focuses on tasks, using the initial lostness equation to give a value of lostness from the start of each task until it is completed. A weighted lostness (L) mean based on the number of objectives per task (x), reflecting the complexity of each task (t), is then used to give a measure of lostness for the full game (L_G):

$$L_G = \frac{\sum_{t=1}^{n} (L_t \times x_t)}{\sum_{t=1}^{n} x_t}, \tag{2}$$

In this measure, the S and N values are logged from the moment a task is started up until it is completed. The R-value is often manually inputted based on data from a perfect playthrough
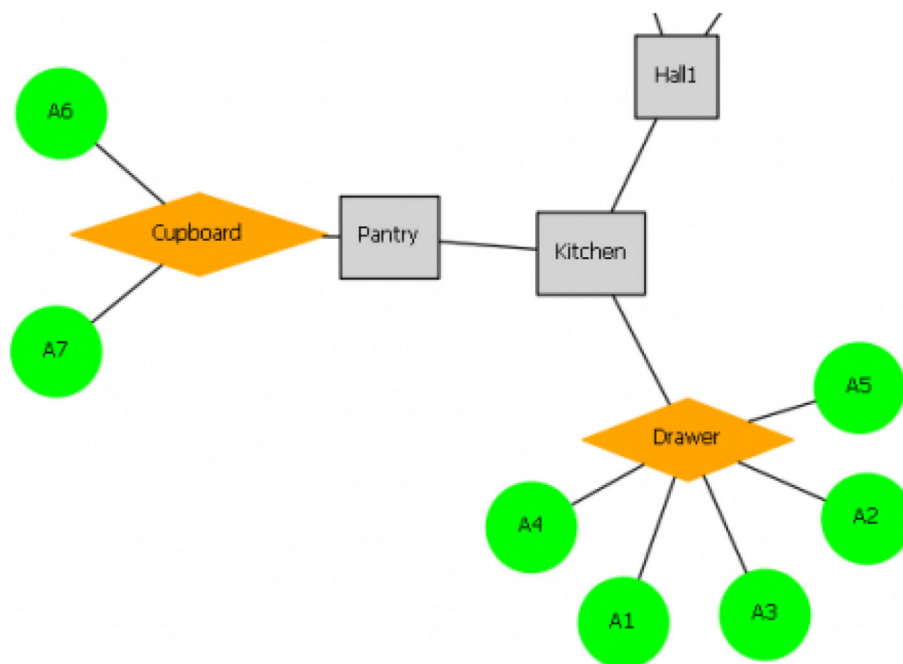
FIGURE 2 | An example of the spatial graph used in *"the Chantry"*. This represents the locations and objects that a player can access [fig:spatial]. Gray nodes represent overall locations (rooms) that a player can visit, orange nodes represent items a player can interact with to reveal objects, and green nodes represent objects that can be picked up and interacted with.

TABLE 1 | Similarities between navigation in hypertext and node-based VR environment.

Typical actions and processes in hypertext systems	Typical actions and processes in a node-based VR game
Parsing a page into top-level schematic regions	Parsing a visual scene into top-level schematic regions
Focus and attend to one of the schematic regions	Focus and attend a promising region
Comprehending and elaborating the screen objects (e.g., hyperlinks) within that region	Examining and elaborating items (e.g., icons and objects) within that region
Evaluating and selecting one of the screen objects	Evaluating and picking up an item
Moving on to another screen object or repeating preceding actions and processes	Moving to another node within the environment or repeating preceding actions and processes

of the game. To give a more accurate overview of the player performance over the full game, a weighted mean is used rather than an ordinary mean, as tasks with more objectives require more searching than shorter tasks to access the learning content, resulting in longer paths being followed. Consequently, lostness could be overstated by low lostness values from simpler tasks masking higher lostness values from more complicated tasks and vice-versa.

Local lostness: Local lostness focuses on objectives, regardless of the task that they belong to. As well as calculating a lostness value for each objective, using the initial lostness equation, the R, S, and N-values of each objective (O) are summed to give a measure of local lostness for the full game (L_L):

$$L_L = \sqrt{\left(\frac{\sum_{o=1}^{n} N_o}{\sum_{o=1}^{n} S_o} - 1\right)^2 + \left(\frac{\sum_{o=1}^{n} R_o}{\sum_{o=1}^{n} N_o} - 1\right)^2}, \quad (3)$$

In this measure, the S and N-values are always being logged from the start of the game and are reset to 0 after an objective has been completed. Because there are probably multiple orders in which objectives can be completed, a path-finding algorithm, such as a breadth-first search algorithm, is usually necessary to calculate the shortest path (R) between objectives.

Rather than only focusing on a limited set of objectives that belong to a task, local lostness considers all objectives in the game. This is because it is considered a mistake to make a game linear when it aims to tell a story and make the player part of it (Hudson, 2011). This also has the additional benefit of including objectives that act as task starting points, which also must be searched for and may contain educational content. As such, this measure encompasses all objectives, which must be completed. Therefore, local lostness can follow the complete path of the player through the game. The R, S, and N-values from each objective are fed into the original lostness formula to give a full game local lostness measure L_L, which reflects the full path taken by a player throughout the whole game, essentially treating the whole game as one long task.

See **Supplementary Materials** for more detailed implementations, algorithms, and examples of both lostness measures being used.

2.3. Potential Other Relevant Factors to Learning in VR

Although it is proposed that measuring disorientation/navigational efficiency is important when the goal of a game is information acquisition to meet learning goals, the type of motivational/emotional experience that a player has whilst playing these kinds of educational games should also be taken into account (Lee et al., 2010). This is backed up by research showing that cognitive and emotional aspects of player involvement with the game environment, particularly enjoyment and deep thinking, also positively influence information processing (Imlig-Iten and Petko, 2018).

Three elements of player experience are thought to be noteworthy when it comes to having a positive effect on learning in VR: (a) presence, i.e., the experience of being integrated into a mediated environment and the sense of being there (Slater, 2002), (b) engagement, i.e., heightened concentration, involvement, and enjoyment (Kim, 2018), and (c) cognitive interest i.e., understanding key topics presented by the game and becoming more interested (Harp and Mayer, 1997).

In many studies, engagement is seen as an important effect of games (and determinant of game playing) and recent research has concluded that both emotional and cognitive engagement catalyze learning in games (Boyle et al., 2012). Furthermore, presence is viewed as being of particular importance in VR games, a medium where increased presence is a defining feature (Steuer, 1992), and was found to aid learning outcomes (Lee et al., 2010). Finally, cognitive interest serves as intrinsic motivation to explore and experience new and unfamiliar things (Van der Sluis et al., 2014), a key part of VR educational games.

For all three aspects, we assume, based on empirical studies (Boyle et al., 2012; Abdul Jabbar and Felicia, 2015), that they are positively influenced by a (VR) game environment and are positively related to learning. Accordingly, we will measure these game experience aspects through three questionnaires that have proven validity: the I-group Presence Questionnaire (Schubert, 2003), the Game Engagement Questionnaire (Brockmyer et al., 2009), and the Perceived Interest Questionnaire (Schraw et al., 1995) to investigate the experience aspects of a player whilst playing the game.

3. RESEARCH QUESTIONS

As mentioned, navigation is key when it comes to information searching to achieve learning goals and disoriented players have trouble completing tasks and objectives, which we expect to have a negative effect on learning. In this regard, lostness is a process-oriented measure of what participants are learning. We expect that a higher performance in navigation should contribute to learning. In other words, if a player is efficient, they should learn better. Accordingly, we will apply and evaluate the lostness hypertext usability measure in an appropriate VR game to measure navigational efficiency and identify this disorientation. Two related measures, "global" and "local" lostness, stemming from the type of information-searching tasks identified by Puerta Melguizo et al. (2012), will be investigated to examine if they provide an online indication of how well players learn from a VR educational game.

Secondly, because we may assume that more factors than lostness will be active, such as the game experience aspects of presence, engagement, and cognitive interest, which have been linked to learning, we will also investigate whether the effect of these game experience aspects influence learning.

Therefore, we will compute the correlations of both lostness measures (global and local) with experience aspects (cognitive interest, engagement, and presence) as well as with learning by participants. This will allow us to examine the effect of both lostness measures as well as of these three additional player experience aspects *on* learning. This brings us to the following research questions: In a VR educational game, when it comes to learning, what is the effect of:

a. Global lostness and local lostness?
b. Experience aspects? (i.e., presence, engagement, cognitive interest)

Because a low lostness score denotes that the participant followed an efficient (perfect in the case of a value of zero) search path and a high score represents a chaotic, imperfect search, a lower lostness value indicates higher navigational efficiency and less disorientation. Consequently, we predict that:

I. There will be a negative correlation between learning and both global and local lostness.
II. There will a positive correlation between learning and the game experience aspects of engagement, presence, and cognitive interest.

The effect of the lostness measures on learning will be analyzed via three steps (using IBM *SPSS* Statistics 24). First, all measures will be investigated through their means and standard deviations (SDs). Next, we carry out a zero-order correlation analysis, which will show if these measures give any kind of indication of whether or not the measures are related to one another. Finally, the contribution of the lostness measures (global and local) and experience aspects (presence, cognitive interest, and engagement) to learning performance will be investigated in multiple linear regression analyses to give a more pure estimation of the effect of lostness on learning when the experience aspects are taken into account.

4. MATERIALS AND METHODS

4.1. Apparatus
4.1.1. Game
Sony PlayStation VR game "The Chantry," a VR educational game using a node-based movement system was used. This game takes place in the house of Dr. Edward Jenner and tells the story of the invention of the smallpox vaccine https://jennermuseum.com/.

To progress through games such as "The Chantry," players interact with closed doors and window shutters. Upon trying to open these, a word appears (e.g., "*Gloucestershire*" in **Figure 3**), which represents an educational story topic, and an audio narrative is then played, containing story information that the player needs to learn about. The player is then presented with a list of objectives e.g., "*County Map*" and "*Last Letter to Bristol*" in **Figure 3**) with the same at the top as on the door/shutter (e.g., "*Gloucestershire*" in **Figure 3**). Each objective in the list (e.g., "*County Map*" and "*Last Letter to Bristol*" in **Figure 3**) is a hint toward an item that must be found. The player must navigate the virtual environment and find the item that the objective is referring to and turn it over in their hands to reveal the same keyword that is written on the door. When one of these items is found, completing the objective, the items reveal additional story information (fact-finding task) in the form of an audio narrative. Achieving all of the objectives completes the task, opening the door/shutter, and another audio narrative plays, revealing additional story information, together forming an information-searching (gathering) task. See **Figure 4** for an example of a task ("*Gloucestershire*" as shown in **Figure 3**) being carried out in *the Chantry*.

In this kind of in-game learning activities, there is always an efficiency aspect, i.e., a shorter path or sequence of nodes that must be visited to complete an objective and, therefore, task. This makes the lostness measure suitable here. See **Supplementary Materials** for how both lostness measures are applied.

4.1.2. Hardware
Each participant wore a Sony PlayStation VR headset (model: CUH-ZVR1) to play the game on a Base PlayStation 4 Development Kit and, to realize maximum comfort, all participants wore their own over-ear headphones for audio. This accurately replicates how VR games are played with complete immersion in the VR environment without distraction, as the participants could only see and hear what was happening in the game. They also made use of the standard DualShock 4 controller (model: CUH-ZCT1). To control the game, participants used their head movements to look at a node and used a single button press on the controller to move to that node or pick up an item. Once picked up, an item is moved and rotated by doing the same action holding on the controller.

4.2. Measurements
Each interaction participants carried out, including jumping to another node, picking up and examining an item, and unlocking something, was captured. As this data was timestamped, it was possible to trace the path that has been taken and calculate the lostness measures.

Knowledge retention was measured through a bespoke knowledge test, consisting of 24 randomized true/false statements (50% true and 50% false). Sixteen of these questions concerned facts, relating to the story (e.g., "Vaccination was already popular in England by 1,800"), and 8 involved spatial aspects, relating to the location of items/rooms in the game (e.g., "The library was very close to the dining room and located on the first floor").

Standard questionnaires, consisting of five-point Likert scales, were used to measure engagement, presence, and cognitive interest. These were: the Game Engagement Questionnaire (Brockmyer et al., 2009) (example item—"I feel like I can't stop playing"), the igroup Presence Questionnaire (Schubert, 2003) (example item—"I felt present in the virtual space"), and the Perceived Interest Questionnaire (Schraw et al., 1995) (example item—"I thought the game's topic was fascinating"). Each of these questionnaires consists of 5-point Likert scales, with high Cronbach's coefficients-alpha, being respectively 0.85, 0.85, and 0.90.

4.3. Procedure
Upon being seated, participants were given health and safety information, which consisted of warnings about possible motion sickness, advice to withdraw if it occurs, and to remain seated and not attempt to physically interact with the virtual environment. both in oral and written form. They were also given instructions on how to play the game. Also, an informed consent form was given. They were asked if everything was clear to them. If not, an additional explanation was provided. After this, they signed the informed consent form. The participants were informed that they would be playing a VR educational game and would be tested after playing to investigate how much knowledge they were able to recall. They were instructed to divide their attention to the whole environment and not to stick to one location. They were then given 30 min to play through the game at their own pace, learning about the story, and completing tasks and objectives. The tasks and objectives in the game, varying in complexity, were the same for all participants and presented in the same order. To control the game, players used their head movements to look at a node and used a single button press on the controller to move to that node or pick up an item, which is moved and rotated by doing the same action whilst holding the controller. Once the participants had played the game for the full 30 min, they immediately completed the knowledge test and the game experience questionnaires. Finally, the participants were debriefed and informed about the nature of both the study and the game.

5. INITIAL STUDY

5.1. Participants
A total of 13 students aged 13–18 (mean: 15, SD: 1.354), 6 males and 7 females, from a University Technical College in England, with differing levels of experience in using VR were selected. They were only permitted to participate if they did not suffer from epilepsy, migraines, or motion sickness when in moving vehicles, in line with recommendations on the use of VR (Sony Interactive Entertainment LLC, 2020).

5.2. Results
5.2.1. Means and Standard Deviations
Firstly, the means and standard deviations (SDs) of all measures (global lostness, local lostness, presence, cognitive interest, and engagement) were compared across participants. The outcomes were as follows (see **Table 2**):
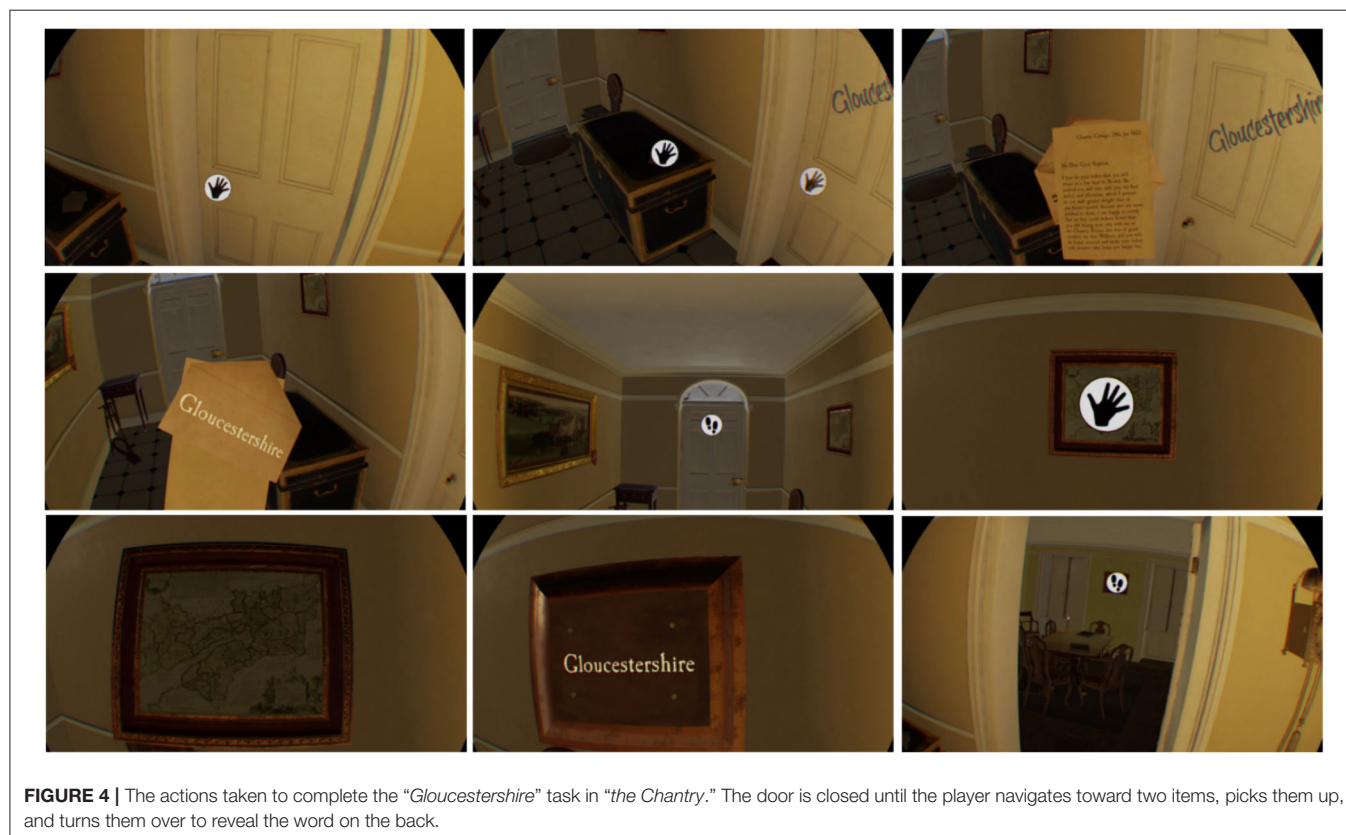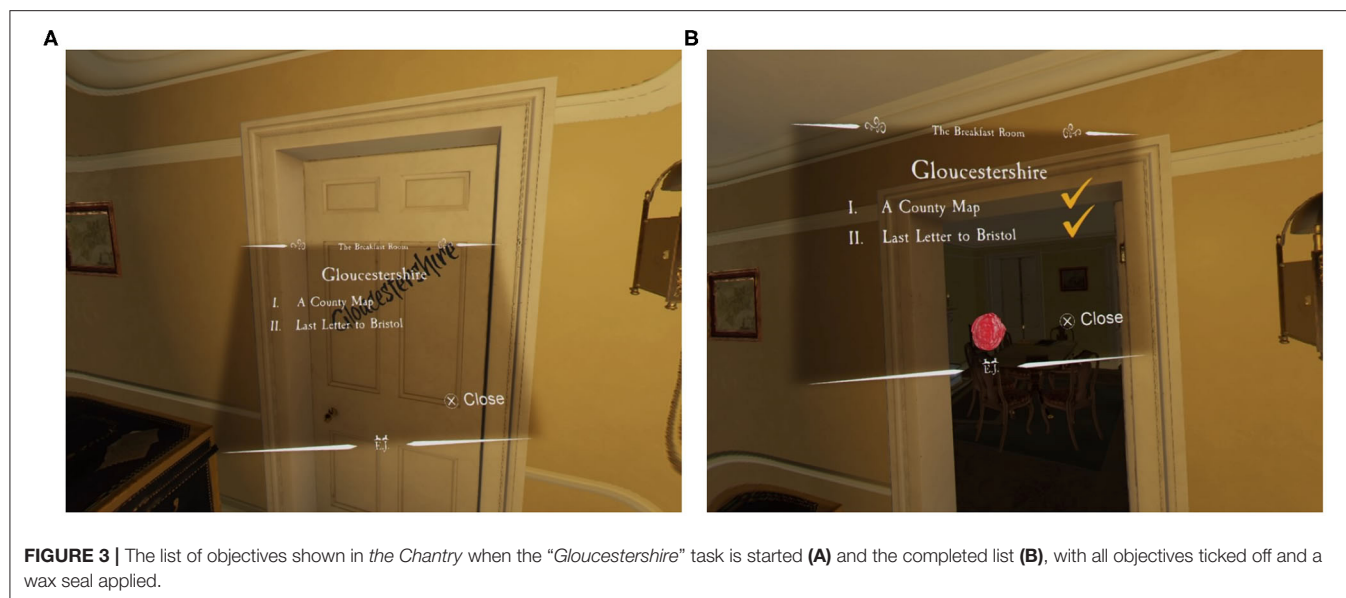
FIGURE 3 | The list of objectives shown in *the Chantry* when the "*Gloucestershire*" task is started **(A)** and the completed list **(B)**, with all objectives ticked off and a wax seal applied.

FIGURE 4 | The actions taken to complete the "*Gloucestershire*" task in "*the Chantry*." The door is closed until the player navigates toward two items, picks them up, and turns them over to reveal the word on the back.

This shows that participants were only moderately effective at the in-game tasks, as the average of both lostness measures was between 0.40 and 0.60; local lostness was 0.464 (SD: 0.112) and global lostness was 0.571 (SD: 0.147). The difference between the global lostness and local lostness is considerable [$t_{(12)} = 4.474, p < 0.001$] and global lostness is, in general, higher than the local lostness. Therefore, "globally" players are more lost than "locally." Despite this, as expected, both lostness measures correlated significantly (0.811, $p < 0.001$).

When it comes to the knowledge test, the mean proportion of correct items on the overall knowledge test was 0.590 (SD: 0.147), somewhat above the chance level. This is in line with the spatial

questions, which were only answered slightly better, with a mean of 0.615 (SD: 0.173).

For experience aspects, the results were moderately positive, as they were around or above the mid-point of the five-point scale. The mean cognitive interest was 3.654 (SD: 0.521), engagement was 2.915 (SD: 0.515), and presence 3.005 (SD: 0.350).

5.2.2. Zero-Order Correlations

The main research question was to investigate if the lostness and game experience measures indicate how well the player learns from a game. To begin with, the zero-order correlations are presented in **Table 3**.

The results of this correlation analysis show that both lostness measures correlate negatively with learning, as is expected, although only the correlation with local lostness is significant ($p = 0.018$) whereas global lostness is not significant ($p = 0.138$). This strong significant (negative) correlation gives weight to the assumption that local lostness can be used as a game analytic measure.

Although the cognitive interest and engagement measures did not show a significant correlation with the knowledge test results, there is, interestingly, a negative significant correlation ($p = 0.031$) between presence and the knowledge test results. This showed that participants learned less the more present they felt.

5.2.3. Multiple Linear Regression Analyses

The main analysis of the current study is to examine whether lostness measures and player experience aspects are related to knowledge acquisition (i.e., performance on the knowledge test). To investigate this, statistical analyses were performed in which the relationship between the final knowledge test as

criterion variable and both lostness measures, as well as cognitive interest, engagement, and presence, are computed as predictors in regression analyses. This statistical technique enables us to examine the relative influence of variables (the predictors) corrected for the influence of the other variables on the criterion variable, the knowledge test.

We conducted multiple linear regression analyses with all these variables, applying the ENTER method offered by *SPSS*. We checked the normality distributions of the variables and their error distributions, using the Shapiro-Wilk test, and found that two variables (spatial knowledge and cognitive interest) deviated significantly from normality ($p < 0.05$). However, regression analysis has been shown to be quite robust even when the normality assumption is violated (Hair et al., 2009, p. 197). Some scholars even claim that normality is not demanded at all (Lumley et al., 2002; Schmidt and Finan, 2018). This seems to hold particularly with larger samples. In this study, due to practical constraints, only 13 participants were observed. Further, we checked the collinearity of the independent (predictor) variables. In all cases the conditions of non-collinearity were satisfied ($p < 0.05$), based on the Variance Inflation Factor values.

The results of the multiple linear regression analyses are shown in **Table 4**.

Multiple $R = 0.855, F_{(5, 12)} = 3.821, p = 0.055$ for Overall Knowledge;

Multiple $R = 0.90, F_{(5, 12)} = 5.94, p = 0.019$ for Spatial Knowledge.

The multiple correlation coefficient concerning overall knowledge is 0.855, thus explaining 73% of the variance on the knowledge test. The main contributing variables, reported in **Table 4**, are local lostness (*Beta-coefficient* $= -0.746$) and presence (*Beta-coefficient* $= -0.510$) with presence being significant ($p = 0.043$) and local lostness showing a trend ($p = 0.092$). Cognitive interest and engagement did not appear to be relevant.

We also checked separately which variables explain spatial knowledge acquisition: we find a multiple correlation coefficient of 0.90, explaining 81% of the variance of the spatial knowledge test. Local lostness (*Beta-coefficient* $= -0.832, p = 0.037$) and presence (*Beta-coefficient* $= -0.641, p = 0.008$) appeared to be the strongest predictors with presence being significant

TABLE 2 | Lostness, knowledge, and in-game experience questionnaire means and standard deviations from the first study.

	Mean	Standard deviation
Global lostness	0.571	0.147
Local lostness	0.464	0.112
Knowledge test (Proportion correct)	0.590	0.147
Cognitive interest	3.654	0.521
Engagement	2.915	0.515
Presence	3.005	0.350

TABLE 3 | Zero-order correlations between knowledge, lostness, and experience measures from the first study.

	Global lostness	Local lostness	Cognitive interest	Engagement	Presence
Knowledge test	−0.435	−0.643*	0.163	0.054	−0.597*
Global lostness	–	0.811**	−0.288	−0.118	−0.059
Local lostness		–	−0.023	0.067	0.132
Cognitive interest			–	0.440	0.093
Engagement				–	0.050

*$p < 0.05$, **$p < 0.001$.

TABLE 4 | *Beta-coefficients* between knowledge, lostness, and experience measures from the first study.

	Beta-coefficient overall knowledge	*Beta-coefficient spatial knowledge*
Global lostness	0.213	0.598
Local lostness	−0.746 (*)	−0.832*
Presence	−0.510*	−0.641*
Cognitive interest	0.231	0.175
Engagement	0.052	−0.058

(*) $0.05 < p < 0.10$, *$p < 0.05$.

TABLE 5 | Lostness, knowledge, and in-game experience questionnaire means and standard deviations from the combined studies.

	Mean	SD
Global lostness	0.605	0.166
Local lostness	0.484	0.102
Knowledge test (proportion correct)	0.576	0.119
Cognitive interest	3.667	0.693
Engagement	2.976	0.563
Presence	3.179	0.450

TABLE 6 | Zero-order correlations between knowledge, lostness, and experience measures from the combined studies.

	Global lostness	Local lostness	Cognitive interest	Engagement	Presence
Knowledge test	−0.444*	−0.597**	0.377*	0.160	−0.268
Global lostness	–	0.669**	−0.382*	−0.251	0.109
Local lostness		–	−0.197	0.031	0.031
Cognitive interest			–	0.525**	0.251
Engagement				–	0.263

*$p < 0.05$, **$p < 0.001$.

($p = 0.008$) along with local lostness ($p = 0.037$). Global lostness, cognitive interest, and engagement were not found to have any significant effect.

6. EXPANDED SAMPLE FROM A SECOND RELATED STUDY

After the first study had been completed, a second empirical study was carried out with the same game, which examined different research questions. It investigated the effect of interaction and story structure on learning (Ferguson et al., 2020a). However, as the control condition in that study followed the same methodology as our first study (same game, same VR equipment, same measurements, same instruction, similar participants), these participants can be combined with the previous participants, to get a more reliable indication of the predictive validity of the lostness measure.

6.1. Participants

As with the first study, a total of 12 students aged 13–18 (mean: 15.67, SD: 1.231), 10 males and 2 females, were selected from a University Technical College in England, with differing levels of experience in using VR were selected. Again, they were informed that they would be playing a VR educational game and would be tested after playing to investigate how much knowledge they were able to recall.

One (female) participant's results were ultimately excluded from the results for being an outlier, having a local lostness value of more than two standard deviations above the mean (Stevens, 1999). Moreover, this participant struggled to play the game compared to other participants and did not make much progress, taking too much time to the beginning of the game and not accessing the educational content to be learned. This left 11 participants (mean age: 15.25, SD: 1.294, 16 males, 8 females). These 11 participants were combined with the 13 participants from the previous study, giving a total of 24.

6.2. Results
6.2.1. Means and Standard Deviations

Once again, the means and standard deviations of all measures (global lostness, local lostness, presence, cognitive interest, and engagement) were compared across participants. The outcomes were as follows (see **Table 5**):

Once again, this shows that participants were only moderately effective at the in-game tasks, as the average of both lostness measures was 0.484 (SD: 0.102) and 0.605 (SD: 0.166), respectively for local and global lostness. The difference between the global lostness and local lostness is considerable [$t_{(23)} = 4.789, p < 0.001$]. Once again, global lostness is higher than the local lostness, strengthening the claim that players are more lost "globally" than "locally." As before, both lostness measures correlated significantly ($0.669, p < 0.001$).

When it comes to the knowledge test, the mean proportion of correct items on the overall knowledge test was 0.576 (SD: 0.119), also somewhat above the chance level yet marginally lower than the results from the first study (−0.014). Compared to the first study, when focusing on the spatial questions, these were answered slightly worse, with a mean of 0.563 (SD: 0.165). So, both overall and spatial knowledge were roughly in line with each other.

For experience aspects, the results were moderately positive, and, as with the results of the first study were around or above the midpoint of the five-point scale. The mean of cognitive interest was 3.667 (SD:0.693), engagement was 2.976 (SD: 0.563), and presence 3.179 (SD: 0.450).

6.2.2. Zero-Order Correlations

Again, but with a larger dataset, we investigated if the lostness and game experience measures indicate how well the player learns from a game. The zero-order correlations are presented in **Table 6**.

Once again, the results of this correlation analyses show that both lostness measures correlate negatively with learning, yet this time the correlation is significant for global lostness ($p = 0.015$) and highly significant for local lostness ($p = 0.001$). These strong significant (negative) correlations give weight to the assumption that both lostness measures can be used as a game analytic measure yet local lostness is a better measure to use.

Unlike the first dataset, cognitive interest significantly correlated with the knowledge test results ($p = 0.035$) showing that participants learned more the more interested they were. Cognitive interest also significantly correlated with engagement ($p = 0.004$) and global lostness ($p = 0.033$). There was no correlation between the knowledge test results and presence, unlike with the first dataset or, once again, engagement.

TABLE 7 | *Beta-coefficients* between knowledge, lostness, and experience measures from the combined studies.

	Beta-coefficient overall knowledge	Beta-coefficient spatial knowledge
Global lostness	0.217	0.476
Local lostness	−0.667**	−0.629*
Presence	−0.399*	−0.473*
Cognitive interest	0.346 (*)	0.255
Engagement	0.158	0.042

(*) $0.05 < p < 0.10$, *$p < 0.05$, **$p < 0.01$.

6.2.3. Multiple Linear Regression Analyses

Once again, a regression analysis is carried out to examine whether the lostness measures and player experience aspects are related to knowledge acquisition (i.e., performance on the knowledge test).

We checked anew the normality distributions of the dependent variables and their error distributions, using the Shapiro-Wilk test, and found the same variables, spatial knowledge and cognitive interest, still deviated from normality ($p < 0.05$), see the earlier remark on this in section 5.2.3. Once again, in all cases the conditions of non-collinearity were satisfied ($p < 0.05$), based on the Variance Inflation Factor values.

The results of the multiple linear regression analyses are shown in **Table 7**.

Multiple $R = 0.752, F_{(5, 23)} = 4.676, p = 0.007$ for Overall Knowledge;

Multiple $R = 0.617, F_{(5, 23)} = 2.213, p = 0.098$ for Spatial Knowledge.

In this dataset, the multiple correlation coefficient concerning overall knowledge is 0.752, thus explaining 56.6% of the variance on the knowledge test, compared to the 73% found in the first study. The main contributing variables, reported in **Table 7**, are local lostness (*Beta-coefficient* = −0.667, $p = 0.007$), presence (*Beta-coefficient* = −0.399, $p = 0.030$), and cognitive interest (*Beta-coefficient* = 0.346, $p = 0.094$). Again, engagement did not appear to be relevant. Interestingly, despite the significant correlation between global lostness and the knowledge test, this was not a significantly contributing variable. This gives further weight to the assumption that local lostness is the better measure for predicting learning.

We also checked separately which variables explain spatial knowledge acquisition: we find a weakly significant multiple correlation coefficient of 0.617, explaining 38.1% of the variance on the spatial knowledge test. Once again, local lostness (*Beta-coefficient* = −0.629, $p = 0.028$) and presence (*Beta-coefficient* = −0.473, $p = 0.031$) appeared to be the strongest and significant predictors. Global lostness, cognitive interest, and engagement were not found to have any significant effect.

7. DISCUSSION

Altogether, two different novel unobtrusive and implicit hypertext navigational efficiency measures (local and global lostness) and experience aspects (presence, engagement, and cognitive interest) were examined in a VR educational game employing a node-based movement system. Learning was measured by a post-game knowledge test consisting of factual and spatial information items. We will first briefly summarize the zero-order correlations. Using a small sample (12 participants) from an initial study (Ferguson et al., 2020b), we found a negative correlation of both navigation efficiency measures with performance on a knowledge test and, therefore, how well the player learned from the game. However, only the correlation with local lostness was significant. A negative significant correlation was also observed between presence and how well the player has learned from the game, the opposite of what was expected. No significant correlations were found for engagement and cognitive interest. The sample size was subsequently expanded (24 participants) by including data from the control condition from a second study (Ferguson et al., 2020a). When examining this larger sample size, we found significant negative correlations with the knowledge test for both lostness measures, confirming Hypothesis I, with strong support particularly for local lostness. Regarding the experience aspects, again, there was no significant correlation of engagement with the knowledge test results, yet a significant correlation was found for cognitive interest. However, the significant negative correlation between presence and knowledge test results from the first study was not present in the larger dataset. Nevertheless, both sets of data led to a partial rejection of Hypothesis II, because only cognitive interest correlates positively with learning while presence in Study 1 correlates even negatively.

Interestingly, cognitive interest was also found to significantly and negatively correlate with global lostness. Participants with higher interest for the key topics presented by the game show higher global navigation efficiency. Apparently cognitive interest is positively associated with global navigation efficiency, and it is worthwhile to note that both correlate positively with knowledge acquisition in the extended study. It must also be addressed that there was no significant effect between engagement and learning, although there is a highly significant correlation between engagement and cognitive interest, which have previously been found to be positively related when it comes to learning (Mazer, 2013). This lack of correlation may be related to the fact that the Game Engagement Questionnaire (GEQ) was developed considering video game violence (Brockmyer et al., 2009) so may not be appropriate to our learning context. Other research also notes that multiple measures of engagement should be used, particularly physiological measurements (Appleton et al., 2006; Boyle et al., 2011). It could be that the use of a different questionnaire, such as the User Experience Scale (Wiebe et al., 2014) or these additional measures would reveal an effect of engagement on learning.

Of more importance than the above correlations are the *Beta-coefficients* from the multiple linear regression analyses. These outcomes enable us to get an estimate of the influence of the predictor variables when they are corrected for the influence of the other predictor variables. Furthermore, they inform us of the relative contribution of the respective predictor variables to the variance in the criterion variable.

Starting with the lostness measures, in the smaller sample and examining the effect of each of the independent variables on the results of the knowledge test, whilst taking all of the other variables into account, we find a weakly significant negative *Beta-coefficient* between local lostness and the overall knowledge test results and a significant *Beta-coefficient* when focusing on the spatial aspects on the knowledge test. When carrying out the same analyses in the larger sample, we again find, for local lostness, negative *Beta-coefficients*, this time highly significant for overall knowledge and significant for spatial knowledge. Global lostness has no significant *Beta-coefficients* in either of the samples. Therefore, a higher local lostness score, that is, when a player has low navigational efficiency and is disoriented when navigating an environment, with more detours and revisits, apparently leads to less knowledge acquisition. Conversely, a player with high navigational efficiency (not disoriented) will learn better. This is evidenced by the highly significant negative correlation and *Beta-coefficients*. Thus, it appears to indicate that local lostness measures how well players process information and that high local lostness indicates difficulties in this information processing and cognitive overload (Gwizdka and Spence, 2007).

When it comes to the experience aspects, in the smaller sample, no significant *Beta-coefficient* was found between both engagement and cognitive interest and learning, although a significant negative *Beta-coefficient* was found between presence and both overall knowledge and spatial knowledge. This result is repeated in the larger sample. These results indicate that, for both overall and spatial knowledge, a person learns less the more present they feel. Slightly different results are found in the larger sample for the other two experience aspects. There was a weakly significant *Beta-coefficient* of cognitive interest with overall knowledge, backing up previous findings (Van der Sluis et al., 2014). The finding that lostness predicts learning performance is expected and aligns with cognitive load theory, which is expected since disorientation, measured by lostness, is known to invoke cognitive overload and have an adverse effect on learning. In contrast, it was expected that any effect of presence on learning would be positive rather than negative which was the case in both studies. Research has found that building a spatial mental model of a situation is expected to lead to a sense of spatial presence (Bailey and Witmer, 1994; Lee et al., 2010). However, other recent research was unable to find a significant relationship (Alsina-Jurnet and Gutiérrez-Maldonado, 2010; Ling et al., 2013; Coxon et al., 2016). It could be that the additional feeling of presence involves cognitive overload. Recent studies have shown that the higher immersion of virtual reality and higher presence can lead to higher cognitive load, resulting in decreased learning (Schrader and Bastiaens, 2012; Makransky et al., 2019; Frederiksen et al., 2020). However, it must be noted that, for both samples, there was no significant correlation between presence and either of the lostness measures. This would indicate that presence is increasing cognitive load independently to lostness. Altogether, in the literature, there are three types of cognitive load identified: intrinsic, which is associated with the complexity of the task, extraneous, which is related to the way that the task is presented, and germane, which is produced by processing and constructing schemas to handle new information for learning

new skills (Sweller et al., 2011). Intrinsic and germane cognitive load are an essential part of the learning process, and it can be assumed that lostness represents extraneous cognitive load, as it is related to the presentation of the task itself. Increased presence, in the form of enhanced visual information provided by VR, could also be leading to increased extraneous cognitive load. This could be an inherent weakness in the use of VR for learning. It could even point toward research stressing that educational games should be designed differently to better make use of VR's advantages when transferring information. Recent research suggests that communicating information visually is preferable when using VR, making the most of its prominent modality (Huang et al., 2019).

Proposals to avoid the problem of cognitive overload, based on cognitive load theory have been both successful and unsuccessful (Andersen et al., 2016a,b). The negative effect that presence has on learning suggests that reducing extraneous cognitive load could increase learning, something often suggested (Sweller et al., 1998). In this respect, the effect of lostness on learning shows that this measure could be used to adapt the game difficulty to the proficiency of players to assist them before cognitive load is too high, players become too overloaded and their learning is negatively affected. Adaptation would reduce extraneous load and would build on our previous research using the same game (Ferguson et al., 2020a), which showed that participants learned significantly more when they were guided through the environment at the expense of a lesser feeling of cognitive interest and presence. Thus, it should be investigated if aiding players when they are identified as being disoriented can lead to the same high levels of learning, by reducing cognitive load, without negatively affecting the experience players felt they had in the game. The reduced extraneous cognitive load may free up cognitive capacity so that extraneous cognitive load caused by presence does not have an impact on learning, further enhancing the experience whilst learning. This will be studied as a priority in our future research. Further research must also be carried out to determine whether the measure can be generalized to other types and genres of games, in other areas, or using other locomotion techniques and different activities.

It must be addressed that there is a strong and highly significant correlation between global and local lostness in both the smaller and larger dataset. This is expected as both measures are different interpretations of lostness, and both have overlapping activities and make use of the same original lostness formula. Yet the constraint of non-collinearity was satisfied. Non-collinearity means the amount of variability of an independent (predictor) variable not explained by the other independent variables (Hair et al., 2009). In other words, when the non-collinearity is high, the independent variables contribute sufficient unique variance. Although global lostness had similar correlations as local lostness, the *Beta-coefficients* with learning, both overall and spatially, are different and for local lostness significant but not for global lostness, showing that local lostness is, empirically, a better measure.

Educational narrative games, by nature, encourage exploration. Players are more inclined to look around and encounter different tasks and objectives, likely carrying out

multiple tasks simultaneously. This comes back to the point that games are not always linear (Hudson, 2011). Local lostness is more appropriate for these types of games as it considers all possible objectives. Conversely, global lostness considers each task independently, giving a value of lostness from the start of each task until the end. Thus, some players may be incorrectly regarded as lost, because they are carrying out more than the minimum number of steps and revisiting different nodes for each task, even though they are completing other necessary tasks and objectives. This is especially true when one task may rely on another task or objective being completed before all task objectives become available. This could be a major reason for the lack of a significant *Beta-coefficient* of global lostness with learning. Another reason for this low *Beta-coefficient* could be that the knowledge test used was not assessing the gathering aspect of information processing by consisting of questions regarding isolated facts. This would make it more suitable for fact-finding tasks and, thus, local lostness. Perhaps having questions more suited to gathering activities, such as mentioning different objectives related to a task, would be more appropriate when using global lostness. Further research is needed to examine this issue. All in all, based on empirical as well of theoretical considerations, local lostness should be preferred as an online implicit (real-time) measure for learning in VR educational games.

8. CONCLUSION

Overall, it was shown in this study that an implicit and unobtrusive hypertext usability measure has high real-time predictive validity when applied to a certain genre of educational VR games. This applies specifically for games where information-searching activities, involving navigation, are an intrinsic part of the learning process and node-based movement is used as the locomotion system. Besides, we also identified another variable, presence, which seems to indicate, given the significant negative *Beta-coefficient* with the knowledge test, that participants get more overloaded as they feel more present and immersed. These results also confirmed the importance of cognitive interest, showing that a person learns more if they are more interested in the topic.

Overall, this opens the door for the local lostness measure to be meaningfully used as an implicit online game analytic in a dynamic adaptive system, offering assistance to players before they suffer from cognitive overload in educational games. This would offer new opportunities for on-line, in-game adaptivity dependent on user performance during play (Alsina-Jurnet and Gutiérrez-Maldonado, 2010; Shute et al., 2017). Perhaps this

player-dependent adaptivity can support the cognitive processing of educational materials and prevent overload at the right time, resulting in more effective learning (Salzman et al., 1999; Lee et al., 2010).

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Sheffield Hallam University Review Ethics Committee (SHUREC). Written informed consent to participate in this study was provided by the participants, and where necessary, the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

CF wrote the manuscript, designed and conducted the experimental study, and analyzed the data. HvO co-wrote the manuscript and oversaw the project. All authors contributed to the article and approved the submitted version.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.578154/full#supplementary-material

REFERENCES

Abdul Jabbar, A. I., and Felicia, P. (2015). Gameplay engagement and learning in game-based learning: a systematic review. *Rev. Educ. Res.* 85, 740–779. doi: 10.3102/0034654315577210

Alsina-Jurnet, I., and Gutiérrez-Maldonado, J. (2010). Influence of personality and individual abilities on the sense of presence

experienced in anxiety triggering virtual environments. *Int. J. Hum. Comput. Stud.* 68, 788–801. doi: 10.1016/j.ijhcs.2010.07.001

Andersen, S. A. W., Mikkelsen, P. T., Konge, L., Cayé-Thomasen, P., and Sørensen, M. S. (2016a). Cognitive load in distributed and massed practice in virtual reality mastoidectomy simulation. *Laryngoscope* 126, E74–E79. doi: 10.1002/lary.25449

Andersen, S. A. W., Mikkelsen, P. T., Konge, L., Cayé-Thomasen, P., and Sørensen, M. S. (2016b). The effect of implementing cognitive load theory-based design principles in virtual reality simulation training of surgical skills: a randomized controlled trial. *Adv. Simul.* 1:20. doi: 10.1186/s41077-016-0022-1

Appleton, J. J., Christenson, S. L., Kim, D., and Reschly, A. L. (2006). Measuring cognitive and psychological engagement: validation of the student engagement instrument. *J. School Psychol.* 44, 427–445. doi: 10.1016/j.jsp.2006.04.002

Bailey, J. H., and Witmer, B. G. (1994). Learning and transfer of spatial knowledge in a virtual environment. *Proc. Hum. Fact. Ergon. Soc. Annu. Meet.* 38, 1158–1162. doi: 10.1177/154193129403801803

Bolter, J. (2000). *Writing Space: Computers, Hypertext, and the Remediation of Print.* Mahwah, NJ: Lawrence Erlbaum Associates. doi: 10.4324/9781410600110

Boyle, E., Connolly, T. M., and Hainey, T. (2011). The role of psychology in understanding the impact of computer games. *Entertain. Comput.* 2, 69–74. doi: 10.1016/j.entcom.2010.12.002

Boyle, E. A., Connolly, T. M., Hainey, T., and Boyle, J. M. (2012). Engagement in digital entertainment games: a systematic review. *Comput. Hum. Behav.* 28, 771–780. doi: 10.1016/j.chb.2011.11.020

Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., and Pidruzny, J. N. (2009). The development of the game engagement questionnaire: a measure of engagement in video game-playing. *J. Exp. Soc. Psychol.* 45, 624-634. doi: 10.1016/j.jesp.2009.02.016

Chen, F. (2016). *Robust Multimodal Cognitive Load Measurement.* Cham: Springer. doi: 10.1007/978-3-319-31700-7

Conklin, J. (1987). Hypertext: an introduction and survey. *Computer* 20, 17–41. doi: 10.1109/MC.1987.1663693

Coxon, M., Kelly, N., and Page, S. (2016). Individual differences in virtual reality: are spatial presence and spatial ability linked? *Virt. Real.* 20, 203–212. doi: 10.1007/s10055-016-0292-x

Drachen, A., and Schubert, M. (2013). "Spatial game analytics," in *Game Analytics: Maximizing the Value of Player Data*, eds M. Seif El-Nasr, A. Drachen, and A. Canossa (London: Springer London), 365–402. doi: 10.1007/978-1-4471-4769-5_17

Edwards, D. M, and Hardman, L. (1989). "Lost in hyperspace: cognitive mapping and navigation in a hypertext environment," in *Hypertext: Theory Into Practice*, ed R. McAleese (Norwood, NJ: Ablex Publishing), 90–105.

Elshout, J. (1976). *Karakteristieke Moeilijkheden in Het Denken (Characteristic Difficulties in Thinking).* Amsterdam: Universiteit Van Amsterdam.

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

Ferguson, C., Van den Broek, E. L., and Van Oostendorp, H. (2020a). On the role of interaction mode and story structure in virtual reality serious games. *Comput. Educ.* 143:103671. doi: 10.1016/j.compedu.2019.103671

Ferguson, C., Van den Broek, E. L., Van Oostendorp, H., de Redelijkheid, S., and Giezeman, G.-J. (2020b). Virtual reality aids game navigation: evidence from the hypertext lostness measure. *Cyberpsychol. Behav. Soc. Network.* 23, 635–641. doi: 10.1089/cyber.2019.0435

Frederiksen, J. G., Sørensen, S. M. D., Konge, L., Svendsen, M. B. S., Nobel-Jørgensen, M., Bjerrum, F., et al. (2020). Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: a randomized trial. *Surg. Endosc.* 34, 1244–1252. doi: 10.1007/s00464-019-06887-8

Furlough, C. S., and Gillan, D. J. (2018). Mental models: structural differences and the role of experience. *J. Cogn. Eng. Decis. Mak.* 12, 269-287. doi: 10.1177/1555343418773236

Gwizdka, J., and Spence, I. (2007). Implicit measures of lostness and success in web navigation. *Interact. Comput.* 19, 357–369. doi: 10.1016/j.intcom.2007.01.001

Habgood, M. P. J., Moore, D., Wilson, D., and Alapont, S. (2018). "Rapid, continuous movement between nodes as an accessible virtual reality locomotion technique," in *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (Reutlingen), 371–378. doi: 10.1109/VR.2018.8446130

Hair, J., Anderson, R., and Babin, B. (2009). *Multivariate Data Analysis.* Upper Saddle River, NJ: Prentice Hall.

Harp, S. F., and Mayer, R. E. (1997). The role of interest in learning from scientific text and illustrations: on the distinction between emotional interest and cognitive interest. *J. Educ. Psychol.* 89, 92–102. doi: 10.1037/0022-0663.89.1.92

Head, M., Archer, N., and Yuan, Y. (2000). World wide web navigation aid. *Int. J. Hum. Comput. Stud.* 53, 301–330. doi: 10.1006/ijhc.2000.0392

Huang, K.-T., Ball, C., Francis, J., Ratan, R., Boumis, J., and Fordham, J. (2019). Augmented versus virtual reality in education: an exploratory study examining science knowledge retention when using augmented reality/virtual reality mobile applications. *Cyberpsychol. Behav. Soc. Network.* 22, 105–110. doi: 10.1089/cyber.2018.0150

Hudson, K. (2011). *Player-Driven Stories: How Do We Get There?* Available online at: https://www.gdcvault.com/play/1014657/Player-Driven-Stories-How-Do (accessed September 20, 2020).

Imlig-Iten, N., and Petko, D. (2018). Comparing serious games and educational simulations: effects on enjoyment, deep thinking, interest and cognitive learning gains. *Simul. Gam.* 49, 401–422. doi: 10.1177/1046878118779088

Karanam, S., and Van Oostendorp, H. (2020). "Cognitive modeling of age and domain knowledge differences in information search," in *Understanding and Improving Information Search: A Cognitive Approach*, eds W. T. Fu and H. van Oostendorp (Cham: Springer International Publishing), 47–68. doi: 10.1007/978-3-030-38825-6_4

Karanam, S., Van Oostendorp, H., and Fu, W. T. (2016). Performance of computational cognitive models of web-navigation on real websites. *J. Inform. Sci.* 42, 94–113. doi: 10.1177/0165551515615842

Kim, S. (2018). *Gamification in Learning and Education: Enjoy Learning Like Gaming.* Cham: Springer.

Kitajima, M., Blackmon, M. H., and Polson, P. G. (2000). "A comprehension based model of web navigation and its application to web usability analysis," in *People and Computers XIV-Usability or Else!*, eds S. McDonald, Y. Waern, and G. Cockton (London: Springer), 357–373. doi: 10.1007/978-1-4471-0515-2_24

Lee, E. A.-L., Wong, K. W., and Fung, C. C. (2010). How does desktop virtual reality enhance learning outcomes? A structural equation modeling approach. *Comput. Educ.* 55, 1424–1442. doi: 10.1016/j.compedu.2010.06.006

Lester, J. C., Spires, H. A., Nietfeld, J. L., Minogue, J., Mott, B. W., and Lobene, E. V. (2014). Designing game-based learning environments for elementary science education: a narrative-centered learning perspective. *Inform. Sci.* 264, 4–18. doi: 10.1016/j.ins.2013.09.005

Lester, J. C., Spires, H. A., Nietfeld, J. L., Minogue, J., Mott, B. W., and Lobene, E. V. (2018). A meta-analysis with examination of moderators of student cognition, affect, and learning outcomes while using serious educational games, serious games, and simulations. *Comput. Hum. Behav.* 80, 158–167. doi: 10.1016/j.chb.2017.10.040

Ling, Y., Nefs, H. T., Brinkman, W.-P., Qu, C., and Heynderickx, I. (2013). The relationship between individual characteristics and experienced presence. *Comput. Hum. Behav.* 29, 1519–1530. doi: 10.1016/j.chb.2012.12.010

Loh, C., Yanyan, S., and Ifenthaler, D. (2015). *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement.* Cham: Springer International Publishing, doi: 10.1007/978-3-319-05834-4

Lumley, T., Diehr, P., Emerson, S., and Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annu. Rev. Public Health* 23, 151–169. doi: 10.1146/annurev.publhealth.23.100901.140546

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

Malone, T. W., and Lepper, M. R. (1987). "Making learning fun: a taxonomy of intrinsic motivations for learning," in *Aptitude, Learning and Instruction III: Conative and Affective Process Analyses*, eds R. Snow and M. Farr (Hillsdale, NJ: Lawrence Erlbaum Associates Publishers), 223–253.

Mazer, J. P. (2013). Associations among teacher communication behaviors, student interest, and engagement: a validity test. *Commun. Educ.* 62, 86–96. doi: 10.1080/03634523.2012.731513

Perez-Colado, I., Alonso-Fernandez, C., Freire, M., Martinez-Ortiz, I., and Fernandez-Manjon, B. (2018). "Game learning analytics is not informagic!," in *2018 IEEE Global Engineering Education Conference (EDUCON)* (Tenerife: IEEE), 1729–1737. doi: 10.1109/EDUCON.2018.8363443

Puerta Melguizo, M. C., Vidya, U., and Van Oostendorp, H. (2012). Seeking information online: the influence of menu type, navigation path complexity

and spatial ability on information gathering tasks. *Behav. Inform. Technol.* 31, 59–70. doi: 10.1080/0144929X.2011.602425

Salzman, M. C., Dede, C., Loftin, R. B., and Chen, J. (1999). A model for understanding how virtual reality aids complex conceptual learning. *Presence* 8, 293–316. doi: 10.1162/105474699566242

Schmidt, A. F., and Finan, C. (2018). Linear regression and the normality assumption. *J. Clin. Epidemiol.* 98, 146–151. doi: 10.1016/j.jclinepi.2017.12.006

Schrader, C., and Bastiaens, T. J. (2012). The influence of virtual presence: effects on experienced cognitive load and learning outcomes in educational computer games. *Comput. Hum. Behav.* 28, 648–658. doi: 10.1016/j.chb.2011.11.011

Schraw, G., Bruning, R., and Svoboda, C. (1995). Sources of situational interest. *J. Read. Behav.* 27, 1–17. doi: 10.1080/10862969509547866

Schubert, T. W. (2003). The sense of presence in virtual environments. *Zeitsch. Medienpsychol.* 15, 69–71. doi: 10.1026//1617-6383.15.2.69

Sharples, M. (1999). *How We Write: Writing as Creative Design.* London: Routledge. doi: 10.4324/9780203272732

Shute, V., Ke, F., and Wang, L. (2017). "Assessment and adaptation in games," in *Instructional Techniques to Facilitate Learning and Motivation of Serious Games*, eds P. Wouters and H. van Oostendorp (Cham: Springer International Publishing), 59–78. doi: 10.1007/978-3-319-39298-1_4

Slater, M. (2002). Presence and the sixth sense. *Presence* 11, 435–439. doi: 10.1162/105474602760204327

Smith, P. A. (1996). Towards a practical measure of hypertext usability. *Interact. Comput.* 8, 365–381. doi: 10.1016/S0953-5438(97)83779-4

Sony Interactive Entertainment LLC (2020). *Playstation Health Warnings.* Available online at: https://www.playstation.com/en-us/network/legal/health-warnings/ (accessed September 20, 2020).

Steuer, J. (1992). Defining virtual reality: dimensions determining telepresence. *J. Commun.* 42, 73–93. doi: 10.1111/j.1460-2466.1992.tb00812.x

Stevens, J. (1999). *Intermediate Statistics: A Modern Approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sweller, J., Ayres, P., and Kalyuga, S. (2011). *Cognitive Load Theory.* New York, NY: Springer. doi: 10.1007/978-1-4419-8126-4

Sweller, J., Van Merrienboer, J. J. G., and Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educ. Psychol. Rev.* 10, 251–296. doi: 10.1023/A:1022193728205

Van der Sluis, F., Van den Broek, E. L., Glassey, R. J., Van Dijk, E. M. A. G., and de Jong, F. M. G. (2014). When complexity becomes interesting. *J. Assoc. Inform. Sci. Technol.* 65, 1478–1500. doi: 10.1002/asi.23095

Van Oostendorp, H., Aggarwal, S., and Karanam, S. (2017). "Development of a cognitive model of web-navigation," in *Reading Comprehension in Educational Settings*, eds J. Leon and I. Escudero (Amsterdam: John Benjamins Publishing Company), 69–192. doi: 10.1075/swll.16.07van

Van Oostendorp, H., and Karanam, S. (2013). "Navigating in a virtual environment with model-generated support," in *Proceedings of the 31st European Conference on Cognitive Ergonomics, ECCE '13* (New York, NY: Association for Computing Machinery), 1–8. doi: 10.1145/2501907.2501948

Van Oostendorp, H., Van der Spek, E. D., and Linssen, J. (2014). Adapting the complexity level of a serious game to the proficiency of players. *EAI Endors. Trans. Ser. Games* 1, 1–8. doi: 10.4108/sg.1.2.e5

Wasserman, J. A., and Banks, J. (2017). Details and dynamics: mental models of complex systems in game-based learning. *Simul. Gam.* 48, 603–624. doi: 10.1177/1046878117715056

Westera, W., Nadolski, R., and Hummel, H. (2014). "Learning analytics in serious gaming: uncovering the hidden treasury of game log files," in *Games and Learning Alliance*, eds A. De Gloria (Cham: Springer International Publishing), 41–52. doi: 10.1007/978-3-319-12157-4_4

Wiebe, E. N., Lamb, A., Hardy, M., and Sharek, D. (2014). Measuring engagement in video game-based environments: investigation of the user engagement scale. *Comput. Hum. Behav.* 32, 123–132. doi: 10.1016/j.chb.2013.12.001

Challenges in Process Dissociation Measures for Moral Cognition

Anton Kunnari[1]*[†], Jukka R. I. Sundvall[2][†] and Michael Laakasuo[2][†]

[1] *Department of Psychology and Logopedics, University of Helsinki, Helsinki, Finland,* [2] *Department of Digital Humanities, Cognitive Science, University of Helsinki, Helsinki, Finland*

The process dissociation procedure (PDP) for moral cognition was created to separately measure two dispositions of moral judgment based on the dual-process theory of moral reasoning: deontological and utilitarian inclinations. In this paper we raise some concerns from a psychometrics perspective regarding the structure, reliability, and validity of the moral PDP as a measure of individual differences. Using two simulation studies as well as a real sample of $N = 1,010$, we investigate the psychometric properties of the moral PDP. We present novel evidence showing that (1) some correlations between PDP parameters are mathematical artifacts, and as such cannot be taken as evidence in support of a theory, (2) there are severe response inconsistencies within dilemma batteries, and (3) reliability estimates for these scores seem to be far below the accepted standards. We discuss some potential theoretical and content-related reasons for these statistical issues and their implications. We conclude that in their current form, PDP measures of utilitarian and deontological tendencies are sub-optimal for assessing individual differences.

Keywords: process dissociation, utilitarianism, deontology, measurement, psychometrics, simulation studies, validity

INTRODUCTION

Valid and accurate measurement is one of the cornerstones of scientific inquiry. The psychological sciences have had concerns about how measurement instruments are created (Hedge et al., 2018), evaluated (Flake et al., 2017), and used (Flake and Fried, 2019). This has resulted in questioning the validity of both custom-made scales (see Flake and Fried, 2019) and more established instruments (Hussey and Hughes, 2020), as well as having likely contributed to the ongoing replication crisis (Loken and Gelman, 2017).

In the present paper we raise some potential pitfalls in a measure of individual differences in utilitarian and deontological moral inclinations, the moral process dissociation procedure (PDP) developed by Conway and Gawronski (2013); from here on referenced as C&G. We describe both this measure and our concerns in length below, but the main concerns are:

1. Issues with content: heterogeneity between moral dilemmas not reflected in how scores are calculated.
2. Insufficient justification for scale structure: lack of psychometric work to examine if items function well together and whether they should be combined.
3. Insufficient reliability: lack of reliability reporting and low internal consistencies observed in the current work.
4. Artefactual validity evidence: some correlations between parameters arise solely due to mathematical formulae used to compute the moral PDP scores.

First, we will cover some background on moral psychology of utilitarianism and deontology in general: the use of moral dilemmas to measure utilitarian/deontological preferences, and the dual-process model of moral judgment. We will briefly cover the consequent methodological discussion that resulted in the development of the moral PDP, and the caveats with this model.

Measuring Utilitarian and Deontological Dispositions

Moral psychology has put great focus on deontological and utilitarian preferences. In brief, deontological thinking sees some acts as forbidden (e.g., the intentional killing of another person) because of their inherent immorality, regardless of potential benefits. In contrast, utilitarian thinking perceives ethicality of an action based on its consequences. These moral preferences are often measured with so-called sacrificial moral dilemmas (see Christensen and Gomila, 2012, for a review). These dilemmas are typically variations of a central theme, present in the classical trolley dilemma, where the moral agent needs to personally cause the death (or other injury) of at least one person if they wish to save a larger group of people from death (or other injury).

Research using these dilemmas was the basis for the dual-process model of moral judgment (see Greene, 2007) which posited that utilitarian and deontological responses were driven by two separate cognitive processes, one fast, automatic and emotion-based (implicit) and one slow and deliberate (explicit). According to the theory, a prospect of sacrificing a person for the greater good leads to a negative emotional response that drives moral disapproval (i.e., a deontological judgment of the proposed harm), and utilitarian reasoning is possible after overriding this response.

However, C&G argued that the traditional framing of the moral dilemmas was confounded, because they did not allow for the differentiation of the contributions of two processes. Scores from these dilemmas represented a bipolar continuum where deontological and utilitarian responding were the opposite endpoints. That is, the utilitarian choice always conflicted with the deontological choice: it was not possible to know whether, e.g., a high "utilitarian" score reflected strong utilitarian tendencies, weak deontological tendencies, or both.

To address this problem C&G created a process dissociation model[1] to measure separately the strength of utilitarian and deontological inclinations. C&G based their model on the one originally developed by Jacoby (1991) in order to clarify the relative contributions of automatic and deliberate memory processes. The stated purpose in C&Gs own words was "to provide a compelling test of the dominant dual-process account of moral judgment" (p. 220). They also stated that calculating separate parameters for utilitarian and deontological inclinations

for each individual in a sample allows "researchers to use [the parameters] as measurement scores in experimental or *individual differences* designs" (p. 220, italics ours). Although the latter is our main concern here, we will also examine the claim about testing the dual-process theory.

The PDP for moral cognition computes two parameters to represent the strength of an individuals' utilitarian (U) and deontological (D) inclinations. This is done by using two dilemma types: congruent (C), where both utilitarianism and deontology should point towards an action being impermissible; and incongruent (IC), where utilitarianism should permit the act but deontology should not. Responses to C and IC dilemmas are used to compute U and D scores according to formulae presented in Simulation study 1.

The model was later refined to include a parameter for action preference in CNI model (consequences, norms, inaction; Gawronski et al., 2016, 2017). Despite this more recent approach, studies continue to use the PDP scoring method of calculating individual utilitarianism and deontology scores for participants (e.g., Białek et al., 2019; Mata, 2019; Bostyn et al., 2020).

PD based models have been thought to clarify whether it is processes underlying deontological or utilitarian reasoning that are related to other individual differences, or affected by manipulations such as cognitive load. The effect of cognitive load as well as sex differences found using the traditional sacrificial dilemmas (Greene et al., 2008; Fumagalli et al., 2010), have been verified using the PDP or similar models (Friesdorf et al., 2015; Gawronski et al., 2017). Using the PDP, experimenters have found that utilitarian but not deontological inclinations are related to other cognitive measures such as the cognitive reflection test (Patil et al., 2020). Thus, the PDP seems to replicate results produced by more traditional measurements, but it allows for more specific inferences about where, specifically, individual differences manifest, and what kinds of processes are affected by experimental manipulations. The moral PDP has been quite popular: at the time of writing, C&G's 2013 paper alone has been cited over 400 times, and at least 30 studies have used the measure.

There have been some concerns about the appropriateness of the sacrificial dilemma method, including the moral PDP, as measures of utilitarianism and deontology. Everett and Kahane (2020) argue that the kind of "utilitarianism" measured both by the traditional approach and by the PDP dilemmas is, in the end, not true utilitarianism but only utilitarianism with some qualifications. For example, the U parameter as measured by the moral PDP is not positively correlated with moral views stating an obligation to maximize good. While we think this is an important discussion, our focus is the more proximal psychometric question: even if the moral PDP only measures "utilitarianism with some qualifications", does it measure it well?

We have found few psychometric examinations on the structure of the moral PDP. Baron and Goodwin (2020) have recently shown that participants often interpret norms and consequences in CNI dilemmas (similar in structure to moral PDP dilemmas) in a way not intended by the experimenters. Due to this, they suggested that correlations between these measures and external variables may stem from systematic

---

[1]PD models have been popular with various dual-process theories (Sherman et al., 2014; Calanchini et al., 2018), and have been used for topics such as memory processes, judgment and decision making, object recognition, (implicit) social attitudes, and source monitoring (see Erdfelder et al., 2009; Payne and Bishara, 2009). However, it is important to note that simply applying these models do not by themselves reveal anything qualitative (like implicitness or explicitness) about these processes, and these claims must be established via manipulations and comparisons with external criteria (Sherman et al., 2014).

variation in *understanding the dilemmas as intended* rather than any meaningful differences in moral thinking. They also noted that some effects may be driven by only one specific dilemma in a dilemma set, which raises concerns of reliability. We have similar theoretical concerns about the PDP dilemmas and will cover this in more detail in subsequent sections.

The moral PDP is claimed to function as a measurement for individual differences, and as such it must "pass" the same psychometric examinations as other scales. These include justification for combining items (here dilemmas or dilemma pairs) into scales and examining their measurement accuracy. Whereas the original PDP was used in an experimental manner, the moral PDP uses it to measure individual differences as well. This type of approach has recently gained attention for causing reliability issues[2] (Hedge et al., 2018). C&G explicitly consider the moral PDP applicable for both purposes. We want to highlight the position we are taking here: we are not attempting to challenge the usefulness of PD models in general, nor the majority of empirical findings they have produced. We specifically examine the properties of the moral PDP as an individual difference instrument. Aggregating responses across participants can cancel out inaccuracies at the individual level, and provide more accurate estimates of process magnitudes at the group level (Calanchini et al., 2018).

Here, we raise psychometric concerns about using the PDP to quantify individual differences in utilitarian and deontological tendencies. We raise these concerns and structure the rest of the introduction in the following order: (1) content issues, (2) justification for scale structure, (3) reliability of measures, and (4) validity evidence. Thereafter, we present two simulation studies as well as an empirical study to examine the validity evidence and psychometric properties for the moral PDP as an indicator of individual differences.

Concern 1: Content Considerations

The core assumptions of the PDP have been questioned and criticized in the context of memory research from early on (see, e.g., Graf and Komatsu, 1994; Curran and Hintzman, 1995; Russo et al., 1998). Critics have pointed to empirical results contradicting some of the core statistical assumptions in the original PDP. However, we have theoretical reservations about applying the PDP to individual differences in utilitarian moral judgment, even if the main premises of the PDP were sound in general.

Baron and Goodwin (2020) have critiqued the more refined version of the moral PDP model, i.e., the CNI model. They argued that different dilemma types in PDP/CNI models permit extensive interpretation by participants, and showed that in some dilemmas there were major disagreements between the participants and the experimenters regarding what the norms and the consequences actually are in individual dilemmas. That is, while an experimenter may interpret a specific response to a dilemma (e.g., not accepting a harm) to be an exemplar

of a certain type of thinking (e.g., following a deontological norm regardless of a utilitarian motive), the responder may not agree (e.g., they have an argument for why they are in fact being utilitarian).

Potential issues in interpreting the dilemmas tie into an issue with the uniformity of stimuli. For the moral PDP, the incongruent dilemmas range from whether to avoid hitting an old lady or a young woman and her child with one's car when it's too late to brake, to whether to kill a young Adolf Hitler in order to prevent the Second World War (Conway and Gawronski, 2013, Appendix A, pp. 231–233). It seems likely that the magnitude of the potential utilitarian benefit (or the strength of the norm against a specific harm) would affect people's responses, but the dilemmas are given equal weight in PDP formulae (see section "Simulation 1"). Something similar applies to the congruent dilemmas as well: although these are intended as situations where the consequences are never good enough to justify the action, there are differences in the kinds of harms and consequences. In other words, there is a hidden "ladder" within the dilemmas, where different dilemmas may test for different levels of utilitarianism or deontology (some more and some less). This is not bad in and of itself, but it is not acknowledged in the calculations (i.e., weights given to the individual dilemmas in how much they measure an inclination), and as such, is not part of the reasoning in the PDP.

The heterogeneity of stimuli ties into a larger question of whether the PDP is appropriate to dissociate processes in moral reasoning. While the PDP has been used as a content-neutral procedure for separating the contributions of two processes behind many different tasks, note that the PDP started as a way of separating contributions of (automatic and conscious) processes in *memory*, specifically word recollection. It seems perfectly reasonable to measure a person's success in a memory task as the number of items recalled from a list. However, it is trickier to measure the *strength* of any moral inclination as the number of specific answers to a series of similar questions about harm. That is, a better performance in memory is characterized by more things remembered, so it makes sense to measure memory performance in this way. A "better performance" in, e.g., utilitarian thinking is characterized by greater acceptance towards utilitarian sacrifice, which does not translate equally well to be measured simply as a number of certain responses unless the heterogeneity in items is accounted for in the calculations.

We understand that a total uniformity of stimuli in the moral PDP would be counterproductive - it would amount to asking the same question ten times. Thus, it makes sense that the levels of utilitarian motive and/or harms vary between the dilemmas, but we feel that this may simply be a concession that becomes necessary because of the structure of the PDP. Since there is no clear ranking of the dilemmas (as in an Item Response Theory approach), or a *priori* knowledge of how strong the deontological norms for each dilemma are, it is unclear how much utilitarian or deontological inclinations actually affect a specific response. In our empirical data, we find not only "hard" and "easy" items in terms of utilitarian responses to the IC dilemmas, but also dilemmas that we would argue show response patterns that undermine their validity.

---

[2]Robustness and replicability of experimental findings is facilitated by low between-subjects variability, but this variability is the very thing that measurement instruments attempt to capture. With low between-subjects variability there is very little to measure.

Concern 2: Justification for Scale Structure

Our second concern is that the internal structure of the moral PDP might not be justified by data. To our knowledge, few tests for the appropriateness of this structure have been done using psychometric methods.[3] In summary, we have some concerns with how combining the C and IC dilemmas into the U and D scales is justified. These include insufficient appraisal of overall structure and item functioning.

The formulae used to compute the U and D scores from responses to C and IC dilemmas are:

$$^4U = P(\text{Unacceptable} \mid \text{Congruent})$$
$$- P(\text{Unacceptable} \mid \text{Incongruent}) \qquad (1)$$

$$D = \frac{P(\text{Unacceptable} \mid \text{Incongruent})}{(1 - U)} \qquad (2)$$

where the probabilities represent within-subject averages of the respective dilemma sets with responses with values 1 indicating unacceptance and 0 acceptance. Additionally, the traditional bipolar scores (TS) reflecting the standard scoring of Greene's (2007) high conflict moral dilemmas can be computed as:

$$\text{Traditional Scoring } (TS) = P(\text{Unacceptable} \mid \text{Incongruent}) \qquad (3)$$

An implicit assumption in PDP formulae is that all the dilemmas are "worthy" of combining together: aggregating just any binary responses would not of course make any sense, so they need to exhibit certain statistical relationships. One potential pitfall is that if some items do not correlate well with the scale sum, that would mean that the implied continuity in the parameters would hardly be justified. The moral PDP also assumes that each dilemma functions equally well. Thus the scoring procedure always gives, e.g., the Car Accident and Time Machine dilemmas (Conway and Gawronski, 2013, Appendix A, pp. 231–233) equal weights, as alluded to above. Scale development for individual differences is typically a labor intensive effort to empirically examine these properties, which we have not seen published in the moral PDP context.

Another concern here is the calculation of the D parameter, which involves division by another variable, namely the complement of the U parameter. To our knowledge, this makes it hard to evaluate how well the data 'justifies' the model, at least by using more traditional psychometric methods such as factor

---

[3] PD attempts to provide measures for any underlying processes in a way which in some ways resembles latent variable modeling: the task responses are used to infer the "hidden variables" that underlie the observed response patterns (Sherman et al., 2014). However, unlike more traditional psychometrics, PD models do not attempt to quantify the relationship of each task (or trial) with the unobserved process (like factor loadings or IRT difficulty parameters), and each will have equal weight in PD scores.

[4] In Appendix A, we show with an algebraic proof that the formula of U can also be expressed as sum of the mean of C dilemmas and mean of reverse-coded IC-dilemmas minus a constant. This alternative formulation enables application of more common psychometric methods such as reliability calculations and factor analysis for appraising U scores. We show some such examinations in a later section where we analyze our empirical data.

analysis or Rasch models. Later, we show that it is possible to evaluate the properties of U using such methods by expressing its formula in a different, but equivalent way (see Appendix A for algebraic proof).

Concern 3: Insufficient Evidence for Reliability

Our next concern is that due to its structure, the PDP model does not easily allow one to check for the internal consistency of items (an index of within-participant agreement for multiple items of the instrument indicating a signal-to-noise ratio). Unlike in, e.g., many personality measures, the items of the PDP are not simply averaged together to form an aggregate score. This is likely the reason we have been unable to find usual indices of reliability such as Cronbach's alpha or McDonald's omega (calculated from hierarchial factor analysis, and unlike alpha, does not assume that each item functions equally well; equals alpha when assumptions of alpha are not violated) reported in the PDP literature. This is problematic because agnosticism about reliability makes evaluating sample sizes in power calculations difficult (see Williams and Zimmerman, 1989), as well as interpreting effect sizes (Wilkinson and Task Force on Statistical Inference APA Board of Scientific Affairs, 1999). Large amounts of measurement error (low reliability) can both make it harder to find existing relationships (type 2 error) and lead to spurious findings (type 1 error), both contributing to non-replicability (Loken and Gelman, 2017).

Later, we calculate reliability coefficients for PDP parameters in two ways. First, the alternative formulation for U mentioned above enables the application of standard psychometric methods. Second, to estimate reliability for D as well, we use the split-half permutation method (see Parsons et al., 2019). We use these methods on our own empirical sample and find quite low reliabilities for both of the parameters.

Concern 4: Validity Evidence

The reliability issue is concerning, because sufficient reliability is a necessary precondition for validity (Cook and Beckman, 2006). We are also concerned about some pieces of the presented validity evidence for the moral PDP. The two main pieces of construct validity evidence are the correlations between different PDP scores, and their relationships with other theoretically relevant constructs (Conway and Gawronski, 2013). In our analyses we examine the former, but discuss the external correlations later.

First, C&G (and later Friesdorf et al., 2015, in a meta-analysis of the moral PDP) argued that the correlations between PDP parameters and the traditional (bipolar) scores (TS) were evidence of the confound in traditional moral dilemmas. U had a strong negative relationship with TS, whereas D had a strong positive relationship with TS. C&G claimed that "this finding not only corroborates the validity of the two PD parameters; it also suggests that the traditional bipolar index indeed confounds two distinct processes [...]" (p. 223). In other words, these correlations are taken as proof that two separate processes strongly drive TS. We will later show that these correlations could stem from the properties of the scoring method alone.

In addition, C&G observed a nonexistent-to-small correlation between U and D parameters and considered this consistent with the separateness of two processes. However, we find this line of reasoning inaccurate because U is directly used in calculating D (see the formulas on the following section), and for this reason these cannot be considered truly independent, even in the absence of a (linear) correlation. This becomes apparent when one looks at the scatterplot between the U and D parameters (**Figures 1C**, **2C**, **3C**) where there is a clear non-random pattern stemming from the scoring formula. In our opinion, considering null correlation as evidence for separateness would require full distributional independence of the variables involved. We will elaborate on this later.

C&G presented several correlations to exogenous variables as evidence of construct validity. D had positive correlations with Empathetic Concern, Perspective-taking, Religiosity, and Moral Identity Internalization (MII), whereas U had a positive correlation with MII and a trending positive relationship with Need for Cognition. TS in turn had the same positive relationships as D with the exception of MII, and a negative trending relationship with MII. In addition, the authors showed that a cognitive load manipulation selectively influenced U but not D. These associations make sense in light of the dual-process model, but we have reservations about them based on our other concerns about how the PDP scoring works, which we cover in the discussion.

Overview of the Data Analysis

Here we show, with two simulations and one empirical data set, that the PDP scoring method produces comparable results whether the data is purely random, simulated to be highly correlated and consistent with the theoretical assumption of lower acceptance of C dilemmas, or actual responses from real people. The results suggest that (1) some correlations between parameters are mathematical artifacts that emerge even from randomly generated data, (2) patterns are similar whether the data is randomly or "ideally" simulated, or actual responders, (3) reliabilities for all parameters calculated from PDP are poor, and (4) bipolar scores from PDP and a high-conflict moral dilemma battery (see Koenigs et al., 2007; Greene et al., 2008) might not reflect the same underlying construct. All the data, materials and scripts are available at https://osf.io/vmy4q/.

SIMULATION 1

The purpose of SIMULATION 1 is to show that the correlations between the TS and U or D parameters of the moral PDP are mathematical artifacts. As stated earlier, C&G created moral PDP to be a "compelling test of the dominant dual-process account of moral judgment", and the authors claimed these correlations confirm the existence of the confound between utilitarian and deontological processes. We show by simulation that these correlations would be very similar even when the data is completely random and there are no differences in response trends between the C and IC dilemmas. We argue that they cannot be considered as evidence for such claims.

Methods

We simulated 10,000 responders with completely random response patterns with R (version 3.6.3) for the simulation. In other words, we simulate 10,000 rows of ten binary values with equal probabilities [meaning that $P(X = 0) = 0.5$ and $P(X = 1) = 0.5$] to represent responses for both C and IC dilemmas. A simulated response of 1 indicates that harm in the dilemma is considered unacceptable, and a response of 0 indicates that it is considered acceptable. The standard PDP scoring procedure is then used to compute U, D, and TS from this artificial data.

Results

Correlations between the scores are presented in **Table 1** and graphically in **Figure 1**, with blue dots indicating what we will term "realistic" responses: accepting the harm in IC dilemmas is more likely than in C dilemmas.[5] We see that in completely random data, the U and D parameters exhibit substantial correlations with the unidimensional TS, in the directions observed by C&G and later ourselves in our empirical data. In **Figure 1A**, observations with high TS and low D parameters are absent below the diagonal and scarce on the upper left. This implies that the scoring does not allow any other results than a positive correlation between these two to emerge, provided the answering patterns have variance. Similarly, TS and U have a forced negative linear relationship, presented in **Figure 1B**.

We also observe the correlation between U and D parameters to be very close to zero. Note that the null correlation between U and D depends on variation in both IC and C dilemmas – of these, the variation in C dilemmas is the more theoretically interesting part. Variation in the IC dilemmas is to be expected as they are dilemmas where the two processes proposed by the dual process model drive different responses. In the C dilemmas, the processes are in agreement about a negative response. If responses to C dilemmas were a constant between participants, this would manifest as a strong negative correlation between U and D (see results from modified Simulation 1 in Appendix B). However, the

---

[5]An important assumption of the PDP is that both of the two processes drive similar negative (i.e., not accepting harm) responses in the congruent condition, but different responses in the incongruent condition. A U score below 0 indicates that this assumption has been violated. Moreover, it indicates a participant who seemingly thinks, on average, that causing harm to others is *more acceptable when there is no aggregate benefit* than when there is such a benefit. We argue responses like this in real data would more likely indicate misunderstood task instructions or an unmotivated, malignant or a bot participant.

TABLE 1 | Correlations in completely randomly generated data with process dissociation procedure (PDP) scoring.

	U	*D*	*U*	*D*
D	0.00		0.02	
TS	−0.71***	0.68***	−0.53***	0.82***

*D = PDP deontology parameter, U = PDP utilitarianism parameter, TS = PDP traditional bipolar utilitarianism-deontology score. The left-hand side of the table uses the whole simulated sample. The right-hand side uses only simulated responders with U greater than 0, a "realistic" response pattern (5831 out of 10000). ***p < 0.001.*
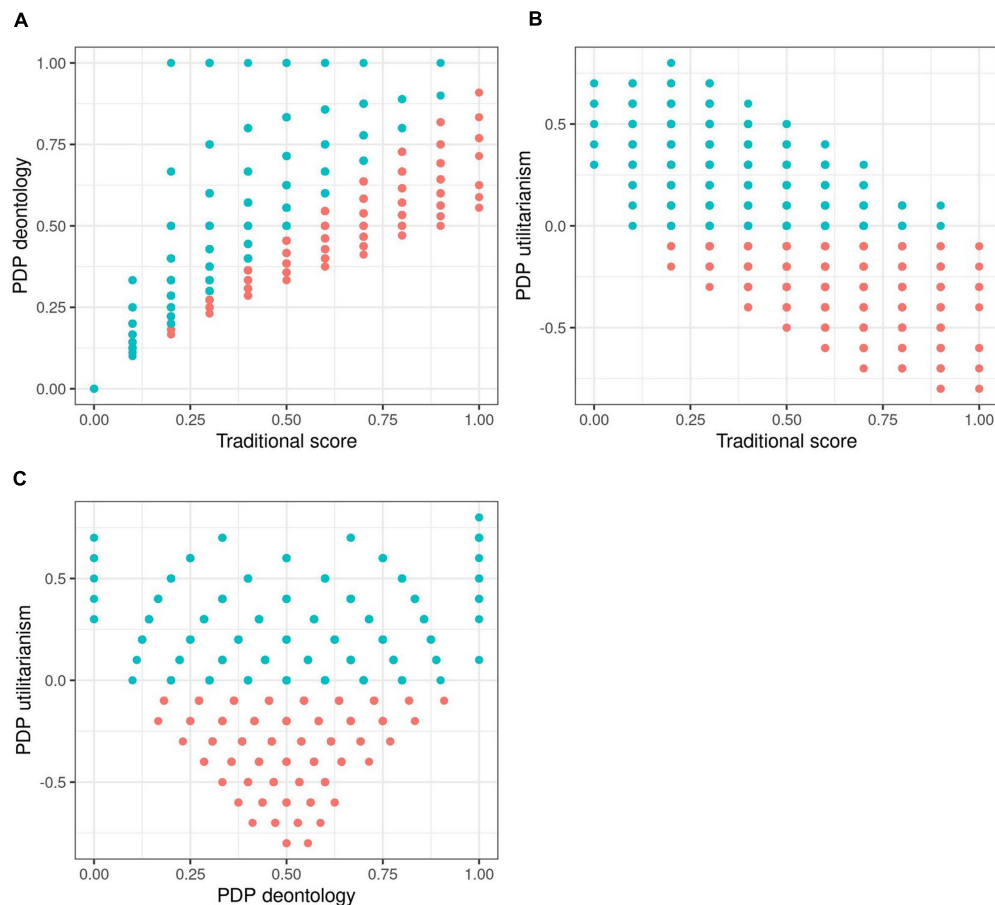
FIGURE 1 | Scatterplots of process dissociation procedure (PDP) scores in randomly generated data absent of any real patterns. Blue dots represent simulated participants with realistic response distributions, with *U* parameter being > 0.

significant correlations between *TS* and the two moral inclination parameters do not depend on variation in the C dilemmas: if *C* is set to be a constant, we would still observe a strong negative correlation between *U* and *TS* and a strong positive correlation between *D* and *TS*.

Since all parameters (including *TS*) are derived from the same sets of dilemmas, there are built-in dependencies between them. Thus, we argue that correlations between *TS* and the other parameters cannot be used as evidence for the separateness of the processes. These correlations would exist as long as there is variance in responses to IC dilemmas, and variance in IC dilemmas in and of itself is not evidence in favor of the dual-process model.

What is really notable in the scatterplots is the fan-shaped joint distribution of *D* and *U* (see **Figure 1C**). Although the correlation between *D* and *U* is zero, there is a clear non-random pattern between the parameters. This dependency is not surprising given *D* is calculated based on the value of *U*. We observe this also with empirical data. Even if there was no linear correlation between *U* and *D*, it is clear from the plot that if some manipulation would affect *U*, it would constrain or relax possible values for *D* as well. When *U* is at 0, *D* can have almost any value; when *U* increases

from 0, the range of possible values for *D* narrow between the extremes. Extreme or midpoint *D* scores can co-occur with a wide range of *U* scores, but, e.g., a *D* score of 0.9 can only co-occur with a *U* score of 0. What follows from the relationship between *U* and *D* is that selectively increasing or decreasing *U* in a group of participants has an effect on the possible distribution of *D* scores. A lower value of *U* implies a wider range of possible values of *D*, despite theory stating the two should be independent. Affecting only *D* requires affecting C and IC dilemmas in the same direction and in the same degree; affecting only *U* is very difficult as *D* is dependent on both *U* and the absolute value of IC dilemmas (see the formulae presented in concern 2 and figures produced by simulation studies). Therefore, affecting a single parameter would require affecting both dilemma types in precisely the right way, and affecting a parameter truly selectively seems unlikely. This then means that it might be difficult to create manipulations that target only one process, and in many cases both are affected even if statistical significance is not achieved.[6]

---

[6]Note that this is an argument about mathematics, that is, about affecting a single process as measured by the instrument. However, there may also be a difficulty in truly affecting only one process conceptually, i.e., in the cognition of a participant,

To summarize, there are several features in the PDP that are artefactual in nature, which are either detrimental to theory-testing or theoretically implausible. These include correlations emerging from the scoring procedure, constraining patterns between different parameters, and a consequent difficulty in manipulating only one type of processing (as measured).

SIMULATION 2

The purpose of Simulation 2 is to examine the same correlation patterns as before, but now in conditions where the data is very consistent. In brief, the results are basically the same as they were in Simulation 1 with purely random data. In Simulation 2, both C and IC dilemmas were simulated to have large correlations within their respective dilemma sets (but not between the dilemma sets). C dilemmas were set to be more unacceptable than IC dilemmas. See **Table 2** and **Figure 2** for the results.

Methods

Again, we simulated 10,000 responders with 20 binary values representing 10 C and 10 IC dilemmas. To ensure that these values would correlate similarly with one another, we initially simulated two sets of 10 correlated gaussian variables (mean = 0 and SD = 1; inter-item $r = 0.50$). We then separately dichotomized these to represent C and IC dilemma batteries by using different cut-off points for each dilemma resulting in different unacceptance rates. First, we randomly sampled the cut-off points for C dilemmas from uniform distribution (min = -1, max = 2). We then sampled the cut-off points for IC similarly, but made sure that the cut-off point for IC dilemma was always lower than for the corresponding C dilemma. We did this by lowering the minimum to -2, and using the previously sampled cut-off point for corresponding C dilemma as the maximum (min = -2, max = corresponding C cut-off point). This ensured that the probability of "unacceptable" responses was always greater for C dilemmas, and resulted in an overall unacceptance rate of 68% for C dilemmas, and 35% for IC dilemmas. We then applied PDP formulae to compute U, D, and TS.

Results

The median correlation for dichotomized C dilemmas was 0.26 and for IC dilemmas 0.20. The correlations between

as there seem to be cases where, e.g., higher empathic concern predicts either more utilitarian or more deontological responding based on the type of moral dilemma (Rosas et al., 2019b).

TABLE 2 | Correlations in simulated data with within-dilemma-set correlations.

	U	D	U	D
D	0.15***		0.11***	
TS	−0.66***	0.57***	−0.59***	0.66***

*D = PDP deontology parameter; U = PDP utilitarianism parameter; TS = PDP traditional bipolar utilitarianism-deontology score; PDP = process dissociation procedure. The left-hand side of the table uses the whole simulated sample. The right-hand side uses only simulated responders with U greater than 0, a "realistic" response pattern (8,850 out of 10,000); *** = p < 0.001.*

the parameters are presented in **Table 2**, and corresponding scatterplots in **Figure 2**. The general pattern is the same: the D and U parameters correlate with TS in the theoretically predicted directions, but have a weak correlation with each other. It would thus seem that the issues we observed with random data are present with simulated consistent responses as well. In practical terms this implies, again, that the correlations between TS and the two parameters are caused by the scoring formulae, and specifically the inclusion of IC.

The range of possible U scores covers the whole spectrum of "realistic" values only when D is at either extreme. This pattern stems from the fact that the probability of accepting the harm in C and IC dilemmas constrains the possible values of the U and D parameters. Namely, 100% acceptance of the harm in IC dilemmas and 0% acceptance of the harm in C dilemmas are the only situations that allow for theoretically realistic responses so that the U parameter is not constrained. 100% acceptance in IC dilemmas corresponds to a D parameter value of 0, and 0% acceptance in C dilemmas corresponds a D parameter value of 1. We wish to emphasize that this constraining is a direct result of the way the PDP scoring works. This may be a lesser issue when using the PDP test for group-average contributions of two separate processes in a task as Jacoby (1991) did, but it leads to issues when using it as an individual differences measure. However, we also find the constraining generally problematic for the theory, given that it means that at face value, the processes as measured by the PDP do not and indeed cannot vary completely independently.

EXAMINATION WITH EMPIRICAL DATA

As established in Simulations 1 and 2, comparing PDP parameters with traditional bipolar scores when all are computed from PDP dilemmas presents a confound because the same items are used for all parameters. Here we examine a dataset collected for other purposes, which contains responses to both the PDP dilemmas and other high-conflict moral dilemmas (HCMDs; Greene et al., 2001; Koenigs et al., 2007). This enables us to examine correlations similar to those presented by C&G and our two simulations, but without the confound between TS and the U and D parameters because a different battery is used for TS. In addition, we use this data to examine the psychometric properties of the moral PDP. To summarize the results, we found low response consistencies in the PDP dilemma sets, which manifest as reliabilities below the accepted standards.

Methods
Participants
A total of 1,043 participants were recruited from Prolific Academic[7] to participate in an online experiment which is to be reported elsewhere. 33 participants were excluded from the sample due to failed attention checks. Of the retained sample, 466

---

[7]http://www.prolific.co/
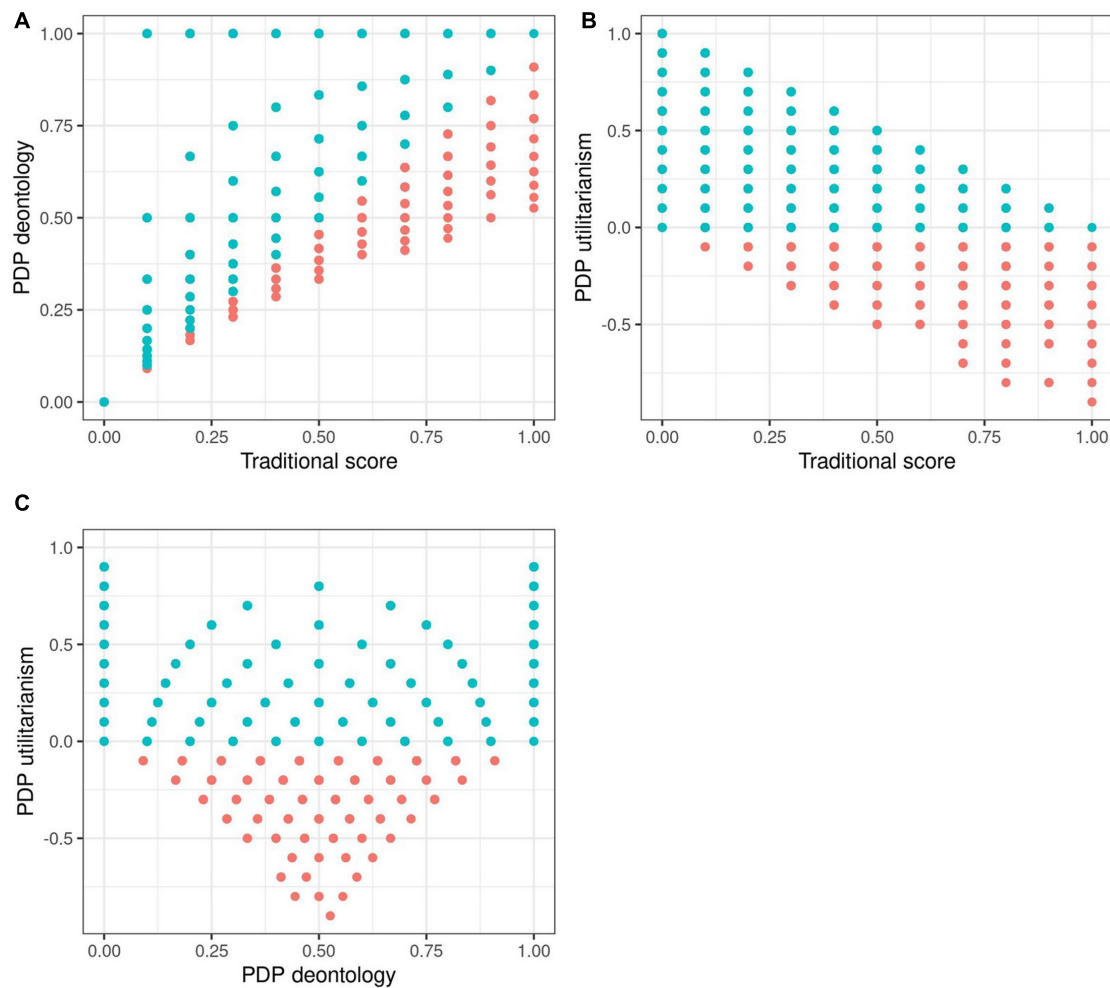
FIGURE 2 | Scatterplots of process dissociation procedure (PDP) scores in simulation 2 simulated data with realistic assumptions. Blue dots represent simulated participants with realistic response distributions, with the *U* parameter being > 0.

were men, 535 women, 4 non-binary, and 5 refused to state their gender. Mean participant age was 37.35 (SD = 13.36).

Materials

PDP moral dilemmas

We presented participants with nine dilemma pairs from the moral PDP (Conway and Gawronski, 2013; we omitted the Crying Baby dilemma pair because a very similar dilemma was in the high-conflict dilemma battery). Each pair described similar situations with different consequences for an action. Participants were asked to indicate if they consider the action unacceptable. In C dilemmas, both consequences and norms encourage inaction; in IC dilemmas, norms encourage inaction while consequences encourage action. Traditional bipolar scores were computed by taking an average of the IC dilemmas. Higher traditional scores represent higher deontological responding and lower utilitarian responding. The *U* and *D* parameters were computed according to formulas from Conway and Gawronski (2013; based on Jacoby, 1991), presented in Simulation 1. Traditional scoring had a

low reliability estimate (Cronbach's alpha = 0.60, McDonald's omega = 0.39; see below for *U* and *D* parameters; corresponding reliability estimates for only C dilemmas were: alpha = 0.71, and omega = 0.56).

High-conflict moral dilemmas

We presented participants with the 12 "high-conflict" (see Koenigs et al., 2007; Greene et al., 2008; see Laakasuo and Sundvall, 2016 for psychometric examination; see also Laakasuo et al., 2017) dilemmas from Greene et al.'s (2004) original dilemma battery: the traditional bipolar deontological/utilitarian scale. We asked participants to rate how acceptable the utilitarian solution to each dilemma was, on a scale from 1 (not at all acceptable) to 7 (totally acceptable), and averaged the responses to compute the bipolar score. Higher scores represent higher utilitarian and lower deontological tendencies (note that this scoring goes the opposite way compared to the PDP traditional score). The scale had a reliability estimate alpha of 0.88, and omega of 0.74.
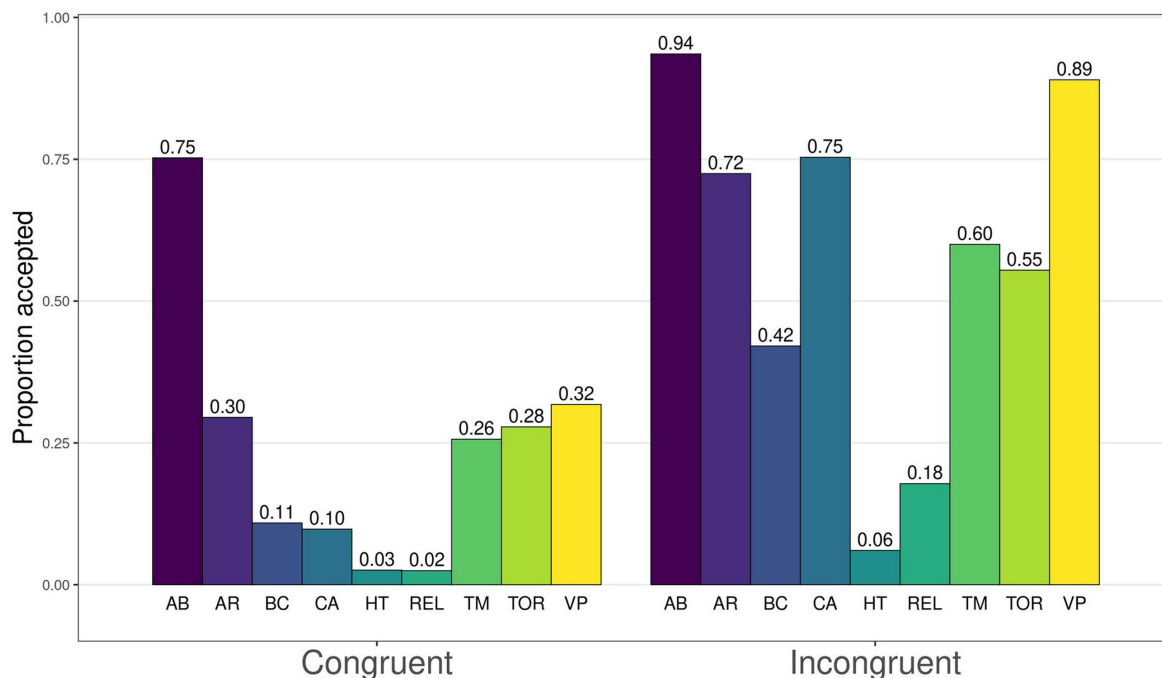
FIGURE 3 | Proportions of accepted harm for congruent and incongruent dilemmas. AB, abortion; AR, animal research; BC, border crossing; CA, car accident; HT, hard times; REL, relationship; TM, time machine; TOR, torture; VP, vaccination policy.

Results

Descriptives

On average, participants accepted the harm 56.8% of the time in IC dilemmas and 23.9% of the time in C dilemmas, broadly replicating C&G's results. Proportions of responses where the participant accepted the harm in a PDP dilemma are presented in **Figure 3** (see Conway and Gawronski, 2013, Appendix A, pp. 231–233, for the content of each dilemma pair). As expected, in all cases harm in C dilemmas was rated unacceptable more often than in IC dilemmas, though with considerable variation in both acceptance rates and the size of the difference within a dilemma pair. We also computed the proportions of participants who either changed or maintained their response (positive or negative) between the C and IC versions of each dilemma (**Table 3**). We did this to provide an overview of the "difficulty"[8] of different dilemmas and potential problematic items. For each dilemma pair, only a small minority (under 5% of participants) gave responses where the C version was found acceptable but the IC version was not, i.e., an obviously problematic response.

In the Car Accident and Vaccine Policy dilemmas, over half of the participants responded "unacceptable" to the C version and "acceptable" to the IC version. Thus, in these dilemmas the congruency manipulation caused a difference for acceptability in a sensible direction for a majority of the sample, suggesting they

were relatively "easy" in terms of utilitarianism (note that it is not necessary for a dilemma pair to have over 50% of participants responding in this way for the dilemma pair to be "good" or appropriate). A further two dilemmas, Animal Research and Time Machine, also had more responses in this category than any other category, but not over half the sample: the responses were more evenly spread between the other response patterns.

In the Border Crossing, Hard Times and Relationship dilemmas, over half of the participants responded "unacceptable" to both versions of the dilemma. In other words, a majority of the participants found the harm unacceptable in these dilemmas whether it had a utilitarian justification or not. Thus, these dilemmas seem to be the "hardest" in terms of utilitarianism. The Torture dilemma also had more "both unacceptable" responses than responses of any other type, but not over half the sample.

It is not necessarily clear why some dilemmas are "easy" or "hard." A majority of "both unacceptable" responses may stem from the relevant deontological norm being very strong. However, it could also be that the utilitarian justification in some of the dilemmas was not very strong, or participants interpreted the dilemma differently from what the developers intended. For example, the utilitarian motivation in the Border Crossing IC scenario is only that a soldier has a *suspicion* that a person approaching a checkpoint intends to bomb the checkpoint and kill an unstated number of people: participants may not find this a clear enough motive to shoot the person. Similarly, a majority of "IC acceptable, C unacceptable" responses may stem from a strong utilitarian motive or a weak deontological norm.

---

[8]Note that we use the word "difficulty" here in a technical sense, referring to how common it is for participants to accept harm in the IC version and reject harm in the C version of a dilemma pair. This is different from subjectively assessed difficulty, i.e., how hard participants may have felt a dilemma was to answer.

TABLE 3 | Proportions of different response patterns to congruent (C) and incongruent (IC) versions of dilemmas.

	AB	AR	BC	CA	HT	REL	TM	TOR	VP
Both unacceptable	0.048	0.260	0.569	0.233	0.928	0.814	0.363	0.436	0.097
Both acceptable	0.736	0.280	0.099	0.085	0.013	0.017	0.220	0.268	0.305
IC acceptable, C unacceptable	0.200	0.444	0.321	0.668	0.048	0.161	0.380	0.286	0.585
IC unacceptable, C acceptable	0.017	0.015	0.010	0.013	0.013	0.008	0.037	0.010	0.013
Difference	0.183	0.429	0.311	0.655	0.034	0.153	0.343	0.276	0.572

AB = Abortion, AR = Animal research, BC = Border crossing, CA = Car accident, HT = Hard times, REL = Relationship, TM = Time machine, TOR = Torture, VP = Vaccination policy, Difference = P(accept|IC) - P(accept|C).

The Abortion dilemma is an outlier, with a large majority (73.6%) of participants responding "acceptable" to both versions of the dilemma. This dilemma seems to be the only "hard" dilemma for rejecting harm in our sample. We suggest this stems either from a majority not actually recognizing a deontological norm against abortion, or from both versions of the dilemma having consequences that people may find worse than or equally as bad as abortion.[9]

In addition to the above, the Animal Research, Time Machine, Torture and Vaccine Policy dilemmas all had over 20% of "both acceptable" responses. For the Animal Research dilemma, we suggest that a norm against animal testing is not nearly as recognized as other deontological norms, and thus even the C version of the dilemma (animal testing for an acne medication) has notable acceptance. For the Vaccine Policy dilemma, we are less sure: accepting the C version of this dilemma means accepting a potentially lethal medicine to a non-lethal case of the flu. This may be a case of participants misinterpreting the dilemma in some way. The Time Machine and Torture dilemmas especially raise questions of validity, given that the C versions of these dilemmas deal with murder and torture, respectively. We argue that at face value, it is not plausible that over 25% of people have deontological and/or utilitarian inclinations that are too weak to condemn murder or torture without a utilitarian motivation. Rather, we think some participants in these two dilemmas may interpret the benefits of the harm (prevention of a child kidnapping for ransom or a bombing that vandalizes private property, respectively) as greater than intended by the developers of the measure.

In sum, in addition to the dilemmas having different levels of "difficulty" (**Figure 3** and **Table 3**), there are several dilemmas where a large number of participants respond in a way that is hard to interpret. Differences in difficulty (inferred from response patterns, not self-reported by participants) are not inherently problematic, but experimenters should be aware of them, and consider what it means to create an average score out of different items without weighing the "easy" and "hard" items differently. At least some of the more problematic responses suggest that some participants may interpret dilemmas differently from what is intended, as suggested by Baron and Goodwin (2020). Given that a null correlation between the U and D parameters

depends on variation between participants in responses to the C dilemmas, issues in interpretation pose a validity problem not only for specific dilemmas. The null correlation, which on its face supports the dual-process model, may arise from variation in C dilemmas that is not caused by moral inclinations but by differences in understanding the items.

Correlations Between PDP Dilemmas

We computed Spearman correlations between all PDP dilemmas to learn about the degree of response consistency (see **Figure 4**). First, in our sample, the median absolute correlation between dilemmas was 0.06, and the average absolute correlation was 0.08. This is concerning because co-variation among items is usually the basis for using them in a measurement scale.

We also observe a similar non-linear relationship between U and D as with the "realistic" responses in Simulations 1 and 2. The "floor" of U was -0.2 in the empirical data (see **Figure 5**), due to a handful of participants (18 out of 1,010) who were more accepting in C than in IC dilemmas. The fact that our empirical data had very few "unrealistic" responses suggests that a majority of responders were generally logical in their responding, i.e., not accepting "useless" harms more than utilitarian ones (see **Figure 4**). Nevertheless, even with real responders, the distribution of U and D shows the theoretically implausible constraining between U and D seen in Simulations 1 and 2. As stated before, at face value, this is hard to reconcile with the claim of independent processes.

Reliability Analysis

Due to the nature of the PDP formulae, computing reliability coefficients (such as Cronbach's alpha or McDonald's omega) from the data is not as straightforward as in most questionnaires. The U scale can be composed as C dilemmas and reverse-coded IC dilemmas minus a constant (see Appendix A for mathematical proof). The constant in the formula should not affect statistical properties other than mean, so we can use this scale in psychometric analyses. This formulation has a Cronbach's alpha of 0.16, and omega 0.08, that is, almost nonexistent (however, see permutation method below). However, if the IC dilemmas are not reversed (contrary to the PDP formula) this reliability coefficient becomes 0.77 (and omega 0.37). This may reflect deontological dispositions driving similar responses across the dilemma types, but either of these coefficients is sub-par. It is unclear what this formulation of the scale would represent: as stated, common psychometric assessments are hard to apply to the PDP.

[9]In the incongruent version, the baby would die anyway if delivered. In the congruent version, it is stated that if the baby is delivered, the mother will spend the rest of her life as a single mother without a job, making life very hard for both her and the child.
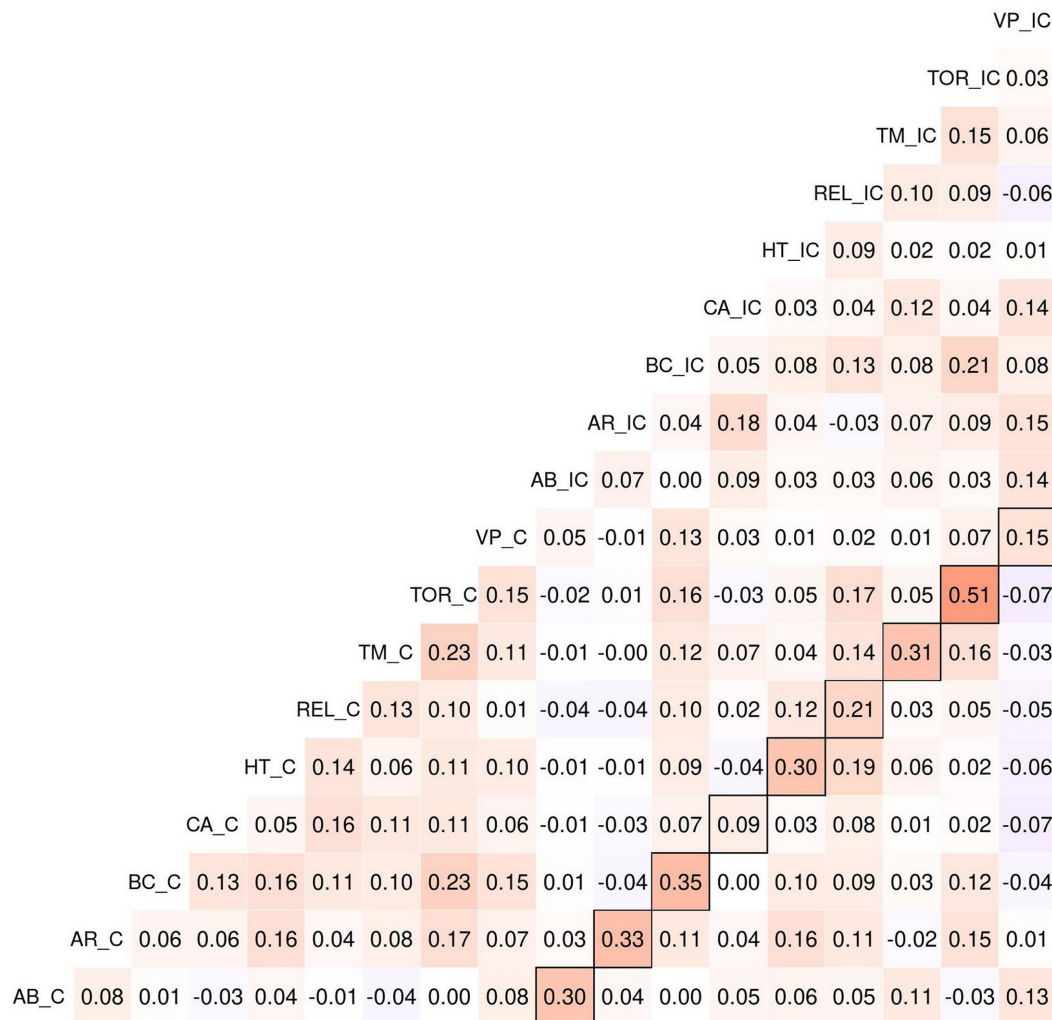
FIGURE 4 | Spearman correlations between process dissociation procedure (PDP) dilemmas in empirical data. Contours indicate correlations for same dilemmas from C and IC dilemma sets. Note that the correlations for the large part are highest between the dilemma pairs and not within the incongruent (IC) and congruent (C) clusters.

Exploratory factor analysis (see Appendix C) for U suggests that almost all dilemmas load positively on a single factor (however, some with very low loadings) rather than IC dilemmas loading negatively. We also used confirmatory factor analysis with DWLS estimation for dichotomous items, but the pattern was essentially the same. If we were to interpret this as a kind of general moral condemnation factor – as deontology drives not accepting the harm in both dilemma types and utilitarianism in the C dilemmas – it is worrying that the loadings of the C dilemmas are not much better than those of the IC dilemmas.[10]

To gauge the reliability of the parameters we also used a permutation approach to split-half reliability for both U and D (as

recommended by Parsons et al., 2019). This method investigates the response consistency of a scale by how similar responses are to the two halves of the same test. We randomly sampled 4 dilemma pairs and computed U and D parameters from these, and did the same for the remaining five dilemma pairs. We then computed the Spearman correlation coefficient between the parameters from the two halves. We iterated this process 10,000 times and then applied the Spearman–Brown correction (see Parsons et al., 2019) to these correlations to take into account underestimation of reliability.

The resulting distribution of corrected estimates is presented in **Figure 6**. These function as direct estimates of reliability for these two scales, and should approximate Cronbach's alpha (see Parsons et al., 2019). The average corrected relationship between the split-halves was 0.30 (SD = 0.05) for U, and 0.33 (SD = 0.07) for D. This reliability estimate for U does not converge well with coefficient alpha calculated above. Still, these results imply serious

---

[10]Similarly to IQ test items, we should be capable of evaluating whether those cognitive tasks (here, moral judgments) load onto the same factor (i.e., that it can be claimed that they measure a similar construct).
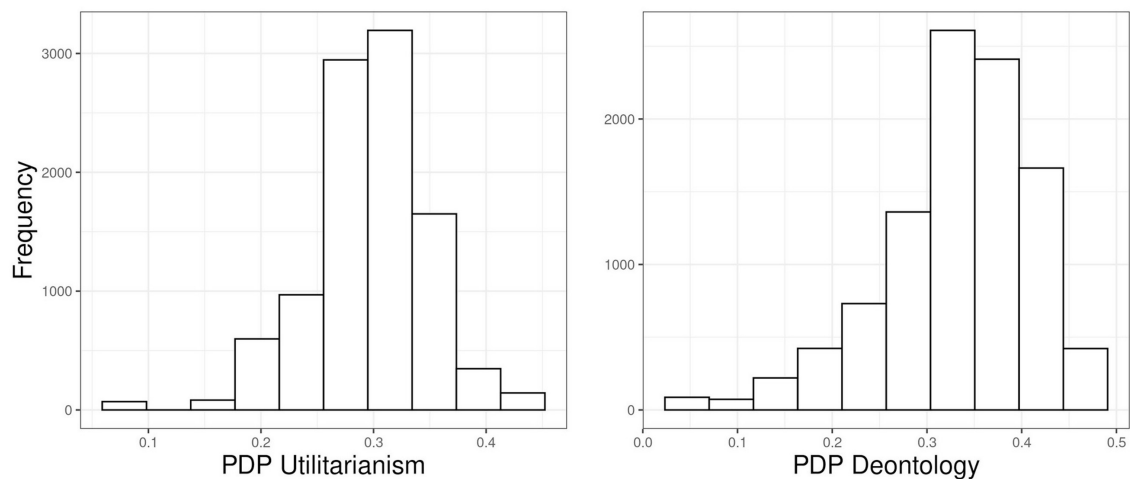
FIGURE 5 | Distributions for reliability estimates for *U* and *D* from split-half permutations.

measurement inconsistencies for both *U* and *D*.[11] For practical purposes, we believe these measures can hardly tap into any common cognitive, motivational or personality factors.

Comparison Between PDP and HCMD

The bipolar scores calculated from the two dilemma batteries (PDP and HCMD) should correlate very highly, as they are intended to be parallel measures of the exact same underlying trait. Additionally, *U* and *D* scores should exhibit similar correlation patterns with the HCMD scores as they do with *TS* computed from the PDP dilemmas.

The correlations are presented in **Table 4** (for scatterplots between PDP scores, see **Figure 6**). We observe correlational patterns that are in the correct direction. These correlations are much lower than within PDP, but still non-trivial. A correlation of 0.51 between the HCMD *TS* and PDP *TS* means that only 26% of reliable variance is shared between the two measures. This implies heterogeneity in the measures that are supposed to reflect the exact same underlying trait, and thus undermining their equivalence. Some of this discrepancy is likely attributable to large measurement error in PDP scores. However, considering both the low reliability of PDP *TS* and this correlation, it seems that we cannot conclude that these two measures appropriately assess the same construct. If they do, and one of the measures has been incorrectly conceived in the past, it is not clear that the problem lies with the HCMD measure, which the moral PDP aimed to improve.

Summary

To summarize the results and implications of the analyses presented above, they indicate the following: first, there are severe response inconsistencies in the PDP dilemma sets that manifest as low correlations inconsistent in sign. Second, all parameters calculated from the PDP dilemma sets have sub-par

reliability when examined with either coefficient alpha, coefficient omega, or the split-half permutation method. Third, there are nevertheless non-trivial correlations to the expected direction between the HCMD bipolar score and the PDP scores. Fourth, the correlation between the two different bipolar scores is still too low to ensure that the two measures tap into the same construct. Based on the results, we are skeptical whether the current formulation of PDP is able to quantify either individual utilitarianism and deontology or their bipolar continuum.

DISCUSSION

In this paper we have presented both theoretical and statistical concerns about the PDP measures of utilitarian and deontological tendencies, and using both simulations and empirical data, examined their psychometric properties. In Simulation 1, we showed that the PDP formulae produce similar correlations between the PDP parameters and the bipolar scoring method as observed in the empirical literature even when the data is randomly generated. In Simulation 2, we showed that the theoretically problematic non-independent distribution of data similar to Simulation 1 is also present when responses are highly reliable and aligned with theory. In our empirical data, we replicated the correlations between parameters found in prior literature and our simulations, and also found issues with reliability and similar distribution patterns as we did in our simulations.

Based on our results, the PDP scoring procedure constrains the possible values for the estimates of utilitarian and deontological inclinations. The scoring also leads to utilitarian and deontological scores that correlate with the traditional bipolar score. It is important to emphasize that these correlations are similar regardless of whether the data are completely random, simulated to be highly internally correlated and theoretically consistent, or actual responses from real people. They are also similar regardless of whether simulated "unrealistic" responses are excluded or not, and whether simulated participants vary

---

[11] In terms of signal-to-noise ratio, a reliability of 0.30 means there is 9% common variance among items, which means, there is 10 times as much noise as there is signal. As a reference, alpha of 0.71 would bring signal to noise ratio to about 1 (meaning that there is as much signal as there is noise).
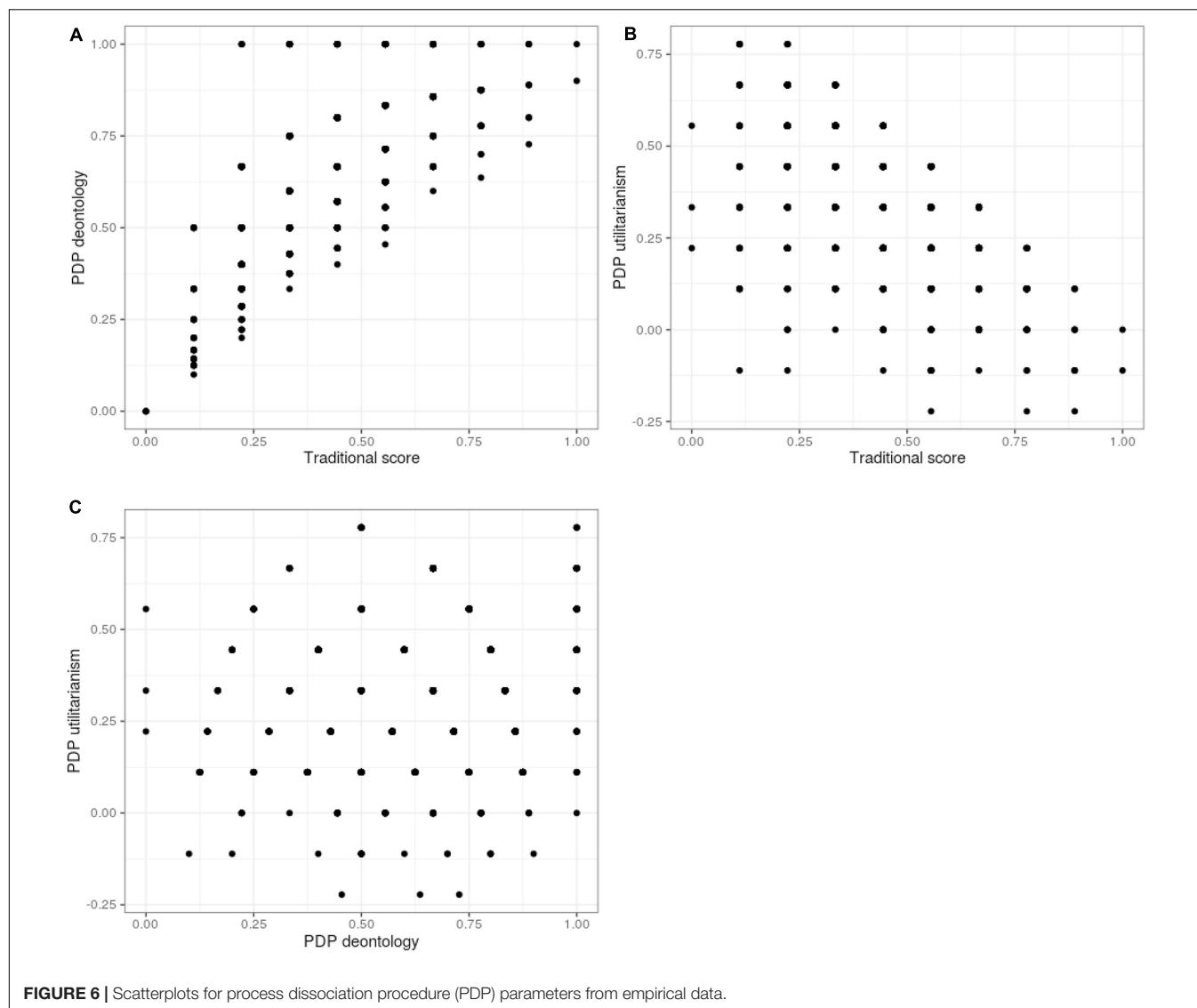
FIGURE 6 | Scatterplots for process dissociation procedure (PDP) parameters from empirical data.

in their responses to the congruent dilemmas or not. All this suggests that such correlations should not be interpreted as sufficient validity evidence for the scale, or an underlying theory. Moreover, in empirical data, the correlation between bipolar scores from the PDP and HCMD batteries is notable but

TABLE 4 | Correlations between process dissociation procedure (PDP) parameters and high-conflict moral dilemma (HCMD) scores.

	U	*D*	**PDP TS**
D	0.12***		
PDP TS	−0.59***	0.70***	
HCMD TS	0.22***	−0.43***	−0.51***

*D = PDP deontology parameter, U = PDP utilitarianism parameter, PDP TS = PDP traditional bipolar utilitarianism-deontology score, HCMD TS = HCMD bipolar utilitarianism-deontology score. Note that HCMD TS scoring is in reverse relative to PDP TS, thus a negative correlation between them is expected. ***p < 0.001.*

low given that the scales should measure the same construct (*r* = 0.51). Correlations above 0.70 are recommended to claim that two instruments measure the same construct (Carlson and Herdman, 2012). This does not mean that the PDP necessarily measures something conceptually radically different from the HCMD battery: the low correlation may be caused by measurement error.

Our empirical data broadly replicates the results reported by C&G with harm in 57% of IC dilemmas 24% of C dilemmas being approved on average. The dilemmas are thus clearly not producing completely random responses from participants, and the difference between the two dilemma types makes theoretical sense. Another argument in favor of this is that in our empirical data we observed only few "unrealistic" responses, i.e., participants more approving of non-utilitarian than utilitarian harm. We do not wish to give the impression that the dilemmas are completely useless. The issue is that despite the theoretically expected difference between C and IC dilemmas, they produce,

at least in our data, very noisy measurements of utilitarian or deontological tendencies within an individual. We argue this noisiness is not due to unmotivated or malignant responders, because as we mentioned previously, our sample had very few "unrealistic" responses. Averaging dilemmas that work very differently (e.g., due to tapping potentially mutually independent norms) produce noisy estimates of inclinations, which is the reason approaches like CFA and SEM are often used.

However, even with different weights on different dilemmas and/or elimination of unclear dilemmas, the PDP scoring itself causes issues. As mentioned above, we also replicated the correlation patterns between the U, D and TS parameters calculated from the PDP dilemmas in our empirical data. In light of the simulations and the examination of parameter reliabilities, our interpretation of these correlations between parameters calculated from PDP dilemmas is that they are mostly artifacts of the scoring procedure. We do not think the correlations between either of the two main parameters and the bipolar score are evidence in favor of the dual-process theory (contra C&G and Friesdorf et al., 2015). The null correlation between U and D could be used as evidence in favor of the dual-process theory, as it does not inevitably follow from the scoring. The lack of a correlation between U and D depends crucially on variation between individuals in responses to the C dilemmas, i.e., variation in what Baron and Goodwin (2020) term "perverse responses" (accepting harm that breaks norms and is not justified by its consequences). Our empirical data agrees with C&G and Friesdorf et al. (2015) in that people do, in fact, vary in their rates of "perverse responses". If responses to the C dilemmas were constant between participants (but not with exactly 0% acceptance, as this would lead to no variance in the D parameter) in addition to variation in the IC dilemmas, the correlation between U and D would be strongly negative. Of course, if responses to the C dilemmas were a constant, there would be little reason to try and dissociate two processes in this way, as then the C dilemmas could simply be dropped from the procedure.

The theoretical issue here is that "perverse responses" may stem from several different factors, not all of them consistent with the theoretical reasoning behind the PDP or similar models, as Baron and Goodwin (2020) argued. That is, participants could have both utilitarian and deontological inclinations that are too weak to condemn the harm, or the norm against the harm could be quite weak – but participants could also simply disagree about the relevant norms, harms and consequences, or read carelessly. For example, does the relatively high acceptance of the congruent version of the Torture dilemma in our sample tell us about weak norms against torture, about disagreement about whether preventing vandalizing of private property is a good enough consequence to justify torture, or simply about misreading the dilemma? Additionally, at face value, it seems to us that there would be differences between some cultural or political groups, such as liberals and conservatives, on some of the dilemmas, regarding whether a norm against, e.g., animal testing or abortion actually exists, or how much of a harm these things are. For the instrument to work as intended, it should be measuring the relative contributions of, e.g., two cognitive processes or personality traits, not cultural effects. Note that

this concern is separate from the wider discussion of whether utilitarianism or deontology measured using dilemma batteries map onto the philosophy of utilitarianism or deontology: our concern is measurement. We believe the HCMD dilemma battery side-steps this issue as a majority of the dilemmas are about causing the death of another human being in order to save others: the norm against killing is quite universal.

Due to the aforementioned issues, we would advise caution when interpreting the results of moral PDP studies. First, low reliabilities can increase risk for both spurious findings and non-findings (Loken and Gelman, 2017). Second, we argue that the constrained distribution patterns between parameters are enough to question results for individual parameters. If selectively affecting either parameter would constrain or relax values the other one could get, there are likely effects for the other parameter as well even if it does not reach statistical significance. Of course, the latter applies to measures derived from PD models more generally if similar scoring procedures are used to compute individual scores. We would advise caution in interpreting correlations between PDP parameters in these cases as their formulae can by themselves create artefactual associations.

Despite the PDP parameters having very low reliabilities, we observed them to have non-trivial correlations with the HCMD score, which can be interpreted in at least two ways. The first one is that the real correlations are large enough to remain notable even after dilution by measurement error. The other one is that there are limitations in our reliability estimation. The permutation method of split-half reliability is supposed to approximate Cronbach's alpha, which acts as a lower-bound of true reliability. Observing very low reliability estimates thus leaves a wide range of possible values for true reliabilities. Thus, we consider the estimates provided here as the bare minimum as we found no other suitable methods for estimating the reliability of D.

We must note that there have been consistent results showing connections between specific moral PDP parameters and external variables such as gender (Friesdorf et al., 2015) or reasoning style (Conway and Gawronski, 2013; Byrd and Conway, 2019; Patil et al., 2020). As we mentioned earlier, an argument in favor of the moral PDP is that it seems to replicate results obtained with the more traditional bipolar measures of utilitarian judgment, and is related to variables measuring, e.g., reasoning in a way that makes sense in light of the dual-process model. However, there is some uncertainty about the dual-process model itself, which raises questions about the extent to which results that make sense in the light of that model support the PDP. Some recent publications question some of the bases of the dual-process model especially when it comes to emotions as a basis of moral judgment (e.g., McAuliffe, 2019; Rosas et al., 2019a,b). Regardless, we argue that at least some of the issues we have brought up here do not stand or fall based on how well a given result with the PDP replicates. Our concerns regarding the distributions and correlations between parameters that are forced by the scoring procedure are not invalidated by well-replicating results. Any relationship between the parameters and external variables does not take away the observation that the constraining between U and D is problematic.

Additionally, we argue that our concern about the reliability of the dilemma sets is not nullified by existing theoretically meaningful correlations between PDP measures and external variables, because the measure may still be noisy. As an analogous point, Gawronski et al. (2017) reported an effect of cognitive load on one of the CNI parameters, but as pointed out by Baron and Goodwin (2020), this seems to be driven by a single item. Given the variance in the PDP dilemmas, something similar may well happen there: individual differences driven mostly by one or two specific dilemmas. Ideally, effects should be assessed on the level of individual dilemmas or dilemma pairs (McGuire et al., 2009). As our empirical data shows (see **Figure 4** for a plot of accepting harm in the C and IC dilemmas), there is great variance between the dilemmas, and it is not *a priori* clear why.

Moreover, even consistent associations between external variables and the whole set of PDP dilemmas may stem from these external factors affecting something else than moral reasoning per se. For example, there could be systematic differences in how carelessly participants respond to the dilemmas, or how often they disagree with the developers of the scale on what counts as a harm, a benefit, or a moral norm. This was pointed out by Baron and Goodwin (2020) in relation to the CNI model and sex differences, but the argument applies to the PDP and to other individual differences as well. For example, an association between the U parameter (but not the D parameter) and success in the cognitive reflection test (Patil et al., 2020) could be because people higher in reflection might read the congruent dilemmas more carefully and give fewer erroneous accepting responses as a result,[12] instead of or in addition to other possible effects. As another example, the foreign language effect, where participants accept harm more often in congruent dilemmas not in their native tongue (see, e.g., Muda et al., 2018) would similarly make sense under the assumption that participants give more "perverse responses" when they misread a dilemma. Almost by definition, any individual difference involves fluctuations in the numbers of positive responses to the congruent dilemmas, which are hard to interpret as they may stem from several different factors. In short, while some effects may replicate well, it is not clear what is being replicated. We do not intend to claim that every well-replicating association between the PDP parameters and theoretically meaningful exogenous variables is an artifact. For example, we find it at face value believable that psychopathic traits are associated with less care for deontological norms (see, e.g., Reynolds and Conway, 2018). In any case, researchers should be careful to make sure that differences in a trait between groups stem from true differences in that trait and not a difference in how a measurement error in an instrument works for those groups.

More generally, while we found that mathematically speaking, things such as factor analysis can be applied to the U parameter of the moral PDP, it is not as clear that this makes sense from a substance perspective. That is, the scoring formula for the U parameter is equivalent to a sum score with reverse-coded IC dilemmas, which makes it possible to apply factor analysis. However, on the substance level, the items in the U parameter are each supposed to measure two separate latent variables, with a difference between the C and IC items in the way they load onto these variables. We are not currently aware of a method like factor analysis that would allow for confirming a structure like this. We have tried to provide a variety of approaches to assessing reliability, but due to the PDP approach being very different from measures psychometrics usually deals with, some of the analyses presented here may not be appropriate for assessing how well the utilitarianism/deontology PDP works. However, if this is the case, we are simply left in the dark: we do not even know how to assess whether moral dilemmas meant to measure utilitarian and deontological thinking do so in a consistent manner.

Notwithstanding the limitations, this paper has made novel methodological contributions in the psychometrics of utilitarianism and deontology. First, we have mathematically shown that the U parameter of the moral PDP can be evaluated using psychometric methods. Second, we have demonstrated that estimated measurement accuracy for both U and D can be assessed with the split-half permutation method. Moreover, we have shown that the PDP scoring formulae can produce very high artefactual correlations that can be misinterpreted as evidence for the dual-process theory.

CONCLUSION

Conway and Gawronski (2013) raised an important methodological issue within moral psychology, which we believe is still very relevant. We have tried to elucidate in this paper why we think this issue cannot be solved with the PDP in its current form. We found several pieces of validity evidence either lacking or artefactual, or when investigated, insufficient. This suggests that either revisions to the model or novel methodologies are required to appropriately test for the existence of two separate processes and measure them.

DATA AVAILABILITY STATEMENT

The dataset, along with other materials, can be found online at https://osf.io/vmy4q/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Helsinki Ethical Review Board in Humanities and Social and Behavioral Sciences. The

---

[12]Holding a hypothetical participant's responses to IC dilemmas constant, and assuming only "realistic" responses, i.e., no negative U parameter, an increase in responding negatively to C dilemmas increases both the U and D parameters, but not to the same degree. The increase in U is always a linear increase of 0.1 for each additional negative response to a C dilemma (assuming 10 dilemma pairs), but the increase in D can be lower or higher than 0.1. Assuming a "realistic" response pattern, a move from the lowest "realistic" rate of negative responses to C dilemmas (given a specific rate of negative responses to IC dilemmas) to the highest will by definition increase U more than it does D, and could thus more likely lead to a significant effect on U but not D.

patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AK and ML conceived the initial study idea and initially analyzed the data. AK and JS drafted the first study and created simulations. ML collected the data. AK prepared the online materials. All authors contributed significantly to the improvement of the manuscript.

FUNDING

REFERENCES

Baron, J., and Goodwin, G. P. (2020). Consequences, norms, and inaction: a critical analysis. *Judgm. Decis. Mak.* 15, 421–442.

Białek, M., Paruzel-Czachura, M., and Gawronski, B. (2019). Foreign language effects on moral dilemma judgments: an analysis using the CNI model. *J. Exp. Soc. Psychol.* 85:103855. doi: 10.1016/j.jesp.2019.103855

Bostyn, D. H., De Keersmaecker, J., Van Assche, J., and Roets, A. (2020). Bright mind, moral mind? Intelligence is unrelated to consequentialist moral judgment in sacrificial moral dilemmas. *Psychon. Bull. Rev.* 27, 392–397. doi: 10.3758/s13423-019-01676-9

Byrd, N., and Conway, P. (2019). Not all who ponder count costs: arithmetic reflection predicts utilitarian tendencies, but logical reflection predicts both deontological and utilitarian tendencies. *Cognition* 192:103995. doi: 10.1016/j.cognition.2019.06.007

Calanchini, J., Rivers, A. M., Klauer, K. C., and Sherman, J. W. (2018). Multinomial processing trees as theoretical bridges between cognitive and social psychology. *Psychol. Learn. Motiv.* 69, 39–65. doi: 10.1016/bs.plm.2018.09.002

Carlson, K. D., and Herdman, A. O. (2012). Understanding the impact of convergent validity on research results. *Organ. Res. Methods* 15, 17–32. doi: 10.1177/1094428110392383

Christensen, J. F., and Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: a principled review. *Neurosci. Biobehav. Rev.* 36, 1249–1264.

Conway, P., and Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *J. Pers. Soc. Psychol.* 104, 216–235. doi: 10.1037/a0031021

Cook, D. A., and Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *Am. J. Med.* 119, 166.e7–e16. doi: 10.1016/j.amjmed.2005.10.036

Curran, T., and Hintzman, D. L. (1995). Violations of the independence assumption in process dissociation. *J. Exp. Psychol. Learn. Mem. Cogn.* 21, 531–547. doi: 10.1037/0278-7393.21.3.531

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., and Nadarevic, L. (2009). Multinomial processing tree models: a review of the literature. *J. Psychol.* 217, 108–124. doi: 10.1027/0044-3409.217.3.108

Everett, J. A., and Kahane, G. (2020). Switching tracks? Towards a multidimensional model of utilitarian psychology. *Trends Cogn. Sci.* 24, 124–134. doi: 10.1016/j.tics.2019.11.012

Flake, J. K., and Fried, E. I. (2019). Measurement schmeasurement: questionable measurement practices and how to avoid them. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/hs7wm

Flake, J. K., Pek, J., and Hehman, E. (2017). Construct validation in social and personality research: current practice and recommendations. *Soc. Psychol. Pers. Sci.* 8, 370–378. doi: 10.1177/1948550617693063

Friesdorf, R., Conway, P., and Gawronski, B. (2015). Gender differences in responses to moral dilemmas: a process dissociation analysis. *Pers. Soc. Psychol. Bull.* 41, 696–713. doi: 10.1177/0146167215575731

Fumagalli, M., Ferrucci, R., Mameli, F., Marceglia, S., Mrakic-Sposta, S., Zago, S., et al. (2010). Gender-related differences in moral judgments. *Cogn. Process.* 11, 219–226.

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., and Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: the CNI model of moral decision-making. *J. Pers. Soc. Psychol.* 113, 343–376.

Gawronski, B., Conway, P., Armstrong, J. B., Friesdorf, R., and Hütter, M. (2016). "Understanding responses to moral dilemmas," in *The Social Psychology of*

Morality, 1st Edn, eds J. P. Forgas, L. Jussim, and P. A. M. Van Lange (Abingdon: Routledge), 91–110. doi: 10.4324/9781315644189-6

Graf, P., and Komatsu, S.-I. (1994). Process dissociation procedure: handle with caution! *Eur. J. Cogn. Psychol.* 6, 113–129. doi: 10.1080/09541449408520139

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108. doi: 10.1126/science.1062872

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends Cogn. Sci.* 11, 322–323. doi: 10.1016/j.tics.2007.06.004

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition* 107, 1144–1154. doi: 10.1016/j.cognition.2007.11.004

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44, 389–400.

Hedge, C., Powell, G., and Sumner, P. (2018). The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* 50, 1166–1186. doi: 10.3758/s13428-017-0935-1

Hussey, I., and Hughes, S. (2020). Hidden invalidity among fifteen commonly used measures in social and personality psychology. *Adv. Methods Pract. Psychol. Sci.* 3, 166–184.

Jacoby, L. L. (1991). A process dissociation framework: separating automatic from intentional uses of memory. *J. Mem. Lang.* 30, 513–541. doi: 10.1016/0749-596X(91)90025-F

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* 446, 908–911. doi: 10.1038/nature05631

Laakasuo, M., and Sundvall, J. (2016). Are utilitarian/deontological preferences unidimensional? *Front. Psychol.* 7:1228. doi: 10.3389/fpsyg.2016.01228

Laakasuo, M., Sundvall, J., and Drosinou, M. (2017). Individual differences in moral disgust do not predict utilitarian judgments, sexual and pathogen disgust do. *Sci. Rep.* 7:45526. doi: 10.1038/srep45526

Loken, E., and Gelman, A. (2017). Measurement error and the replication crisis. *Science* 355, 584–585. doi: 10.1126/science.aal3618

Mata, A. (2019). Social metacognition in moral judgment: decisional conflict promotes perspective taking. *J. Pers. Soc. Psychol. Soc. Psychol.* 117, 1061–1082. doi: 10.1037/pspa0000170

McAuliffe, W. H. B. (2019). Do emotions play an essential role in moral judgments? *Think. Reason.* 25, 207–230. doi: 10.1080/13546783.2018.1499552

McGuire, J., Langdon, R., Coltheart, M., and Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *J. Exp. Soc. Psychol.* 45, 577–580.

Muda, R., Niszczota, P., Białek, M., and Conway, P. (2018). Reading dilemmas in a foreign language reduces both deontological and utilitarian response tendencies. *J. Exp. Psychol. Learn. Mem. Cogn.* 44, 321–326. doi: 10.1037/xlm0000447

Parsons, S., Kruijt, A.-W., and Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive behavioural measurements. *Adv. Methods Pract. Psychol. Sci.* 2, 378–395. doi: 10.1177/2515245919879695

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., et al. (2020). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *J. Pers. Soc. Psychol.* doi: 10.1037/pspp0000281

Payne, B. K., and Bishara, A. J. (2009). An integrative review of process dissociation and related models in social cognition. *Eur. Rev. Soc. Psychol.* 20, 272–314. doi: 10.1080/10463280903162177

Reynolds, C. J., and Conway, P. (2018). Not just bad actions: affective concern for bad outcomes contributes to moral condemnation of harm in moral dilemmas. *Emotion* 18:1009.

Rosas, A., Bermudez, J. P., and Aguilar-Pardo, D. (2019a). Decision conflict drives reaction times and utilitarian responses in sacrificial dilemmas. *Judgm. Decis. Mak.* 14, 555–564.

Rosas, A., Viciana, H., Caviedes, H., and Arciniegas, A. (2019b). Hot utilitarianism and cold deontology: insights from a response patterns approach to sacrificial and real world dilemmas. *Soc. Neurosci.* 14, 125–135.

Russo, R., Cullis, A. M., and Parkin, A. J. (1998). Consequences of violating the assumption of independence in the process dissociation procedure: a word fragment completion study. *Mem. Cogn.* 26, 617–632. doi: 10.3758/BF0321 1382

Sherman, J. W., Krieglmeyer, R., and Calanchini, J. (2014). "Process models require process measures," in *From Dual-Process Theories of the Social Mind*, 1st Edn, eds J. W. Sherman, B. Gawronski, and Y. Trope (New York, NY: The Guilford Press), 121–138.

Wilkinson, L., and Task Force on Statistical Inference APA Board of Scientific Affairs (1999). Statistical methods in psychology journals. *Am. Psychol.* 54, 594–604.

Williams, R. H., and Zimmerman, D. W. (1989). Statistical power analysis and reliability of measurement. *J. Gen. Psychol.* 116, 359–369. doi: 10.1080/00221309.1989.9921123

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

APPENDIX A

Algebraic Proof for the Alternative Expression for the U Formula

For any individual i, the formula can be expressed as:

$$1. \quad U_i = \bar{x}_{C_i} - \bar{x}_{IC_i}$$

Opening the formula for the latter average gives:

$$2. \quad U_i = \bar{x}_{C_i} - \frac{1}{k}\left(x_{IC_{i1}} + x_{IC_{i2}} + \ldots + x_{IC_{ik}}\right)$$

By reversing the sign in each IC dilemma response the formula becomes:

$$3. \quad U_i = \bar{x}_{C_i} + \frac{1}{k}\left(-x_{IC_{i1}} - x_{IC_{i2}} - \ldots - x_{IC_{ik}}\right)$$

By substituting $-x_{IC_{ij}} = -1 + \left(1 - x_{IC_{ij}}\right)$ for each incongruent dilemma k:

$$4. \quad U_i = \bar{x}_{C_i} + \frac{1}{k}\left(\left(-1 + \left(1 - x_{IC_{i1}}\right)\right) + \left(-1 + \left(1 - x_{IC_{i2}}\right)\right) + \ldots + \left(-1 + \left(1 - x_{IC_{ik}}\right)\right)\right)$$

Opening one layer of parentheses we get

$$5. \quad U_i = \bar{x}_{C_i} + \frac{1}{k}\left(-1 + \left(1 - x_{IC_{i1}}\right) - 1 + \left(1 - x_{IC_{i2}}\right) + \ldots - 1 + \left(1 - x_{IC_{ik}}\right)\right)$$

All -1 terms inside parentheses sum to -k:

$$6. \quad U_i = \bar{x}_{C_i} + \frac{1}{k}\left(-k + \left(1 - x_{IC_{i1}}\right) + \left(1 - x_{IC_{i2}}\right) + \ldots + \left(1 - x_{IC_{ik}}\right)\right)$$

-k can be taken out of the parentheses, where it becomes -1 as it is multiplied by $\frac{1}{k}$:

$$7. \quad U_i = \bar{x}_{C_i} + \frac{1}{k}\left(\left(1 - x_{IC_{i1}}\right) + \left(1 - x_{IC_{i2}}\right) + \ldots + \left(1 - x_{IC_{ik}}\right)\right) - 1$$

Let's notate $x_{IC_{Rik}} = 1 - x_{IC_{Rk}}$ as reverse-coded item k for person i

$$8. \quad U_i = \bar{x}_{C_i} + \frac{1}{k}\left(x_{IC_{Ri1}} + x_{IC_{Ri2}} + \ldots + x_{IC_{Rik}}\right) - 1$$

Finally we can express the second term as an average again.

$$9. \quad U_i = \bar{x}_{C_i} + \bar{x}_{IC_{Ri}} - 1$$

Therefore U is the sum of average congruent response and average reverse-coded incongruent response minus a constant.

APPENDIX B

Results From Simulation 1 When C Is Held Constant at 0.75

TABLE B1 | Correlations between process dissociation procedure (PDP) parameters when congruent (C) is held constant at 0.75 and incongruent (IC) is random.

	U	D	U	D
D	−0.94***		−0.95***	
TS	−1***	0.94***	−1***	0.95***

D = PDP deontology parameter, U = PDP utilitarianism parameter, TS = PDP traditional bipolar utilitarianism-deontology score. The left-hand side of the table uses the whole simulated sample. The right-hand side uses only simulated responders with U greater than 0, a "realistic" response pattern (9405 out of 10000). ***p < 0.001.

APPENDIX C

Factor Analysis for U Parameter Scale

Dilemma	Loading
Abortion (C)	0.06
Animal Research (C)	0.31
Border Crossing (C)	0.40
Car Accident (C)	0.21
Hard Times (C)	0.30
Relationship (C)	0.26
Time Machine (C)	0.37
Torture (C)	0.57
Vaccine Policy (C)	0.24
Abortion (IC)	0.05
Animal Research (IC)	0.10
Border Crossing (IC)	0.41
Car Accident (IC)	0.10
Hard Times (IC)	0.23
Relationship (IC)	0.33
Time Machine (IC)	0.22
Torture (IC)	0.47
Vaccine policy (IC)	0.01

Factor analysis was conducted on tetrachoric correlation matrix due to variables being binary. No variables were reversed for this analysis. 9% variance was accounted for by the model; RMSR = 0.07.

APPENDIX D

Confirmatory Factor Analysis for U

Dilemma	CFA Loading
Abortion (C)	0.06
Animal research (C)	0.32
Border crossing (C)	0.35
Car accident (C)	0.18
Hard times (C)	0.19
Relationship (C)	0.18
Time machine (C)	0.38
Torture (C)	0.64
Vaccine policy (C)	0.21
Abortion (IC)	0.06
Animal research (IC)	0.17
Border crossing (IC)	0.39
Car accident (IC)	0.10
Hard times (IC)	0.14
Time machine (IC)	0.24
Torture (IC)	0.55
Relationship (IC)	0.28
Vaccine policy (IC)	0.03

DWLS estimation (for dichotomous and categorical variables) was used in this confirmatory factor analysis (CFA). CFI: 0.65.

Understanding the Multidimensional and Dynamic Nature of Facial Expressions Based on Indicators for Appraisal Components as Basis for Measuring Drivers' Fear

Meng Zhang*, Klas Ihme, Uwe Drewitz and Meike Jipp

Institute of Transportation Systems, German Aerospace Center/Deutsches Zentrum für Luft- und Raumfahrt (DLR), Braunschweig, Germany

Facial expressions are one of the commonly used implicit measurements for the in-vehicle affective computing. However, the time courses and the underlying mechanism of facial expressions so far have been barely focused on. According to the Component Process Model of emotions, facial expressions are the result of an individual's appraisals, which are supposed to happen in sequence. Therefore, a multidimensional and dynamic analysis of drivers' fear by using facial expression data could profit from a consideration of these appraisals. A driving simulator experiment with 37 participants was conducted, in which fear and relaxation were induced. It was found that the facial expression indicators of high novelty and low power appraisals were significantly activated after a fear event (high novelty: $Z = 2.80$, $p < 0.01$, $r_{contrast} = 0.46$; low power: $Z = 2.43$, $p < 0.05$, $r_{contrast} = 0.50$). Furthermore, after the fear event, the activation of high novelty occurred earlier than low power. These results suggest that multidimensional analysis of facial expression is suitable as an approach for the in-vehicle measurement of the drivers' emotions. Furthermore, a dynamic analysis of drivers' facial expressions considering of effects of appraisal components can add valuable information for the in-vehicle assessment of emotions.

Keywords: fear, facial expression, action units, in-vehicle, component process model

INTRODUCTION

Over the past decade, affective computing came into the focus of research for driver monitoring systems, because some emotions are supposed to impact drivers' cognitive capabilities necessary for driving and risk perception (Jeon et al., 2011). Therefore, detecting and mitigating driver emotions by using affective computing in an emotion-aware system may ensure driving safety (Ihme et al., 2019). One idea of such a system is to interpret the user's emotional state and provide assistance to support users to reduce the negative consequences of certain emotional states (Klein et al., 2002; Tews et al., 2011; Jeon, 2015; Löcken et al., 2017; Ihme et al., 2018). Furthermore, in the context of high-level automated driving functions, an automated assessment of emotions could allow adapting driving styles or warnings to the drivers' current emotional state to maximize drivers' comfort and optimize the driving experience (Techer et al., 2019). For instance, fear, which refers to the emotional responses evoked by processing threatening stimuli (Schmidt-Daffy et al., 2013), can be

regarded as an indicator of experienced risk (Fuller, 1984). Hence, the recognition of fear could help automated driving functions to adapt their speed to reduce the feeling of risk and regain trust on it. However, theoretically, emotions usually have been regarded as a static state rather than a dynamic process (Scherer, 1984, 2019). Accordingly, emotions' time courses and the underlying mechanism have been barely focused in practical applications, especially in the driving context, in which a dynamic interpretation of drivers' spontaneous emotion is required. Interestingly, a recent study confirmed that the recognition of emotions from dynamic facial expressions was more accurate than from static ones (Namba et al., 2018), which suggests considering the multidimensional and dynamic nature of emotions for affective computing may increase the possibility for a practical implementation of emotion-aware systems. Therefore, investigating the multidimensional and dynamic nature of drivers' emotion would contribute to the development of reliable in-vehicle emotion measurement.

The Component Process Model (CPM) provides a comprehensive theoretical framework for the multidimensional and dynamic interpretation of emotions (Scherer, 1984, 2009a; Scherer et al., 2019). According to the CPM, a given situation would be appraised with multidimensional criteria (the appraisal components), which would follow a fixed order. Furthermore, the result of the individual's appraisals would impact the different components autonomic physiology, action tendencies, motor expressions and subjective feeling (Scherer, 2009b). There are four main appraisal components, which were supposed to happen in sequence: 1, novelty, which means, how sudden or unfamiliar the individual perception of the given situation is; 2, pleasantness, which represents positive or negative feelings about the given situation; 3, goal significance, which represents the impact of the situation on an individual goal; 4, coping potential/power, which represents whether the situation is controllable. Particularly, the appraisal components of pleasantness and power are assumed to be the determinants of valence and power suggested by dimensional emotion theorists (Scherer et al., 2019). The appraisal component of novelty, on the other hand, is suggested to be an additional dimension to the dimensional emotion space (Scherer, 2009b). In the assumption of the CPM, the results of these appraisal components would specifically impact the autonomic nervous system (e.g., changing in heart rate) and somatic nervous system (e.g., changing in facial expressions or voice) (Scherer et al., 2019). Thus, multidimension and dynamics in facial expressions and autonomic nervous system activity can be used as indicators for the presence of certain appraisal processes rendering multidimensional and dynamic interpretation of emotion possible.

Implicit measurements are required for the in-vehicle affective computing. In previous studies, drivers' emotions have been assessed by using voice (Abdic et al., 2016) or facial temperature (Zhang et al., 2019). Besides, facial expressions were one of the commonly used implicit measurements of drivers' emotion (Malta et al., 2011; Abdic et al., 2016; Ihme et al., 2018). Specifically, camera-based approaches of facial expression analysis appear suitable for in-vehicle emotion collection,

because these are contactless and unobtrusive. However, up to now, approaches for in-vehicle assessment of emotions based on facial expressions neglect the time courses of and the mechanisms underlying the facial expressions. According to the CPM, it is assumed that the occurrence of a facial expression is a sequential-cumulative process, which is triggered by appraisal components in sequence (Scherer et al., 2018). For instance, fear can be interpreted as an emotion with high novelty and low power (Scherer et al., 2018). Thus, a fearful facial expression may firstly consist of a raised eyebrow representing "unpredictable" and then a dropped jaw representing "out of control." The Facial Action Coding System (FACS, Ekman and Friesen, 1978; Ekman et al., 2002) can be used to describe facial expressions systematically based on activity in atomic units of facial action, the action units (AUs). Interestingly, a recent paper by Scherer et al. (2018) integrates empirical evidence to determine the relationship between activation in certain AUs and appraisal components based on the FACS (see **Table 1**). To add, in a facial electromyography (EMG) study by Gentsch et al. (2015), corrugator and frontalis regions were revealed to indicate goal significance, while activity in the cheek was supposed to be influenced by coping potential/power. Furthermore, the study suggested that appraisal components drive facial expressions in a fixed sequence and that the effects of power appraisal follow goal significance. Still, besides the work by Gentsch et al. (2015), the empirical evidence for this approach is scarce, so that it needs to be verified especially for the assessment of emotions in applied settings. Therefore, the aim of this study was to investigate whether multidimensional analysis of facial expression is a suitable approach for the in-vehicle measurement of the drivers' emotions. Furthermore, the possibility of dynamic analysis of drivers' facial expressions considering effects of appraisal components is investigated. In this study, we chose fear as the target emotion, because recognition of fear could help to adapt the driving style of automated vehicles to reduce the subjective feeling of risk. In order to present the distinct time difference between appraisals, we focused on the first and last appraisal components results: high novelty and low power to reveal the dynamic process in facial expressions of fear. For this, we induced fear as experimental condition and relaxation as control condition in a realistic driving simulation and extracted participants' facial expressions from camera recordings. Based on the aforementioned considerations, we assumed that the activation in specific AUs (1, 2, 4, 5, and 7) indicates high novelty and low power. We also assumed that activation in specific AUs (15, 20 25, and 26) indicating low power follow the activation in AU indicating high novelty.

MATERIALS AND METHODS

Design

The two target emotional states (fear and relaxation) were induced during two automated driving scenarios in a within-participants design. We assessed participants' facial muscle activity from camera recordings based on the FACS.

TABLE 1 | On basis of CPM prediction of appraisal and action units (AU) for fear (adapted table from Scherer et al., 2018).

Cumulative sequence of appraisal	Appraisal components	Appraisal predicted results	AUs predicted
1	Novelty	high	1, 2, 4, 5, 7, 26, 38
2	Pleasantness	open	
3	Goal significance	high	4, 7, 23, 17
4	Power	low	1, 2, 5, 15, 20, 25, 26

AU: 1 inner brow raiser, 2 outer brow raiser, 4 brow lowerer, 5 upper lid raiser, 7 lid tightener, 15 lip corner depressor, 17 chin raiser, 20 lip stretcher, 25 lips part, 26 jaw drop, 38 nostril dilator.

Participants

In total, 50 volunteers took part in this driving simulator study. All of them had a Western European cultural background, lived in Northern Germany and had German as first language. Thirty-seven of these [thirteen females, age range from 18 to 62 years, mean (M) = 31 years, standard deviation (SD) = 11 years] completed the relevant emotion-induction experimental sessions and their faces were validly recorded on camera, so that they could be included into the data analyses. Thirteen participants were excluded because of incomplete self-report questionnaires (three) and due to technical problems with the face detection from the video signals (ten).

Before the start of the study, the participants were informed about the video recording, potential risks of driving in simulators (e.g., the experience of simulator sickness) according to the simulator safety concept and the rough duration of the experiment. The participants were informed that they could take a break or abort their participation at any time. All participants provided written informed consent to take part in the study and the video recording. As reimbursement for their time, the participants received 10 € (12 $) per commenced hour for their participation. After finishing, the participants were informed about the true goal of the experiment (evoking certain emotions) and the necessity to conceal this goal with a cover story (see below).

Set-Up

The study took place in the DLR's Virtual Reality (VR) laboratory consisting of a realistic 360° projection and steering wheel as well as gas and brake pedals. Video data of the participants' faces were recorded from the front with a network camera (Abus, Wetter, Germany) with a frame rate of 15 frames per second and a resolution of 1,280 × 720 pixels. In order to reduce the influence of changing light and ensure constant lighting, an LED band was mounted above participants' head. The driving simulation was realized using the Virtual Test Drive software from Vires (Vires Simulationstechnologie, Bad Aibling, Germany).

Procedure and Scenario

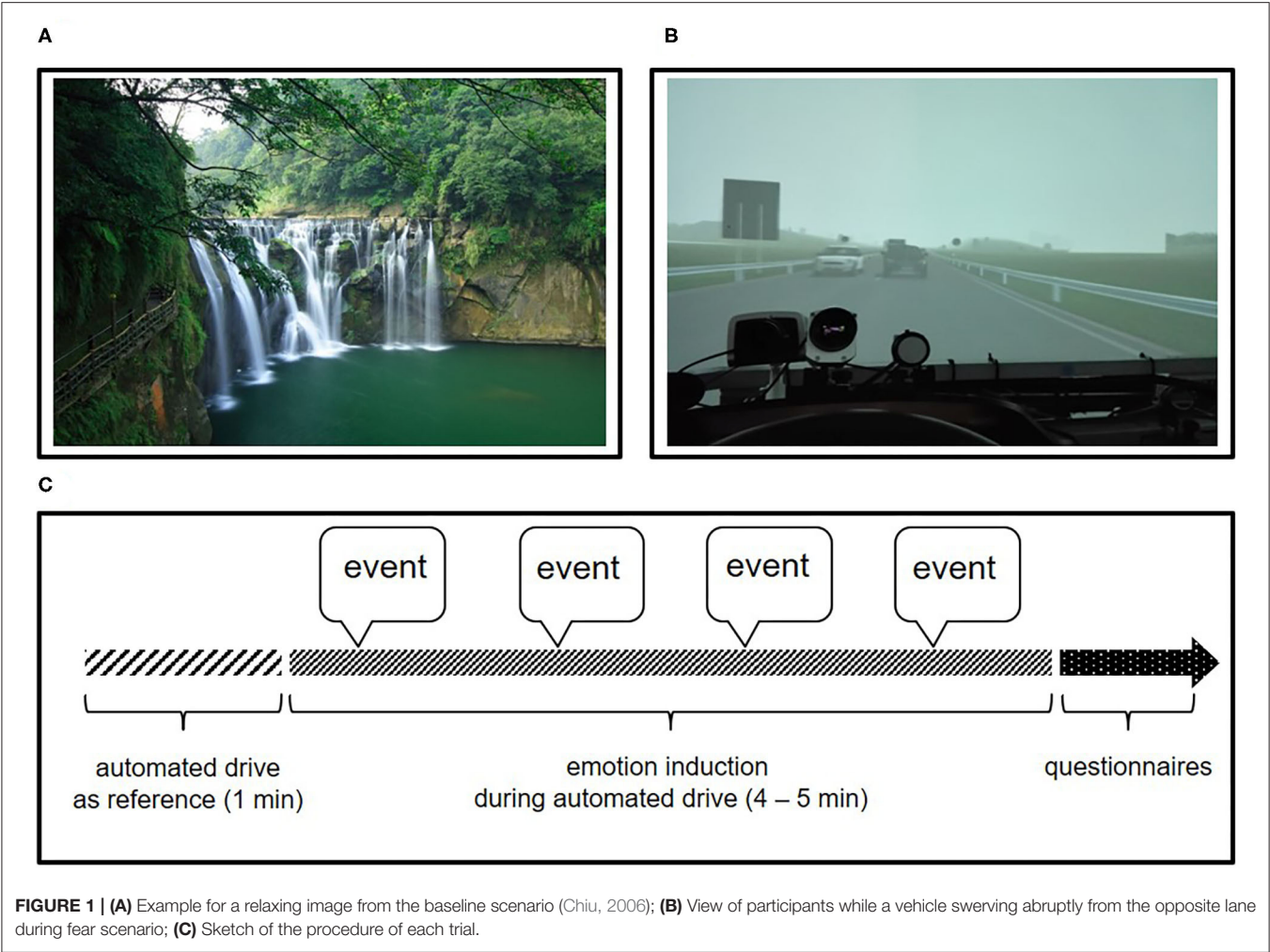With the instructions for the experiment, the participants were presented with a cover story which was supposed to obscure the true background of the study (induction and measuring of emotion). According to the cover story, the aim of the study was to investigate the influence of secondary tasks during an automated drive on the driving performance during a subsequent manual drive. Six emotional states (fear, frustration, joy, sadness, surprise and uncertainty) and relaxation were to be induced in the experiment and the respective tasks to trigger the emotions were included into the cover story. During the experiment, the participant sat alone in the cockpit of the vehicle mock-up and experienced all emotion induction phases in sequence (during the breaks between the drives, the participants had contact to the experimenter). The presentation order of the emotion induction phases was randomized across the participants to reduce the impact of potential ordering effect. For the induction, an automated driving scenario was used, in which the participants where driven by the car in automated driving mode at a given speed along a given route. The drive always started 5 s after the corresponding scenario was activated, with the drive going smoothly for the first minute without any emotional event happening. This period was used to collect a reference for the emotion induction afterwards. Then, the emotional events took place at the given time in the rest of drive (see **Figure 1C**). Here, we focused on fear and relaxation, which were, respectively, regarded as the experimental condition (Fear) and baseline condition (BL).

The fear induction took place on a route of 6 km (3.7 mi) length consisting of a highway section with three lanes and a country road section with one lane per direction. At 1:47 and 3:54 min after the beginning of the scenario, the automated driving vehicle was involved in an accident, which was caused by a vehicle swerving abruptly from the opposite lane (see **Figure 1B**). Both events were associated with loud noise (collision and loud braking). In addition, in order to distract the participants and to enhance the experienced fear, a text message in the form of an SMS was presented 5 s before the accident on the right in the field of vision during the drive. The drive had a total duration of ∼285 s. In this paper, the first event of the experimental condition was considered for further analysis, because it was expected to induce more intensive emotion than the subsequent events; furthermore, the onset time of the first event was comparable between control and experimental condition.

The scenario of relaxation took place on a 4 km (2.5 mi) country road with one lane per direction. Relaxation was supposed to be induced using the large-scale presentation of nature photographs as events. Four large-scale images, each with a presentation time of 50 s, were shown (see **Figure 1A**). The presentation was accompanied by relaxing music. The journey had a total duration of ∼265 s. Again, the first event, which was presented after 1 min of driving, was considered for further analysis.

Self-Report Questionnaires

After each driving scenario, the participants were asked to complete self-report questionnaires to assess their emotional experience during the drives. For this, we used the Positive and

FIGURE 1 | (A) Example for a relaxing image from the baseline scenario (Chiu, 2006); **(B)** View of participants while a vehicle swerving abruptly from the opposite lane during fear scenario; **(C)** Sketch of the procedure of each trial.

Negative Affect Schedule (PANAS) and an adapted version of Self-Assessment Manikin (SAM, Bradley and Lang, 1994).

The PANAS [Original: Watson et al., 1988; German version: Krohne et al., 1996] is composed of 20 adjectives describing ten positive and ten negative emotions, on a Likert scale (1 - very slightly, 2 - a little, 3 - moderately, 4 - quite a bit, 5 - extremely). We focused our analysis on the "scared" and "relaxed," which were semantically closest to our target emotions.

The SAM uses pictures to represent emotional responses on the three dimensions valence (pleasure-displeasure), arousal (calm-activity) and dominance/power (control-out of control). A question of experienced novelty was additionally added as the fourth dimension. Each dimension was represented by a Likert scale from one to nine (1 - very slightly to 9 - extremely).

Action Units

For extracting the frame-to-frame activity of the facial AUs, we used the Attention Tool FACET Module (FACET, iMotions), which is a face and AU detection software based on the FACS. This software can track and quantify changes in AUs frame by frame and was validated in studies comparing with

TABLE 2 | Description of action units, which were used in analysis.

AU	Description
1	Inner brow raiser
2	Outer brow raiser
4	Brow lowerer
5	Upper lid raiser
7	Lid tightener
15	Lip corner depressor
20	Lip stretcher
25	Lips part
26	Jaw drop

human coders (Krumhuber et al., 2019) and comparing with facial Electromyography (EMG) recording (Kulke et al., 2020). ~300,000 frames (37 participants * (285 + 265 s) * 15 Hz) were encoded with FACET, whereby AU 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 18, 20, 23, 24, 25, 26, and 28 were the variables. Each AU was assigned a numerical value, which is originally called evidence in the FACET software. For better understanding, we use the

FIGURE 2 | Mean and standard error of ratings in PANAS "relax" and "scared" (left) as well as in SAM scales and the added dimension of novelty (right) in Baseline (BL, light blue) and Fear (dark blue).
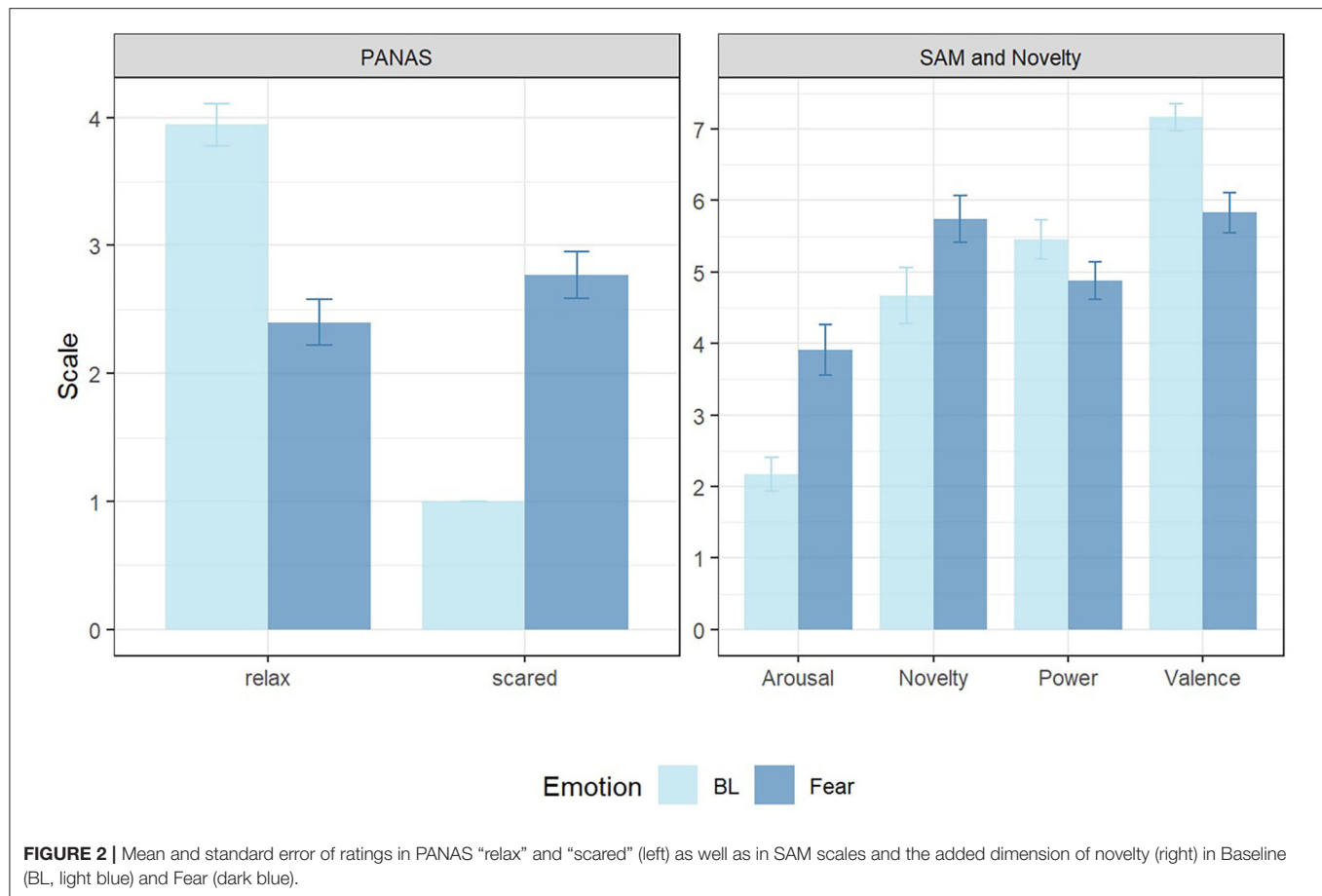
term AU index in the remainder of this work. The AU index is a raw data value that relates to the likelihood of an AU occurring. In order to reduce the difference between the participants, all output of encoding was scaled within every drive and adjusted, whereby the average AU index of the AUs in the first minute of the respective drive was subtracted from the AU index of the AUs in the experimental or control condition.

In order to quantify the changing in different components, we used the linear average value of certain relevant AUs as a compound measure to indicate components. We assumed that using of compound of AUs could increase the signal-to-noise ratio of the appraisal components. In a previous EMG study, facial muscle activity in the frontalis region was revealed to be related with the appraisal component of novelty (Sequeira et al., 2009). According to this assumption, the prediction of the CPM (Scherer et al., 2018, **Table 1**) and avoiding overlap between components, the compound (linear average) of upper facial AUs 1, 2, 4, 5, and 7 was used to indicate the appraisal component of high novelty and the compound of the lower facial AUs 15, 20, 25, and 26 for the appraisal component of low power (see **Table 2** for a semantic description of the AUs). AU 38 was not used because it was not covered by the software package used for facial AU analysis.

In order to reveal the temporal dynamics on both components, the AU compounds were segmented from

event onset to 5 s after event onset (BL: picture presentation, Fear: Swerving vehicle occurrence). On one hand, the mean value of the AU compounds in 5 s were calculated and compared between Fear and BL. On the other hand, the means were also aggregated in subsequent windows of 100 millisecond length in order to reveal the changing over time.

Statistical Analyses

According to the results of Shapiro-Wilk normality tests, neither subjective rating's data nor the mean value of AU compounds were normally distributed (dimension of novelty: $W = 0.94$, $p < 0.01$; dimension of valence: $W = 0.90$, $p < 0.001$; dimension of arousal: $W = 0.87$, $p < 0.001$; dimension of power: $W = 0.92$, $p < 0.001$; PANAS- "scared": $W = 0.74$, $p < 0.001$; PANAS- "relax": $W = 0.9$, $p < 0.001$; AU Compound of high novelty: $W = 0.96$, $p < 0.05$; AU Compound of low power: $W = 0.90$, $p < 0.001$). Therefore, a Wilcoxon test for dependent samples was implemented for the comparison between Fear and BL and the results were presented as Z-score. The condition with two levels Fear and BL was the only factor. The significance level of $\alpha = 0.05$ was used for the overall test. For determining the effect size, the computational parameter $r_{contrast}$ recommended by Rosenthal et al. (1994) was used. Hereby, the effect size is low if $r_{contrast} < 0.1$, medium if $r_{contrast} < 0.3$ and large if $r_{contrast} > 0.5$.

TABLE 3 | The rating on PANAS (descending ordered by the magnitude of difference between Fear and BL).

Items	M (Fear)	M (BL)	M (Fear - BL)	Z
Scared	2.77	1.00	1.77	4.83***
Surprised	2.94	1.36	1.58	4.76***
Alert	3.77	2.64	1.13	4.58***
Active	2.97	1.92	1.05	3.94***
Insecure	2.26	1.22	1.04	3.88***
Attentive	3.83	2.97	0.86	3.89***
Upset	1.77	1.00	0.77	3.63***
Afraid	1.77	1.03	0.74	3.45***
Interested	3.09	2.42	0.67	3.28**
Nervous	1.91	1.25	0.66	3.55***
Frustrated	1.69	1.06	0.63	3.85***
Jittery	1.63	1.08	0.55	2.76**
Distressed	1.60	1.14	0.46	2.14*
Determined	2.74	2.31	0.43	2.64**
Angry	1.43	1.00	0.43	2.11*
Ashamed	1.29	1.00	0.29	2.13*
Sad	1.17	1.14	0.03	0.22
Proud	1.31	1.47	−0.16	−1.56
Enthusiastic	1.91	2.14	−0.23	−1.17
Inspired	1.71	1.97	−0.26	−1.01
Excited	1.60	2.08	−0.48	−2.06*
Relax	2.40	3.94	−1.54	−4.27***

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

In order to identify time points at which the AU compound between Fear and BL begin to diverge, the data in the time interval of 5 s were analyzed pointwise by F-tests, which is believed to provide relevant rather than trivial differences between two functional linear models (Shen and Faraway, 2004). Using the "ERP" package (Causeur et al., 2014) with Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) in the R programming language, it was ensured that the false discovery rate (FDR) was controlled at a preset level α.

RESULTS

Manipulation Check

The participants' rating on the PANAS item "relax" was significantly higher in the BL scenarios than in the Fear scenarios according to a Wilcoxon test for dependent samples ($Z = -4.27$, $p < 0.001$, $r_{contrast} = 0.7$). On the contrary, the rating on the PANAS item "scared" was significantly higher in the fear scenarios comparing the BL scenarios ($Z = -4.83$, $p < 0.001$, $r_{contrast} = 0.79$) (see **Figure 2**). Furthermore, the ratings on PANAS item "surprised," "alert," "active," "insecure," "attentive," "upset," "afraid," "interested," "nervous," "frustrated," "jittery," "distressed," "determined" "angry" and "ashamed" were also significantly higher in Fear (see **Table 3**). Significant differences between Fear and BL scenarios were also found in the SAM dimensions arousal ($Z = 3.51$, $p < 0.001$, $r_{contrast} = 0.58$) and valence ($Z = -3.43$, $p < 0.001$, $r_{contrast} = 0.56$). No significant

difference was found for the rating on SAM's power dimension. However, a trend for a difference was revealed ($Z = -1.94$, $p = 0.053$, $r_{contrast} = 0.32$). Additionally, the participants' rating on the dimension novelty was significantly higher in the Fear scenarios according to a Wilcoxon test for dependent samples ($Z = 2.45$, $p < 0.05$, $r_{contrast} = 0.4$) (see **Figure 2**).

Action Units Compounds

Figure 3 shows the changing of relevant AUs in subsequent windows of 100 millisecond length for 5 s after event onset. Generally, there was difference between the changing in Bl and Fear on several AUs: AU 1, 2, 4, 5 as well as 15 were activated more in Fear than BL before 2.5 s, while after 2.5 s AU 20, 25, and 26 were activated more in Fear than BL.
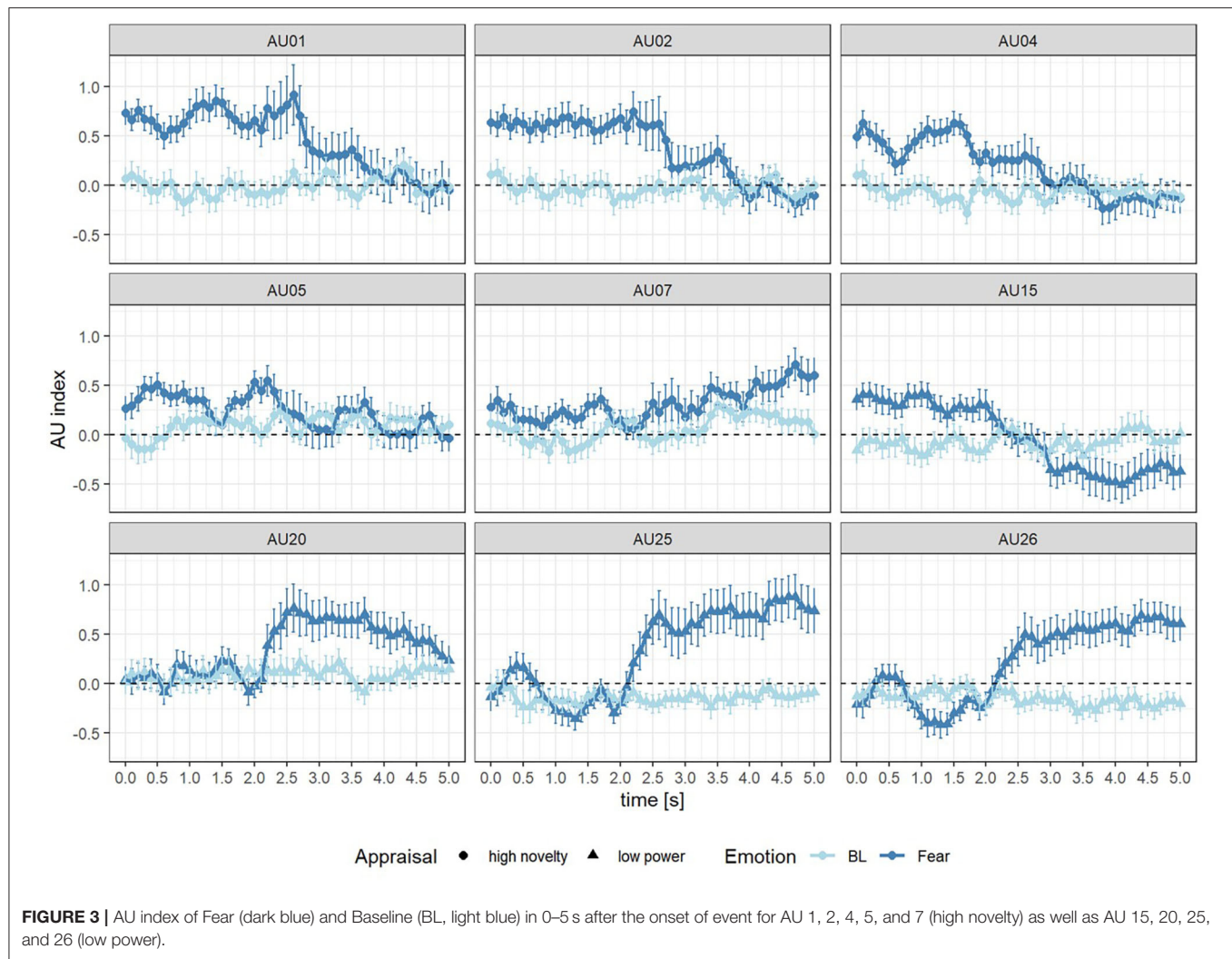
The linear average of the changing of the compounds for high novelty (AU 1, 2, 4, 5, and 7) and low power (AU 15, 20, 25, and 26) was calculated for 5 s after event onset. The Wilcoxon test for dependent samples indicated that the changing of novelty was significantly higher in Fear scenario ($Z = 2.80$, $p < 0.01$, $r_{contrast} = 0.46$). Significant differences between Fear and BL scenarios were also found in compound low power ($Z = 2.43$, $p < 0.05$, $r_{contrast} = 0.50$).

For the dynamics of the compounds high novelty and low power, the following results were obtained: The AU compound of high novelty was continuously significantly activated from 0 to 2.6 s after the onset of Fear compared to BL events (see **Figure 4A** and **Table 4**). The activation of the AU compound of low power started at 2.5 s, at which the difference between Fear and BL was significant ($M = 0.44$, $F = 2.99$, $p < 0.05$). Activation of the AU compound of low power in Fear could be discontinuously found between 2.5 and 4.6 s after event onset (see **Figure 4B** and **Table 4**).

DISCUSSION

The goal of this study was to investigate whether multidimensional analysis of facial expression can be a suitable as basis for the in-vehicle measurement of the drivers' emotion. Especially, we were interested whether we can capture the dynamics of facial expressions by considering effects of appraisal components. We found that the facial expression indicators of high novelty and low power were significantly activated after fear events. Furthermore, after fear events, the activation of high novelty occurred earlier than the activation of low power.

According to the self-report the experimental manipulation was successful. The PANAS item "scared" had a higher value in Fear scenarios, while the participants' rating on the PANAS item "relax" was higher in BL. The results provided evidence that the induction of fear and relaxation was successful. The evidence for a successful manipulation of the experiment was also found in the SAM and novelty scales. Fear is supposed to be located lower on the power and the valence dimension and higher on the arousal and the novelty dimension (Fontaine et al., 2007; Gillioz et al., 2016). The subjective ratings on the dimension of valence were lower and the ratings on the dimension of arousal and novelty were higher in Fear than BL. Besides, the subjective ratings on the dimension of power were descriptively

FIGURE 3 | AU index of Fear (dark blue) and Baseline (BL, light blue) in 0–5 s after the onset of event for AU 1, 2, 4, 5, and 7 (high novelty) as well as AU 15, 20, 25, and 26 (low power).

lower (not significantly, though). This may be due to the fact that the SAM's representativeness of emotional dimensions is still a question at issue (Schmidtke et al., 2014), so that the understanding of the SAM dimension of power could differ between participants. However, in total the manipulation check suggests that we successfully induced the emotional state of fear and relaxation in our driving simulator study.

According to the analysis of the time difference between AU compounds, we confirmed that facial expressions could be multidimensionally and dynamically analyzed. The activation of AU 1, 2, 4, and 5 was earlier than AU 20, 25, and 26. However, the activation of AU 7 and 15 was not as excepted. This may be due to the fact that the mapping between AUs and appraisal components is not always unique and AU7 and AU15 were also considered as the indicator of the appraisal component of unpleasantness (Scherer et al., 2018). Generally, it was revealed that the activation of the AUs in the upper face, which served as the indicator of high novelty, occurs earlier than the activation of the AUs in lower face, which served as the indicator of low power. On the one hand, the results on the dynamics of facial

expression provided evidence for the existence of novelty and power appraisals as proposed by the CPM. On the other hand, the temporal difference between novelty and power appraisal was verified. It was consistent with the prediction of the CPM that the appraisal of novelty occurs earlier than the appraisal of power. The results of this study suggest that emotions could be multidimensionally dynamically assessed through different dimensions at different times.

Besides the multidimensional and dynamic interpretation of emotions, the CPM interprets the individual differences in emotional reactions. According to the CPM, emotions are triggered by individual appraisals, which depend on the individual's goals, values and coping potential (Scherer, 2009b). In other words, the same event could produce an emotion with different time course and intensity or even a different emotion. With regard to the difference in time, we used a 5 s time window to ensure that every onset of event-related facial expressions could be collected. The comparably small standard error for the components of the different experimental conditions suggests that the variance of the underlying individual appraisals was low
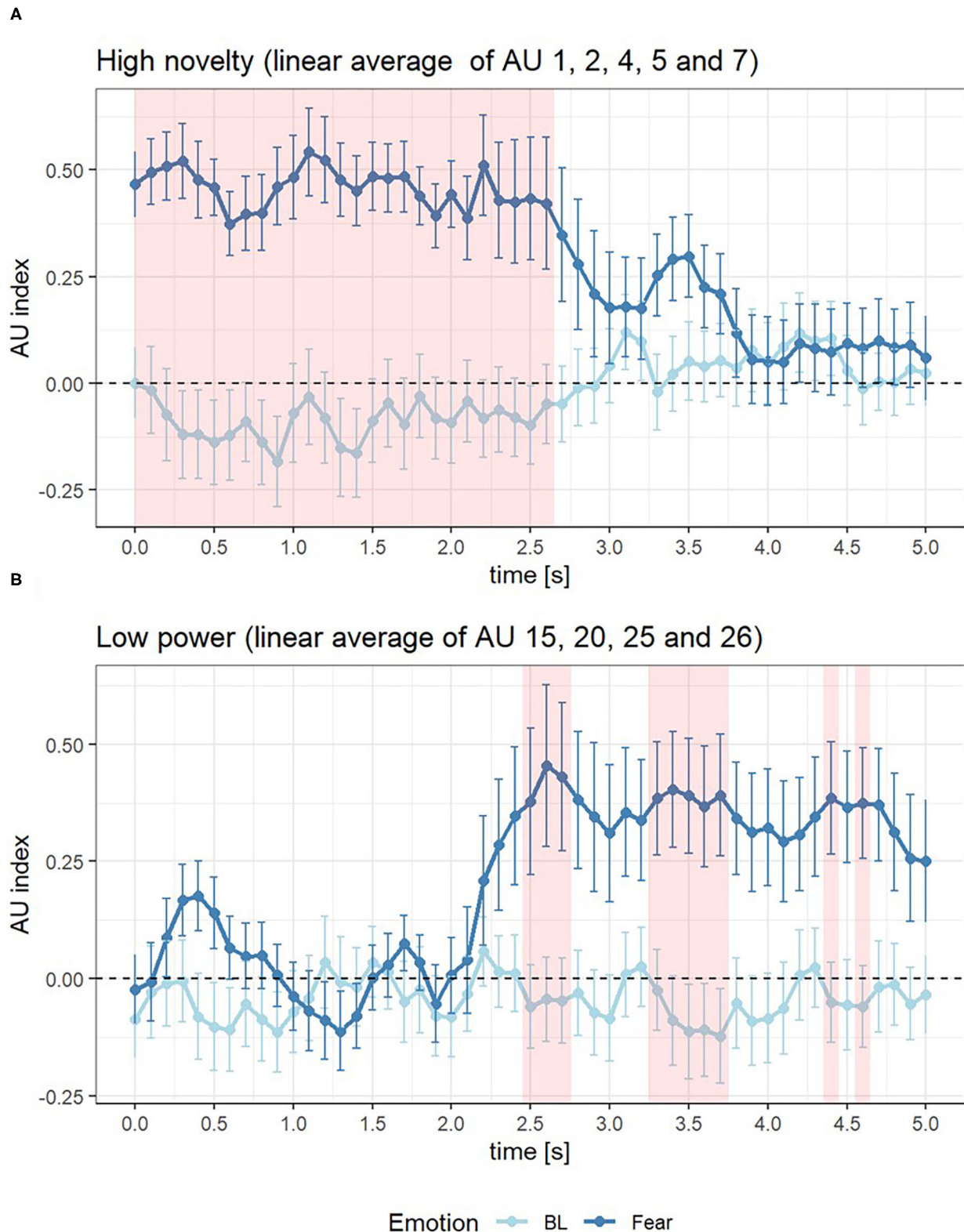
FIGURE 4 | AU index of Fear (dark blue) and Baseline (BL, light blue) in 0–5 s after the onset of event for component of high novelty **(A)** and low power **(B)**, where the red area represents the time interval when the AU index difference between Fear and BL >0.

TABLE 4 | Mean AU index difference of Fear and BL (*M*) and results of F-tests (*F*) for each time point of high novelty (AU 1, 2, 4, 5, and 7) and low power (AU 15, 20, 25, and 26).

High novelty (Fear - BL)						Low power (Fear - BL)					
Time	*M*	*F*	Time	*M*	*F*	Time	*M*	*F*	Time	*M*	*F*
0.0	0.47	3.93**	2.6	0.47	2.52*	0.0	0.06	0.56	2.6	0.50	3.30*
0.1	0.51	3.97**	2.7	0.40	2.10	0.1	0.02	0.19	2.7	0.48	3.09*
0.2	0.58	4.47**	2.8	0.29	1.54	0.2	0.10	0.81	2.8	0.41	2.48
0.3	0.64	5.00***	2.9	0.22	1.20	0.3	0.17	1.54	2.9	0.42	2.35
0.4	0.60	4.30**	3.0	0.14	0.81	0.4	0.26	2.44	3.0	0.39	2.33
0.5	0.60	5.07***	3.1	0.06	0.36	0.5	0.24	2.40	3.1	0.35	2.19
0.6	0.50	3.81**	3.2	0.08	0.46	0.6	0.18	1.81	3.2	0.31	2.08
0.7	0.49	3.98**	3.3	0.27	1.91	0.7	0.10	1.06	3.3	0.41	2.87*
0.8	0.54	4.24**	3.4	0.27	1.90	0.8	0.14	1.42	3.4	0.49	3.11*
0.9	0.64	4.48**	3.5	0.25	1.68	0.9	0.12	1.19	3.5	0.50	3.05*
1.0	0.55	3.84**	3.6	0.19	1.32	1.0	0.03	0.30	3.6	0.48	2.82*
1.1	0.57	4.08**	3.7	0.16	1.09	1.1	−0.03	−0.23	3.7	0.51	3.00*
1.2	0.60	4.19**	3.8	0.08	0.55	1.2	−0.12	−0.98	3.8	0.39	2.60
1.3	0.63	4.17**	3.9	−0.02	−0.15	1.3	−0.11	−0.82	3.9	0.40	2.57
1.4	0.61	4.73**	4.0	0.00	0.03	1.4	−0.06	−0.58	4.0	0.41	2.59
1.5	0.57	4.50**	4.1	−0.04	−0.24	1.5	−0.03	−0.34	4.1	0.36	2.19
1.6	0.53	4.00**	4.2	−0.02	−0.18	1.6	0.01	0.14	4.2	0.30	1.94
1.7	0.58	4.41**	4.3	−0.02	−0.12	1.7	0.13	1.24	4.3	0.32	1.99
1.8	0.47	4.03**	4.4	−0.03	−0.25	1.8	0.06	0.56	4.4	0.44	2.89*
1.9	0.47	3.62*	4.5	0.06	0.54	1.9	0.03	0.25	4.5	0.42	2.60
2.0	0.53	4.07**	4.6	0.09	0.69	2.0	0.09	0.85	4.6	0.43	2.85*
2.1	0.43	2.96*	4.7	0.10	0.74	2.1	0.07	0.58	4.7	0.39	2.56
2.2	0.59	3.54**	4.8	0.08	0.64	2.2	0.15	1.02	4.8	0.33	2.19
2.3	0.49	2.72*	4.9	0.05	0.39	2.3	0.27	1.83	4.9	0.31	2.07
2.4	0.51	2.80*	5.0	0.03	0.25	2.4	0.34	2.19	5.0	0.29	1.86
2.5	0.53	3.12**				2.5	0.44	2.99*			

*$p < .05$; **$p < .01$; ***$p < .001$.

in our study, which may be explained by the fact that the cover story and instructions ensured that participants had similar goals during the drives.

With respect to the ecological validity of this study, there are a few issues worth mentioning. We created an event producing a relatively strong emotional reaction in order use this strong reaction to evaluate whether it is in general possible to use the CPM to model facial reactions of drivers/users in a realistic setting (such as a driving simulation). We see this as a first step to employ the CPM as basis for in-vehicle emotion recognition and acknowledge that further research with less intense emotional episodes as well as during real-world driving is needed. In addition, a driving simulator setup with less ecological validity compared to real-world driving was chosen to have more control about the environmental conditions (e.g., weather). However, in general results from driving simulators have been shown to be transferable to real driving (see Shechtman et al., 2009; Helland et al., 2013). To add, the setting with the automated driving is a realistic setting given the current developments in the automotive domain, so that humans in vehicles will soon be able to engage in other tasks than controlling the car (like in the scenarios chosen).

In addition, it has to be noted that interpreting facial expressions alone is mostly not sufficient to know why the driver experiences a certain emotion and therefore also not sufficient to select the appropriate intervention strategy to support the driver. In a complex setting such as driving, we cannot say based on the facial expression alone whether the driver is fearful due to information she or he has received from a telephone conversation partner or due to the "risky" driving style of the automation. Therefore, it is also necessary to create a representation of the context to derive the need of the driver in a very situation as basis for the provision of the best possible intervention strategy (e.g., Drewitz et al., 2020). For instance, if appraisals pointing to fear have been detected, a virtual in-vehicle assistant could check whether parameters in the environment assessment, such as time-to-collision to the vehicle in front, indicate the occurrence of critical traffic events and, based on previous situations, could determine how likely the fear results from these. In case these probabilities are high, a specific intervention like a more defensive driving style could be chosen. If no relation to the vehicle exterior is likely and no other information about potential causes for the fearful state of the driver are present,

the AI assistance could offer more general help or even ask for the cause. To sum up, a specific support of the user needs more information than solely the interpretation of the facial expressions, however, a detection of the facial expression is an important step to be able to interpret the emotions of the driver in the first place.

The main limitation of this study is the way fear was induced in the context of driving. A dynamic assessment of emotions requires an event-related analysis with a distinct onset. Thus, we needed to set up an emotional event to induce fear and define a time point as the onset of this event. In this study, the traffic accident was regarded as the event of fear and the SMS 5 s before the accident was regarded as a distraction, which was assumed to intensify the fear against the accident. However, according to the results, in which the AU compound of high novelty was already activated at the 0 s after the event onset (see **Table 4**), the appraisal component of high novelty might have started earlier than the accident itself due to the SMS. Hence, future work using this approach should consider an event with a more distinct onset. Additionally, although the CPM model would predict similar facial expressions when similar appraisals are experienced, this generalizability across different events, e.g., the facial expression after a traffic accident as used here compared to expressions after other fearful (with less intensity) or other emotional events (such as something surprising) with similar underlying novelty and power appraisals needs to be evaluated in future work.

Another limitation is reliability of the software used for coding of AU activations. In order to simulate the in-vehicle facial expression recognition at the application level, we used the software package FACET to automatically quantify changes of AUs. Although the software was confirmed to have a high positive correlation with EMG recordings (Kulke et al., 2020), it is assumed that the recognition performance of spontaneous facial expressions in video is much lower than in photos (Stöckli et al., 2018). Hence, in order to control the reliability of software coding, human coding could verify the performance of automated facial expression coding in future research.

CONCLUSIONS

This research provides a new perspective on affective computing. For automated assessment, emotions were previously mostly regarded as a state with a single constant facial expression. However, facial expressions, especially in wild contexts such as driving, are dynamic processes resulting from underlying different appraisals. Models for emotion measurements from facial expressions need consider this multidimensional and dynamic nature. For the affective computing not only, the intensity and the duration of facial expressions is relevant, but also the temporal course of the activations of the different AUs in the facial expression, especially because only a minority of AUs can be unambiguously associated to specific emotions (Mehu and Scherer, 2015). Instead of chasing a certain pattern of facial expressions for a specific emotion, a dynamic perspective provides a multidimensional and multi-time domain solution, which can improve a robust and reliable measurement of drivers' emotion.

DATA AVAILABILITY STATEMENT

The dataset analyzed during the current study is available from the corresponding author on reasonable request.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MZ, KI, and UD: conceptualization and methodology. MZ: data curation, investigation, visualization, and writing – original draft. KI: project administration. KI, UD, and MJ: supervision. MZ, KI, UD, and MJ: writing – review & editing. All authors contributed to the article and approved the submitted version.

REFERENCES

Abdic, I., Fridman, L., McDuff, D., Marchi, E., Reimer, B., and Schuller, B. (2016). "Driver frustration detection from audio and video in the wild," in *IJCAI'16: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ed G. Brewka (Klagenfurt: Springer), 237.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

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

Causeur, D., Perthame, E., and Sheu, C.-F. (2014). "ERP : an R package for event-related potentials data analysis," in *UseR! 2014* (Los Angeles, CA: UCLA).

Chiu, W. (2006). *ShiFengWaterFall_002.jpg (JPEG-Grafik, 3456 × 2304 Pixel) - Skaliert (41%).* Available online at: https://upload.wikimedia.org/wikipedia/commons/5/51/ShiFengWaterFall_002.jpg (accessed October 23, 2020).

Drewitz, U., Ihme, K., Bahnmüller, C., Fleischer, T., La, H., Pape, A.-A., et al. (2020). "Towards user-focused vehicle automation: the architectural approach of the AutoAkzept Project," in *HCI in Mobility, Transport, and Automotive Systems. Automated Driving and In-Vehicle Experience Design Lecture Notes in Computer Science*, ed. H. Krömker (Cham: Springer International Publishing), 15–30.

Ekman, P., and Friesen, W. V. (1978). *Manual for the Facial Action Coding System.* Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., and Hager, J. C. (2002). *The Facial Action Coding System, 2nd Edn.* Salt Lake City, UT: Research Nexus eBook.

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychol. Sci.* 18, 1050–1057. doi: 10.1111/j.1467-9280.2007.02024.x

Fuller, R. (1984). A conceptualization of driving behaviour as threat avoidance. *Ergonomics* 27, 1139–1155. doi: 10.1080/00140138408963596

Gentsch, K., Grandjean, D., and Scherer, K. R. (2015). Appraisals generate specific configurations of facial muscle movements in a gambling task: evidence

for the component process model of emotion. *PLoS ONE* 10:e0135837. doi: 10.1371/journal.pone.0135837

Gillioz, C., Fontaine, J. R. J., Soriano, C., and Scherer, K. R. (2016). Mapping emotion terms into affective space: further evidence for a four-dimensional structure. *Swiss J. Psychol.* 75, 141–148. doi: 10.1024/1421-0185/a000180

Helland, A., Jenssen, G. D., Lervåg, L.-E., Westin, A. A., Moen, T., Sakshaug, K., et al. (2013). Comparison of driving simulator performance with real driving after alcohol intake: a randomised, single blind, placebo-controlled, cross-over trial. *Accid. Anal. Prev.* 53, 9–16. doi: 10.1016/j.aap.2012.12.042

Ihme, K., Preuk, K., Drewitz, U., and Jipp, M. (2019). "Putting people center stage – to drive and to be driven," in *Fahrerassistenzsysteme 2018 Proceedings*, ed T. Bertram (Wiesbaden: Springer Fachmedien), 98–108.

Ihme, K., Unni, A., Zhang, M., Rieger, J. W., and Jipp, M. (2018). Recognizing frustration of drivers from face video recordings and brain activation measurements with functional near-infrared spectroscopy. *Front. Hum. Neurosci.* 12:327. doi: 10.3389/fnhum.2018.00327

Jeon, M. (2015). Towards affect-integrated driving behaviour research. *Theoret. Issues Ergon. Sci.* 16, 553–585. doi: 10.1080/1463922X.2015.1067934

Jeon, M., Yim, J.-B., and Walker, B. N. (2011). "An angry driver is not the same as a fearful driver: effects of specific negative emotions on risk perception, driving performance, and workload," in *The 3rd International Conference on Automotive User Interfaces and Vehicular Applications (AutomotiveUI'11)* (Salzburg), 137–142.

Klein, J., Moon, Y., and Picard, R. W. (2002). This computer responds to user frustration: theory, design, and results. *Interact. Comp.* 14, 119–140. doi: 10.1016/S0953-5438(01)00053-4

Krohne, H. W., Egloff, B., Kohlmann, C.-W., and Tausch, A. (1996). Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS). *Diagnostica* 42, 139–156. doi: 10.1037/t49650-000

Krumhuber, E. G., Küster, D., Namba, S., Shah, D., and Calvo, M. G. (2019). Emotion recognition from posed and spontaneous dynamic expressions: human observers versus machine analysis. *Emotion.* doi: 10.1037/emo0000712. [Epub ahead of print].

Kulke, L., Feyerabend, D., and Schacht, A. (2020). A comparison of the affectiva imotions facial expression analysis software with EMG for identifying facial expressions of emotion. *Front. Psychol.* 11:329. doi: 10.3389/fpsyg.2020.00329

Löcken, A., Ihme, K., and Unni, A. (2017). "Towards designing affect-aware systems for mitigating the effects of in-vehicle frustration," in *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct - AutomotiveUI '17* (Oldenburg: ACM Press), 88–93.

Malta, L., Miyajima, C., Kitaoka, N., and Takeda, K. (2011). Analysis of real-world driver's frustration. *IEEE Transact. Intell. Transport. Syst.* 12, 109–118. doi: 10.1109/TITS.2010.2070839

Mehu, M., and Scherer, K. R. (2015). Emotion categories and dimensions in the facial communication of affect: an integrated approach. *Emotion* 15, 798–811. doi: 10.1037/a0039416

Namba, S., Kabir, R. S., Miyatani, M., and Nakao, T. (2018). Dynamic displays enhance the ability to discriminate genuine and posed facial expressions of emotion. *Front. Psychol.* 9:672. doi: 10.3389/fpsyg.2018.00672

Rosenthal, R., Cooper, H., and Hedges, L. (1994). "Parametric measures of effect size," in *The Handbook of Research Synthesis*, eds H. M. Cooper and L. V. Hedges (New York, NY: Russell Sage Foundation), 231–244.

Scherer, K. R. (1984). "On the nature and function of emotion: a component process approach," in *Approaches to Emotion*, eds K. R. Scherer and P. Ekman (Hillsdale, NJ: Erlbaum).

Scherer, K. R. (2009a). Emotions are emergent processes: they require a dynamic computational architecture. *Philos. Transact. R. Soc. B Biol. Sci.* 364, 3459–3474. doi: 10.1098/rstb.2009.0141

Scherer, K. R. (2009b). The dynamic architecture of emotion: evidence for the component process model. *Cogn. Emot.* 23, 1307–1351. doi: 10.1080/02699930902928969

Scherer, K. R. (2019). Towards a prediction and data driven computational process model of emotion. *IEEE Transact. Affect. Comput.* doi: 10.1109/TAFFC.2019.2905209. [Epub ahead of print].

Scherer, K. R., Ellgring, H., Dieckmann, A., Unfried, M., and Mortillaro, M. (2019). Dynamic facial expression of emotion and observer inference. *Front. Psychol.* 10:508. doi: 10.3389/fpsyg.2019.00508

Scherer, K. R., Mortillaro, M., Rotondi, I., Sergi, I., and Trznadel, S. (2018). Appraisal-driven facial actions as building blocks for emotion inference. *J. Person. Soc. Psychol.* 114, 358–379. doi: 10.1037/pspa0000107

Schmidt-Daffy, M., Brandenburg, S., and Beliavski, A. (2013). Velocity, safety, or both? How do balance and strength of goal conflicts affect drivers' behaviour, feelings and physiological responses? *Accid. Anal. Prev.* 55, 90–100. doi: 10.1016/j.aap.2013.02.030

Schmidtke, D. S., Schröder, T., Jacobs, A. M., and Conrad, M. (2014). ANGST: affective norms for German sentiment terms, derived from the affective norms for English words. *Behav. Res. Methods* 46, 1108–1118. doi: 10.3758/s13428-013-0426-y

Sequeira, H., Hot, P., Silvert, L., and Delplanque, S. (2009). Electrical autonomic correlates of emotion. *Int. J. Psychophysiol.* 71, 50–56. doi: 10.1016/j.ijpsycho.2008.07.009

Shechtman, O., Classen, S., Awadzi, K., and Mann, W. (2009). Comparison of driving errors between on-the-road and simulated driving assessment: a validation study. *Traffic Injury Prev.* 10, 379–385. doi: 10.1080/15389580902894989

Shen, Q., and Faraway, J. (2004). An F test for linear models with functional responses. *Stat. Sin.* 14, 1239–1257. Available online at: http://www.jstor.org/stable/24307230

Stöckli, S., Schulte-Mecklenbeck, M., Borer, S., and Samson, A. C. (2018). Facial expression analysis with AFFDEX and FACET: A validation study. *Behav, Res.* 50, 1446–1460. doi: 10.3758/s13428-017-0996-1

Techer, F., Ojeda, L., Barat, D., Marteau, J.-Y., Rampillon, F., Feron, S., et al. (2019). Anger and highly automated driving in urban areas: the role of time pressure. *Transport. Res. Part F* 64, 353–360. doi: 10.1016/j.trf.2019.05.016

Tews, T.-K., Oehl, M., Siebert, F. W., Höger, R., and Faasch, H. (2011). "Emotional human-machine interaction: cues from facial expressions," in *Human Interface and the Management of Information. Interacting with Information Lecture Notes in Computer Science,* eds M. J. Smith, and G. Salvendy (Orlando, FL: Springer), 641–650.

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Person. Soc. Psychol.* 54:1063. doi: 10.1037/0022-3514.54.6.1063

Zhang, M., Ihme, K., and Drewitz, U. (2019). Discriminating drivers' emotions through the dimension of power: Evidence from facial infrared thermography and peripheral physiological measurements. *Transport. Res. Part F.* 63, 135–143. doi: 10.1016/j.trf.2019.04.003

Scheduling and Rescheduling Operations Using Decision Support Systems: Insights From Emotional Influences on Decision-Making

Mélody Mailliez[1]*, Olga Battaïa[2] and Raphaëlle N. Roy[1]

[1] *Institut Supérieur de l'Aéronautique et de l'Espace, ISAE-SUPAERO, Université de Toulouse, Toulouse, France, [2] Kedge Business School, Talence, France*

For many years, manufacturers have focused on improving their productivity. Production scheduling operations are critical for this objective. However, in modern manufacturing systems, the original schedule must be regularly updated as it takes places in a dynamic and uncertain environment. The modern manufacturing environment is therefore very stressful for the managers in charge of the production process because they have to cope with many disruptions and uncertainties. To help them in their decision-making process, several decision support systems (DSSs) have been developed. A recent and enormous challenge is the implementation of DSSs to efficiently manage the aforementioned issues. Nowadays, these DSSs are assumed to reduce the users' stress and workload because they automatically (re)schedule the production by applying algorithms. However, to the best of our knowledge, the reciprocal influence of users' mental state (i.e., cognitive and affective states) and the use of these DSSs have received limited attention in the literature. Particularly, the influence of users' unrelated emotions has received even less attention. However, these influences are of particular interest because they can account for explaining the efficiency of DSSs, especially in modulating DSS feedback processing. As a result, we assumed that investigating the reciprocal influences of DSSs and users' mental states could provide useful avenues of investigation. The intention of this article is then to provide recommendations for future research on scheduling and rescheduling operations by suggesting the investigation of users' mental state and encouraging to conduct such research within the neuroergonomic approach.

Keywords: scheduling, rescheduling, decision support systems, incidental emotions, uncertainty, decision-making, modern manufacturing environment

INTRODUCTION

Establishing production schedules demands a detailed description and knowledge of the production process and requires handling a large amount of information (Rossit et al., 2019). Indeed, scheduling decisions are "a complex cognitive process that comprises a considerable number of interrelated subtasks" (Dimopoulos et al., 2012; p. 8–9; see Cegarra, 2008, for a cognitive typology of scheduling situations). In the modern manufacturing environment, the original schedule must be regularly updated because it takes place in a very dynamic and uncertain environment. In such an environment, it is not possible to create procedures for every disruption that might occur. Unexpected events inevitably happen and affect the original schedule requiring a

time-pressured response (Battaïa et al., 2019). For example, taking too long to deal with unexpected events can lead to a significant decrease in the manufacturing system performance (Jepson et al., 2017). Production scheduling and rescheduling operations are crucial to maintain and increase the manufacturing productivity and effectiveness (Vossing, 2017), especially in identifying conflicts in production lines and anticipating the occurrence of unexpected events (Vieira et al., 2003; Rossit et al., 2019; see also Larco Martinelli et al., 2019 for a sequence of schedulers' action).

Decision support systems (DSSs) have been developed to support managers and more generally any kind of operators in charge of scheduling activities in their decision-making process (Manzey et al., 2012; Zikos et al., 2018; Onnasch and Hösterey, 2019). DSSs could help users in their analysis of the situation to reach an optimal and effective solution (Riveiro et al., 2014). Huge interest has been demonstrated in investigating the effect of DSSs on user performance (e.g., Ferris et al., 2010). DSSs could be useful for managers and operators in improving on-time delivery, increasing their responsivity (Herrmann, 2006) by providing information that may release users' cognitive demands (Lee and Seong, 2009; Onnasch and Hösterey, 2019). For instance, when using DSSs, individuals might understand and identify potential mistakes more easily (Lee and Seong, 2009). DSSs are also supposed to reduce the mental workload as they automatically reschedule the production planning (e.g., Onnasch and Hösterey, 2019). For example, Navarro et al. (2018) have shown that the subjective workload increased when participants performed all tasks exclusively by themselves compared to when they used fully automated tasks (see also Röttger et al., 2009).

However, DSSs do not necessarily imply an improvement in individuals' performance. It has been shown that they take more time to intervene when using DSSs because they have to recover situation awareness (Lee and Seong, 2009; van der Kleij et al., 2018). Although there is a strong interest in human factors within the scheduling and rescheduling literature (see, e.g., Sanderson, 1989; Crawford and Wiers, 2001; for reviews), there is also a need to deepen the understanding of operators' cognitive processes and performance (see also Smith and Geddes, 2003). This issue could be resolved by bridging the gap between laboratory and field studies. Hence, neuroergonomics, a recent field at the crossroads of several fields of study such as neuroscience, cognitive engineering, and human factors, proposes to examine the brain mechanisms that underlie human–technology interaction. Especially, this approach aims to investigate the cognitive and neural processes in the context of carrying out various real-world tasks under investigation, rather than under reduced isolated conditions that occur only in the laboratory (Callan and Dehais, 2019). In light of this statement, perspectives on overlooked factors are presented in the following section. We therefore highlight how the neuroergonomic approach can substantially improve the understanding of operators' mental states and performance during DSS use while reducing the gap between laboratory and field studies.

FIGURE 1 | Representation of the reciprocal influence of DSSs and users' mental state during (re)scheduling production operations. Line A: influence of users' mental state (i.e., cognitive and affective state) on DSS use. Line B: influence of DSSs on users' mental state (i.e., cognitive and affective state).
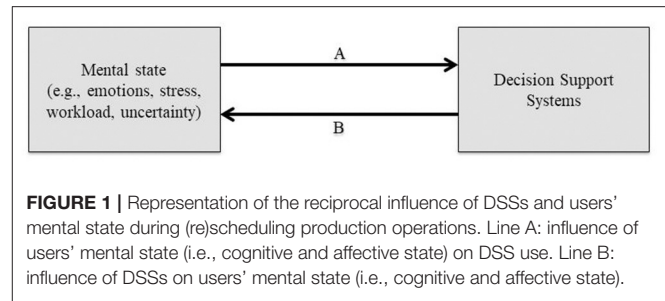
PERSPECTIVES AND DISCUSSION

DSS efficiency could depend on both situational and environmental factors (Lee and Seong, 2009). We therefore assume that investigating the reciprocal influence of DSSs and operators' mental state (**Figure 1**) could provide useful avenues of investigation. Strong interest has been devoted to investigate technical, social, and cognitive factors influencing the adoption and use of systems (Stein et al., 2015). However, less interest has been devoted to the role of emotional factors in user behavior (Thüring and Mahlke, 2007; Stein et al., 2015). We will discuss some of the key findings in usability research aiming to understand how individuals and their mental state influence their engagement with information systems. We then aim to highlight the relative importance of investigating the influence of operators' emotional state on (re)scheduling decision-making. Particularly, we believe that the influence of unrelated emotions (i.e., incidental emotions) on feedback processing can account for the non-systematic improvement in operators' performance when using DSSs. Investigating the relative influence of DSSs and operators' mental state also involves characterizing the influence of DSSs on users' mental state. As a result, we will highlight how neuroergonomics can improve this latter line of investigation (**Figure 1**).

Emotional Influences on DSS Use

Although research on usability is not new (e.g., Sagar and Saha, 2017 for a review), the interest to understand users' subjective experience while they interact with technological artifacts is more recent (e.g., Stein et al., 2015; Jung et al., 2017). It is then widely acknowledged that taking into account the feelings of users' experience is of crucial importance. Particularly, emotion could be one of the main dimensions of user experience (e.g., Thüring and Mahlke, 2007; Bargas-Avila and Hornbæk, 2011; Saariluomaand and Jokinen, 2014; Jeon, 2017). Studies looking at information technologies (e.g., DSSs) as an affect-inducing stimulus and influencing behavior (**Figure 1**, Line B) have to be distinguished from the ones looking at how affect influences information technology use (**Figure 1**, Line A; Stein et al., 2015). Subjective emotional experience shaped by the interaction with technologies depends on various factors related to the relationship between individuals' differences (e.g., coping, task events, and design) and emotions (Jokinen, 2015; Stein et al., 2015). A well-designed system is experienced as more positive, less arousing, more pleasant, goal conductive, and less novel

(Thüring and Mahlke, 2007). According to Jokinen (2015), users appraise a significant task event and respond to it with coping strategies, which in return influence their performance (Jokinen, 2015).

Individuals' performance could also be affected by emotions induced prior to the interaction with information technologies. Although a large body of work investigates how information technologies induce emotions and how they influence subsequent behavior, little has been done in determining how incidental emotions (e.g., induced prior to the interaction) shape users' performance. A positive mood (or emotion) could be positively associated with a perceived ease of use (Cenfetelli, 2004), habit formation in technology use (Lankton et al., 2010), and appealing stimulus detection speed (Reppa et al., 2020). Aesthetic appeal could even counterbalance the effects of negative mood (Reppa et al., 2020). These results are in line with the fact that the emotional state under which individuals perform decision-making tasks modulates their performance (e.g., Blanchette and Richards, 2010). Users' affective state is largely ignored in the current DSS literature even if it has been demonstrated that individuals in a positive mood use their DSSs significantly more efficiently (**Figure 1**, Line A) than individuals in a neutral mood (Djamasbi, 2007). However, the authors take a valence-based approach (as the majority of usability research). This approach holds that all types of positive mood (or positive emotions) have the same influence, which differs from the performance triggered by all types of negative mood (or negative emotions; see Blanchette and Richards, 2010 for a review). The valence-based approach has been challenged by the seminal work of Lerner and Keltner (2000, 2001), who demonstrated that two emotions unrelated to the decision-making process, i.e., incidental emotions (e.g., induced prior to the decision-making tasks), and sharing the same valence could lead to different decisions.

This emotion-specific framework [i.e., The Appraisal Tendency Framework (ATF); Lerner and Keltner, 2001; see also Han et al., 2007] assumes that each incidental emotion can be defined by its score for a set of appraisals (pleasantness, anticipated effort, control, responsibility, attentional activity, and certainty, according to Smith and Ellsworth, 1985). For example, anger is defined by a high degree of certainty (i.e., the extent to which individuals understand what is happening and are able to predict future events; Smith and Ellsworth, 1985) and by individual control. By contrast, fear is defined by a low degree of certainty and situational control. These features could activate "a predisposition to appraise future events in line with the central appraisal dimensions that triggered the emotion" (Han et al., 2007, p. 160). A series of experiments showed that the appraisal of control associated with incidental emotions could mediate the link between emotions and decisions. Consequently, we posit that investigating the effect of incidental specific emotions could represent a fruitful avenue to deepen the understanding of operators' (re)scheduling decisions and the use of DSSs (**Figure 1**, Line A). It can therefore be expected that incidental-specific emotions influence the use of DSSs. For example, it can be hypothesized that incidental happiness leads to better DSS use than incidental hope, especially due to the different feedback

processing triggered by those two incidental emotions (Mailliez et al., 2020).

Emotional Influences and Feedback Processing

An explanation of why specific emotions could lead to different feedback processing—and thus performance—may stem from the type of information processing triggered by the appraisal of certainty (i.e., heuristic or deliberative processing). It has been shown that incidental negative emotions are associated with opposite patterns of performance (Bagneux et al., 2012, 2013; Bollon and Bagneux, 2013). Better performance is observed with emotions associated with a high degree of certainty (Bagneux et al., 2012, 2013; Bollon and Bagneux, 2013; Iyilikci and Amado, 2018). Particularly, incidental emotions associated with a high degree of certainty may trigger a heuristic feedback processing, whereas incidental emotions associated with a low degree of certainty might trigger a deliberative one (Tiedens and Linton, 2001).

When feedback is deliberatively processed, the number of times that a decision leads to a positive or negative outcome is used in controlled cognitive processes such as rethinking the decision strategy (Schiebener and Brand, 2015). Deliberative processing is strongly dependent on the individuals' available cognitive resources (Evans and Stanovich, 2013). Scheduling decisions are cognitively demanding and complex (Berglund and Karltun, 2007; Larsen and Pranzo, 2019). However, individuals' cognitive resources may be limited (Bechara and Damasio, 2005); hence, it could be impossible to define which decision is better than another. In contrast to deliberative processing, the heuristic processing of feedback might allow the processing of emotional cues shaped by the association between feedback (positive vs. negative outcomes) and the elicited emotions (Bechara and Damasio, 2005). Heuristic processing might enable the individuals to process emotionally charged information via the automaticity of emotions (Kahneman and Frederick, 2007). Heuristic processing is therefore less cognitively demanding and allows processing a greater amount of information. Consequently, the heuristic feedback processing might be more effective than deliberative processing to achieve better performance in sequential decision-making, such as (re-) scheduling ones.

While studies on feedback processing have largely demonstrated their influence on sequential decision-making (e.g., Brand et al., 2007; Schiebener and Brand, 2015), the effect of feedback processing on (re)scheduling has received little interest. Moreover, it has been stated that DSSs can replace or automate certain cognitive processes, leading to an increase in individuals' information processing capacities (Djamasbi, 2007). While DSS feedback studies highlight the effect of feedback on the use of DSSs (e.g., Lim et al., 2005; Djamasbi and Loiacono, 2008), they do not include the potential influence of specific emotions on feedback processing. Including such an incidental influence is particularly important as it can mediate the operators' performance. Particularly, emotional influences on feedback processing (**Figure 1**, interaction between Lines

A and B) could be one of the factors explaining both better performance in (re)scheduling decisions and DSS efficiency, especially because emotional influences may trigger different information processing strategies (i.e., deliberative vs. heuristic).

To summarize, previous research on usability focused on how information technologies induce emotions and how these emotions influence the subsequent behavior (**Figure 1**, Line B). Influences of emotions that are not shaped by the interaction (i.e., incidental emotions) are less investigated (**Figure 1**, Line A), and interactions between these emotional influences even less so (**Figure 1**, interaction between Lines A and B). Scholars seem to focus on a specific period (e.g., during or before the interaction). It could be the consequence of a reluctance to theorize and operationalize the users' emotional experience (Jokinen, 2015). This reluctance might stem from the fact that user experience is considered as holistic (Boehner et al., 2007; Jokinen, 2015). This consideration and the valence-based approach taken by studies do not help to deepen the understanding of why and how users' emotional state and generally users' mental state influence the interaction with information technologies. We argued that taking an emotion-specific framework as forecasted by the ATF (and more recently, the emotion-imbued choice; Lerner et al., 2015) could provide an interesting path to consider both the contextual complexity (e.g., users' and technologies' characteristics) while taking into account the different emotional influences and their interaction. Beyond the theorization, operationalization could have everything to gain from being set in a neuroergonomic perspective.

Toward Physiology-Based Mental State Assessment for DSS Use Characterization

Users' subjective experience can be thought of as private and immediate. The use of questionnaires to elicit users' emotional state has been widely debated (Schorr, 2001). One may argue that individuals are reporting their general knowledge concerning emotions, not their current emotional states (Jokinen, 2015). It could therefore be very difficult to put user experience into words (Dennett, 1988). Emotions (especially their associated appraisals) would be responsible for changes in individuals' physiology (Scherer, 2009). The investigation of emotional influences on (re)scheduling decisions can therefore take place within the larger context of deepening our understanding of DSS users' mental state through the lens of neuroergonomics.

Results about the effect of DSSs on users' mental state, especially their workload, remain equivocal, as it was mainly demonstrated at a subjective level. As Charles and Nixon (2019) highlighted in their systematic literature review, there is no single measure that discriminates mental workload, but there is a variety of physiological and behavioral data. A deeper understanding of the relationship between operators' mental workload using DSSs, as well as the reciprocal influence between DSSs and operators' mental state, will therefore be improved by the neuroergonomic approach. This approach is of particular interest as it allows going further than the classical approach based on subjective and behavioral measures by using physiological measures such as cardiac and cerebral activity

markers. Indeed, behavioral metrics, although objective ones, might not reflect all mental processes that take place as illustrated by the inverted U-shaped performance curve observed under varying levels of arousal and task demands (VaezMousavi et al., 2009), as well as by the absence of difference reported between several difficulty levels for very low or very high task demands (Mehler et al., 2009). Hence, individuals may perform adequately but at a great cognitive cost, which might harm them in the long term, and impede their capability to deal with other task-external solicitations.

The neuroergonomic approach therefore allows assessing operators' mental state during operations (e.g., manufacturing ones) rather than afterward or by interrupting the task such as done with subjective measures acquired through questionnaires. Particularly, by using psychophysiological measures such as cardiac and cerebral activity ones, one could monitor operators' stress and workload level during (re)scheduling operations, as well as other cognitive and affective mental states as already studied in the ground and aerial transportation domains (see Borghini et al., 2014; Dehais and Callan, 2019 for reviews). Usual metrics for workload and stress assessment include heart rate and heart rate variability, as well as the power in various frequency bands (e.g., alpha power at parietal sites) recorded through electroencephalography (EEG; e.g., alpha [8 12] Hz) (Roy et al., 2013, 2020; Roy and Frey, 2016). Regarding affective state assessment, the same cardiac features (i.e., heart rate and heart rate variability), as well as the power in various EEG frequency bands and connectivity between electrodes (e.g., in the gamma band, >30 Hz), are metrics known to reflect arousal and valence (Wu et al., 2010; Chen et al., 2015). Usability studies that include physiological measurements have also started to be run (Hu et al., 2000; Brocke et al., 2013; Bhatt et al., 2019). It should be noted that physiological activity associated with mental processes can be recorded during operations (e.g., scheduling and rescheduling operations) and analyzed offline. However, the most striking advantage of this approach is that physiological measures can also be recorded and analyzed in an online manner.

Indeed, the use of such an online analysis approach has enabled researchers to deepen their mental states assessment in ecological settings, as well as to design better interfaces and support tools. Particularly, the use of machine learning tools has recently allowed researchers and engineers to develop adaptive systems that take physiological data as inputs. Such systems that enable cognitive and affective computing are often called biocybernetics systems or passive brain–computer interfaces (Fairclough, 2009; Zander and Kothe, 2011). By enabling the estimation of certain mental states (e.g., mental workload, fatigue, attentional level, emotional state) and modifying the interaction with the user, such bioadaptive or neuroadaptive systems provide a new means to increase safety and performance in operational environments (Lotte and Roy, 2019). Examples of countermeasures that could be implemented to deal with inadequate stress or mental workload levels are a modification of the interface, as performed in the Air Traffic Controller context (Aricò et al., 2016; Saint-Lot et al., 2020). Other solutions from the Human-Unmanned Aerial Systems interaction domain are to dynamically modify the automation level of the (re)scheduling

task (Ruff et al., 2002) or even to dynamically reallocate the load between teammates (Walters and Barnes, 2002). To put in a nutshell, the neuroergonomic approach seems a fruitful avenue of investigation to deepen the understanding of the reciprocal influence between DSSs and users' mental state (**Figure 1**, Lines A and B).

CONCLUSION

Although the complementary strength of DSSs and individuals has been demonstrated in previous work (see MacCarthy et al., 2001 for a review), to date there has been little work on characterizing the reciprocal influence of their mental state from an emotional perspective. Besides characterizing the influence of DSSs on users' mental state, factors such as the influence of incidental emotions and its interaction with feedback processing have received even less attention. However, they are of particular interest as they could represent factors that can improve individuals' and managers' performance. We argue that incidental emotions could mediate the effect of DSSs on users' mental state. This investigation cannot be carried out without considering the reciprocal influence of DSSs on mental (cognitive and emotional) state (and *vice versa*). As a perspective, the neuroergonomic approach is introduced. This approach is of particular interest for the human factors and the engineering communities that can benefit from new tools to better characterize (re)scheduling-induced mental states during

DSS use. The striking advantage of this emerging approach is that it allows a psychophysiological assessment in both an offline and online manner.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

MM contributed in the initial conception of the work, drafting the article, revisions of the article, and final approval of the version to be published. RR contributed in drafting the neuroergonomics sub-section, revisions of the article, and final approval of the version to be published. OB contributed in critical revisions of the article and final approval of the version to be published. All authors contributed to the article and approved the submitted version.

FUNDING

REFERENCES

Aricò, P., Borghini, G., Di Flumeri, G., Colosimo, A., Bonelli, S., Golfetti, A., et al. (2016). Adaptive automation triggered by EEG-based mental workload index: a passive brain-computer interface application in realistic air traffic control environment. *Front. Hum. Neurosci.* 10:539. doi: 10.3389/fnhum.2016.00539

Bagneux, V., Bollon, T., and Dantzer, C. (2012). Do (un)certainty appraisal tendencies reverse the influence of emotions on risk taking in sequential tasks? *Cogn. Emotion* 26, 568–576. doi: 10.1080/02699931.2011.602237

Bagneux, V., Font, H., and Bollon, T. (2013). Incidental emotions associated with uncertainty appraisals impair decisions. *Motiv. Emot.* 37, 818–827. doi: 10.1007/s11031-013-9346-5

Bargas-Avila, J. A., and Hornbæk, K. (2011). "Old wine in new bottles or novel challenges? A critical analysis of empirical studies of user experience," in *Conference on Human Factors in Computing Systems – Proceedings* (Gaithersburg, MD), 2689–2698. doi: 10.1145/1978942.1979336

Battaïa, O., Sanmartin, L., and Pralet, C. (2019). Dealing with disruptions in low-volume manufacturing: a constraint programming approach. *52nd CIRP Conf. Manufact. Syst.* 81, 1372–1375. doi: 10.1016/j.procir.2019.04.030

Bechara, A., and Damasio, A. R. (2005). The somatic marker hypothesis: a neural theory of economic decision. *Games Econ. Behav.* 52, 336–372. doi: 10.1016/j.geb.2004.06.010

Berglund, M., and Karltun, J. (2007). Human, technological and organizational aspects influencing the production scheduling process. *Int. J. Product. Econ.* 110, 160–174. doi: 10.1016/j.ijpe.2007.02.024

Bhatt, S., Agrali, A., McCarthy, K., Suri, R., and Ayaz, H. (2019). Web usability testing with concurrent fNIRS and eye tracking. *Neuroergonomics* 2019, 181–186. doi: 10.1016/B978-0-12-811926-6.00030-0

Blanchette, I., and Richards, A. (2010). The influence of affect on higher level cognition: a review of research on interpretation, judgement, decision making and reasoning. *Cogn. Emotion* 24, 37–41. doi: 10.1080/02699930903132496

Boehner, K., DePaula, R., Dourish, P., and Sengers, P. (2007). How emotion is made and measured. *Int. J. Human Comput. Stud.* 65, 275–291. doi: 10.1016/j.ijhcs.2006.11.016

Bollon, T., and Bagneux, V. (2013). Can the uncertainty appraisal associated with emotion cancel the effect of the hunch period in the Iowa Gambling Task? *Cogn. Emotion* 27, 376–384. doi: 10.1080/02699931.2012.712947

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., and Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* 44, 58–75. doi: 10.1016/j.neubiorev.2012.10.003

Brand, M., Grabenhorst, F., Starcke, K., Vandekerckhove, M. M. P., and Markowitsch, H. J. (2007). Role of the amygdala in decisions under ambiguity and decisions under risk: evidence from patients with Urbach-Wiethe disease. *Neuropsychologia* 45, 1305–1317. doi: 10.1016/j.neuropsychologia.2006.09.021

Brocke, J., Vom Riedl, R., and Léger, P.-M. (2013). Application strategies for neuroscience in information systems design science research. *J. Comput. Inform. Syst.* 53, 1–13. doi: 10.1080/08874417.2013.11645627

Callan, D. E., and Dehais, F. (2019). "Neuroergonomics for aviation," in *Neuroergonomics* (Academic Press), 55–58.

Cegarra, J. (2008). A cognitive typology of scheduling situations: a contribution to laboratory and field studies. *Theor. Issues Ergon. Sci.* 9, 201–222. doi: 10.1080/14639220601095379

Cenfetelli, R. T. (2004). "Getting in touch with our feelings towards technology," in *Academy of Management Proceedings* (Manor, NY). doi: 10.5465/ambpp.2004.13862511

Charles, R. L., and Nixon, J. (2019). Measuring mental workload using physiological measures: a systematic review. *Appl. Ergon.* 74, 221–232. doi: 10.1016/j.apergo.2018.08.028

Chen, M., Han, J., Guo, L., Wang, J., and Patras, I. (2015). "Identifying valence and arousal levels via connectivity between EEG channels," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (Beijing), 63–69. doi: 10.1109/ACII.2015.7344552

Crawford, S., and Wiers, V. C. S. (2001). "From anecdotes to theory: a review of existing knowledge on the human factors of production scheduling," in *Human Performance in Planning and Scheduling*, eds B. L. MacCarthy and J. R. Wilson (CRC Press), 28.

Dehais, F., and Callan, D. (2019). "A neuroergonomics approach to human performance in aviation," in *Improving Aviation Performance through Applying Engineering Psychology*, ed C. Buitrago Torralvo (Boca Raton, FL: CRC Press), 123–141. doi: 10.4324/9780429492181-6

Dennett, D. C. (1988). "Quining qualia," in *Consciousness in Contemporary Science*, eds A. J. Marcel and E. Bisiach (New York, NY: Oxford University Press), 381–414.

Dimopoulos, C., Cegarra, J., Gavriel, G., Chouchourelou, A., and Papageorgiou, G. (2012). "Interdisciplinary modelling of scheduling environments: a case study," in *4th Production and Operations Management World Conference (PandOM 2012)* (Amsterdam), 1–10.

Djamasbi, S. (2007). Does positive affect influence the effective usage of a Decision Support System? *Decis. Support Syst.* 43, 1707–1717. doi: 10.1016/j.dss.2006.09.002

Djamasbi, S., and Loiacono, E. T. (2008). Do men and women use feedback provided by their Decision Support Systems (DSS) differently? *Decis. Support Syst.* 44, 854–869. doi: 10.1016/j.dss.2007.10.008

Evans, J. S. B. T., and Stanovich, K. E. (2013). Dual-process theories of higher cognition. *Perspect. Psychol. Sci.* 8, 223–241. doi: 10.1177/1745691612460685

Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interact. Comput.* 21, 133–145. doi: 10.1016/j.intcom.2008.10.011

Ferris, T., Sarter, N., and Wickens, C. D. (2010). "Cockpit automation," in *Human Factors in Aviation*, eds E. Salas and D. Maurino (Waltham, MA: Academic Press), 479–503. doi: 10.1016/B978-0-12-374518-7.00015-8

Han, S., Lerner, J. S., and Keltner, D. (2007). Feelings and consumer decision making: the appraisal-tendency framework. *J. Consum. Psychol.* 17, 158–168. doi: 10.1016/S1057-7408(07)70023-2

Herrmann, J. W. (2006). "Improving production scheduling : Integrating organizational, decision-making, and problem solving perpectives," in *Industrial Engineering Research Conference* (Orlanda, FL).

Hu, J., Nakanishi, M., Matsumoto, K., Tagaito, H., Inoue, K., Shima, K., et al. (2000). "A method of usability testing by measuring brain waves," in *Proceedings of the International Symposium on Future Software Technology* (Nanjing), 159–164.

Iyilikci, E., and Amado, S. (2018). The uncertainty appraisal enhances the prominent deck B effect in the Iowa gambling task. *Motiv. Emot.* 42, 1–16. doi: 10.1007/s11031-017-9643-5

Jeon, M. (2017). "Emotions and affect in human factors and human–computer interaction: taxonomy, theories, approaches, and methods," in *Emotions and Affect in Human Factors and Human-Computer Interaction*, ed M. Jeon. (London: Academic Press), 3–26. doi: 10.1016/B978-0-12-801851-4. 00001-X

Jepson, J. M., Kirytopoulos, K., and London, K. (2017). Exploring project managers' perception of stress when working in increasingly complex construction projects. *Construct. Econ. Building* 17, 47–67. doi: 10.5130/AJCEB.v17i3.5567

Jokinen, J. P. P. (2015). Emotional user experience: traits, events, and states. *Int. J. Hum. Comput. Stud.* 76, 67–77. doi: 10.1016/j.ijhcs.2014. 12.006

Jung, T., Kaß, C., Schramm, T., and Zapf, D. (2017). So what really is user experience? An experimental study of user needs and emotional responses as underlying constructs. *Ergonomics* 60, 1601–1620. doi: 10.1080/00140139.2017.1341555

Kahneman, D., and Frederick, S. (2007). Frames and brains: elicitation and control of response tendencies. *Trends Cogn. Sci.* 11, 45–46. doi: 10.1016/j.tics.2006.11.007

Lankton, N. K., Wilson, E. V., and Mao, E. (2010). Antecedents and determinants of information technology habit. *Inform. Manage.* 47, 300–307. doi: 10.1016/j.im.2010.06.004

Larco Martinelli, J. A., Fransoo, J., Gharehgozli, A., and Wiers, V. (2019). The scheduler's balancing act of sensing and reacting: a behavioural perspective on scheduling. *Int. J. Product. Res.* 58, 3944–3955. doi: 10.1080/00207543.2019.1636326

Larsen, R., and Pranzo, M. (2019). A framework for dynamic rescheduling problems. *Int. J. Product. Res.* 57, 16–33. doi: 10.1080/00207543.2018.14 56700

Lee, S. J., and Seong, P. H. (2009). Experimental investigation into the effects of decision support systems on operator performance. *J. Nucl. Sci. Technol.* 46, 1178–1187. doi: 10.1080/18811248.2009.97 11631

Lerner, J. S., and Keltner, D. (2000). Beyond valence: toward a model of emotion-specific influences on judgement and choice. *Cogn. Emotion* 14, 473–493. doi: 10.1080/026999300402763

Lerner, J. S., and Keltner, D. (2001). Fear, anger and risk. *J. Pers. Soc. Psychol.* 81, 146–159. doi: 10.1037/0022-3514.81.1.146

Lerner, J. S., Li, Y., Valdesolo, P., and Kassam, K. S. (2015). Emotion and decision making. *Ann. Rev. Psychol.* 66, 799–823. doi: 10.1146/annurev-psych-010213-115043

Lim, K. H., O'Connor, M. J., and Remus, W. E. (2005). The impact of presentation media on decision making: does multimedia improve the effectiveness of feedback? *Inform. Manage.* 42, 305–316. doi: 10.1016/j.im.2003.12.012

Lotte, F., and Roy, R. N. (2019). Brain–computer interface contributions to neuroergonomics. *Neuroergonomics* 2019, 43–48. doi: 10.1016/B978-0-12-811926-6.00007-5

MacCarthy, B. L., Wilson, J. R., and Crawford, S. (2001). Human performance in industrial scheduling: a framework for understanding. *Hum. Factors Ergon. Manuf.* 11, 299–320. doi: 10.1002/hfm.1016

Mailliez, M., Bollon, T., Graton, A., and Hot, P. (2020). Can the induction of incidental positive emotions lead to different performances in sequential decision-making? *Cogn. Emotion* 34, 1509–1516. doi: 10.1080/02699931.2020.1760213

Manzey, D., Reichenbach, J., and Onnasch, L. (2012). Human performance consequences of automated decision aids: the impact of degree of automation and system experience. *J. Cogn. Eng. Decision Making* 6, 57–87. doi: 10.1177/1555343411433844

Mehler, B., Reimer, B., Coughlin, J. F., and Dusek, J. A. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transport. Res. Rec.* 2138, 6–12. doi: 10.3141/2138-02

Navarro, J., Heuveline, L., Avril, E., and Cegarra, J. (2018). Influence of human-machine interactions and task demand on automation selection and use. *Ergonomics* 61, 1601–1612. doi: 10.1080/00140139.2018.1501517

Onnasch, L., and Hösterey, S. (2019). Stages of decision automation: impact on operators' role, awareness and monitoring. *Proc. Human Factors Ergon. Soc. Ann. Meet.* 63, 282–286. doi: 10.1177/1071181319631126

Reppa, I., McDougall, S., Sonderegger, A., and Schmidt, W. C. (2020). Mood moderates the effect of aesthetic appeal on performance. *Cogn. Emotion* 35, 15–29. doi: 10.1080/02699931.2020.1800446

Riveiro, M., Helldin, T., and Falkman, G. (2014). "Influence of meta-information on decision-making: lessons learned from four case studies," in *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, CogSIMA 2014* (San Antonio, TX), 14–20. doi: 10.1109/CogSIMA.2014.6816534

Rossit, D. A., Tohmé, F., and Frutos, M. (2019). Production planning and scheduling in Cyber-Physical Production Systems: a review. *Int. J. Comput. Integr. Manufact.* 32, 385–395. doi: 10.1080/0951192X.2019.1605199

Röttger, S., Bali, K., and Manzey, D. (2009). Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task. *Ergonomics* 52, 512–523. doi: 10.1080/00140130802379129

Roy, R. N., Charbonnier, S., and Campagne, A. (2013). "Probing ECG-based mental state monitoring on short time segments," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Osaka), 6611–6614. doi: 10.1109/EMBC.2013.6611071

Roy, R. N., Drougard, N., Gateau, T., Dehais, F., and Chanel, C. P. C. (2020). How can physiological computing benefit human-robot interaction? *Robotics* 9:100. doi: 10.3390/robotics9040100

Roy, R. N., and Frey, J. (2016). "Neurophysiological markers for passive brain-computer interfaces," in *Brain-Computer Interfaces 1*, eds M. Clerc, L. Bougrain, and F. Lotte (Hoboken, NJ: John Wiley and Sons, Inc), 85–100. doi: 10.1002/9781119144977.ch5

Ruff, H. A., Narayanan, S., and Draper, M. H. (2002). Human interaction with levels of automation and decision-aid fidelity in the supervisory control of multiple simulated unmanned air vehicles. *Presence* 11, 335–351. doi: 10.1162/105474602760204264

Saariluomaand, P., and Jokinen, J. P. P. (2014). Emotional dimensions of user experience: a user psychological analysis. *Int. J. Hum. Comput. Interact.* 30, 303–320. doi: 10.1080/10447318.2013.858460

Sagar, K., and Saha, A. (2017). A systematic review of software usability studies. *Int. J. Inform. Technol.* 3, 1–24. doi: 10.1007/s41870-017-0048-1

Saint-Lot, J., Imbert, J.-P., and Dehais, F. (2020). "Red alert: a cognitive countermeasure to mitigate attentional tunneling," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY), 1–6. doi: 10.1145/3313831.3376709

Sanderson, P. M. (1989). The human planning and scheduling role in advanced manufacturing systems: an emerging human factors domain. *Hum. Factors* 31, 635–666. doi: 10.1177/001872088903100602

Scherer, K. R. (2009). The dynamic architecture of emotion: evidence for the component process model. *Cogn. Emotion* 23, 1307–1351. doi: 10.1080/02699930902928969

Schiebener, J., and Brand, M. (2015). Decision making under objective risk conditions: a review of cognitive and emotional correlates, strategies, feedback processing, and external influences. *Neuropsychol. Rev.* 25, 171–198. doi: 10.1007/s11065-015-9285-x

Schorr, A. (2001). "Subjective measurement in appraisal research," in *Appraisal Processes in Emotion: Theory, Methods, Research*, eds K. R. Scherer, A. Schorr, and Tom Johnstone (Oxford: Oxford University Press), 331–349.

Smith, C. A., and Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *J. Pers. Soc. Psychol.* 48, 813–838. doi: 10.1037/0022-3514.48.4.813

Smith, P. J., and Geddes, N. D. (2003). "A cognitive systems engineering approach to the design of Decision Support systems," in *Handbook of Human-Computer Interaction*, ed E. Lawrence (Upper Saddle River, NJ: Prentice-Hall, Inc), 656–676.

Stein, M. K., Newell, S., Wagner, E. L., and Galliers, R. D. (2015). Coping with information technology. *MIS Q.* 39, 367–392.

Thüring, M., and Mahlke, S. (2007). Usability, aesthetics and emotions in human-technology interaction. *Int. J. Psychol.* 42, 253–264. doi: 10.1080/00207590701396674

Tiedens, L. Z., and Linton, S. (2001). Judgment under emotional certainty and uncertainty: the effects of specific emotions on information processing. *J. Pers. Soc. Psychol.* 81, 973–988. doi: 10.1037/0022-3514.81.6.973

VaezMousavi, S. M., Barry, R. J., and Clarke, A. R. (2009). Individual differences in task-related activation and performance. *Physiol. Behav.* 98, 326–330. doi: 10.1016/j.physbeh.2009.06.007

van der Kleij, R., Hueting, T., and Schraagen, J. M. (2018). Change detection support for supervisory controllers of highly automated systems: effects on performance, mental workload, and recovery of situation awareness following interruptions. *Int. J. Ind. Ergon.* 66, 75–84. doi: 10.1016/j.ergon.2018.02.010

Vieira, G. E., Herrmann, J. W., and Lin, E. (2003). Rescheduling manufacturing systems: a framework of strategies, policies, and methods. *J. Scheduling* 6, 39–62. doi: 10.1023/A:1022235519958

Vossing, M. (2017). Towards managing complexity and uncertainty in field service technician planning. *IEEE 19th Conf. Bus. Inform.* 1, 312–319. doi: 10.1109/CBI.2017.50

Walters, B., and Barnes, M. J. (2002). Manpower, skill and fatigue analysis of Future Unmanned Aerial Vehicle (UAV) environments. *Proc. Human Factors Ergon. Soc. Ann. Meet.* 46, 1905–1909. doi: 10.1177/154193120204602308

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

Zander, T. O., and Kothe, C. (2011). Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *J. Neural Eng.* 8:025005. doi: 10.1088/1741-2560/8/2/025005

Zikos, S., Krinidis, S., Ioannidis, D., Tzovaras, D., Ziazios, K., and Metaxa, I. (2018). "Efficient location-aware scheduling of maintenance tasks in shop floors," in *Proceedings - 2017 4th International Conference on Mathematics and Computers in Sciences and in Industry, MCSI 2017* (Erasmus MC), 119–123. doi: 10.1109/MCSI.2017.28

frontiers
in Neuroergonomics

Comparing Explicit and Implicit Measures for Assessing Cross-Cultural Food Experience

Daisuke Kaneko [1,2]*, Ivo Stuldreher [3,4], Anne J. C. Reuten [3], Alexander Toet [3], Jan B. F. van Erp [3,4] and Anne-Marie Brouwer [3]

[1] Kikkoman Europe R&D Laboratory B.V., Wageningen, Netherlands, [2] Microbiology and Systems Biology, Netherlands Organisation for Applied Scientific Research TNO, Zeist, Netherlands, [3] Perceptual and Cognitive Systems, Netherlands Organisation for Applied Scientific Research TNO, Soesterberg, Netherlands, [4] Human Media Interaction Lab, University of Twente, Enschede, Netherlands

The present study investigated the potential of implicit physiological measures to provide objective measures of affective food experience in contrast to explicit self-report ratings in a cross-cultural context. Dutch and Thai participants viewed 120 food images portraying universal food image categories (regular and molded food) and cultural food image categories (typically Dutch and Thai food). The universal food images were taken as ground truth high and low valence stimuli, where we assumed no genuine difference in affective experience between nationalities. In contrast, for the cultural food images, we did expect a genuine difference between nationalities. Participants were asked to rate valence, arousal and liking of each food image. In addition, heart rate (HR) and phasic electrodermal activity (EDA) responses to the images were recorded. Typically Asian and Western response biases were found for explicit ratings of regular and molded food with an extreme response style for Dutch, and a middle response style for Thai participants. However, such bias was not observed in HR. For cultural food image categories, HR showed the hypothesized interaction between participant nationality and food image category, reflecting the expected genuine difference between nationalities in affective food experience. Besides presenting participants with images, we also asked participants to taste typically Thai and Dutch drinks. Similar to images, a significant interaction between participant nationality and cultural food category was found for HR. An interaction was also found for sip size, while this was not seen in explicit measures. We attribute this to differences in the moment that these measures were taken. In this study, phasic EDA did not appear to be a sensitive measure of affective food experience, possibly since stimuli mostly differed in valence rather than arousal. To conclude, our study constitutes an example where cultural bias negatively affected the accuracy of self-reports, and only the implicit physiological measures followed the prior expectations of genuine food experience, indicating the potential of these measures to study cross-cultural food experience.

Keywords: cross-cultural, response bias, explicit, implicit, physiological measures, HR, EDA, sip size

INTRODUCTION

To predict whether consumers will choose a certain food product, their emotional response when experiencing this product is considered to be an important predictor (Dalenberg et al., 2014; Gutjar et al., 2015; Köster and Mojet, 2015; Samant et al., 2017). With the current globalizing trend, measuring these emotions cross-culturally is important for international food marketers (Rozin, 1988, 2006; Meiselman, 2013, 2015).

Assessment of food-evoked emotions is predominantly based on explicit measures ("conscious" self-report ratings) rather than on implicit measures ("unconscious" physiological and behavioral measures) (Lagast et al., 2017; Kaneko et al., 2018). Explicit measures are relatively easy to apply, practical for quantitative analysis, and widely believed reliable for most sensory and psychological studies (Lawless and Heymann, 2010; Dorado et al., 2016). Many large international companies explore food product-elicited emotions across different countries by simply translating emotion questionnaires into multiple languages (Meiselman, 2015). However, the drawback of using explicit self-report questionnaires for assessing affective experience cross-culturally is that cultural background influences how people "self-report," what emotional language they use, and how they use rating scales to describe their own food-evoked emotions (Meiselman, 2015; Van Zyl and Meiselman, 2015, 2016; Silva et al., 2016; Ares, 2018). A review by Meiselman (2015) questioned whether simple translations between languages capture the local meaning of emotional words in a questionnaire well enough – people raised in different cultures may not experience the same emotions evoked by the same stimuli, and evoked emotions may not be expressed in the same manner. For example, Uchida and Kitayama (2009) analyzed American and Japanese descriptions of "happiness" and showed that Japanese associate happiness with "social harmony" while Americans associate it with positive experience of personal achievement. Concerning culture-dependent use of rating scales, Western respondents have been found to have an "Extreme Response Style" (ERS) (using the extremes of rating scales), whereas Asian respondents more often use a "Middle Response Style" (MRS) (using the neutral part of the scale) (Chen et al., 1995; Harzing, 2006; Kaneko et al., 2019b). Problems with translating between languages, and intercultural differences in terms of using rating scales, could potentially be overcome by implicit measures, which reflect fast, non-conscious, and uncontrollable responses (Soto et al., 2005; Lagast et al., 2017; Ares, 2018; Kaneko et al., 2018).

Several physiological measures have been studied in the context of probing affective experience when tasting and viewing food or food images. Among them, heart rate (HR) and electrodermal activity (EDA) are the most often used implicit physiological measures in recent consumer research. These measures have been shown to distinguish between tasting different beverages, chocolates, liked, and disliked food (de Wijk et al., 2012; Danner et al., 2014; Torrico et al., 2018; Kaneko et al., 2019a). Outside the food domain, several studies have compared implicit physiological responses to different types of stimuli between individuals from different cultures. One study showed weaker electrodermal responses to disgust-eliciting film clips in

Asian-American compared to European-American participants (Soto et al., 2016). No difference between cultural groups was found for physiological responses to stimuli such as acoustic startle [Chinese-American and Mexican-American groups; (Soto et al., 2005)], emotional films [Chinese-American and European-American groups; (Tsai et al., 2000)], and reliving of intense emotional episodes [Hmong-American and European-American groups; (Tsai et al., 2002)]. These results may be taken to mean that Asian-Americans and other Americans "really" differed in emotional experience when watching disgust-eliciting movies, and not when experiencing the other types of stimuli.

To the best of our knowledge, only two studies in the food domain used implicit measures to investigate psychophysiological effects of food products between different nationalities [Asian who spent <2 years in Australia and Australian groups; (Torrico et al., 2018, 2019)]. These studies used rating scales as well as a camera to monitor heart rate, skin temperature and facial expressions to investigate cross-cultural effects of viewing universal and culture-specific food (Torrico et al., 2018, 2019). Their results indicated that rated food liking is positively correlated to familiarity, and that skin temperature differentiates between cultural groups when tasting culture-specific food samples (Torrico et al., 2019) while no physiological differences between cultural groups exist when tasting (universal) chocolate samples (Torrico et al., 2018). These results are in line with the idea that physiological measures reflect the "true" emotion: cultural groups do not differ when tasting universal stimuli, and they do differ (in skin temperature) when tasting samples that are expected to genuinely elicit different emotions. No physiological effects were found besides skin temperature, but we should note that camera-based analysis is usually less precise and suffers more from artifacts caused by (chewing) movements, head orientation, and lighting conditions compared to traditional sensors (Kranjec et al., 2014; Bach et al., 2015; Hassan et al., 2017).

In the present study we investigated whether implicit physiological measures (HR and phasic EDA recorded using traditional sensors) can contribute to comparing affective food experiences across cultures objectively without cultural response biases that affect explicit self-report methods. We compared explicit and implicit responses between two cultural participant groups, Dutch (representative for ERS) and Thai (representative for MRS), toward universal food images (regular and molded food) and cultural food images (typically Dutch and Thai food). We selected universal and cultural food image categories so that we could assume a genuine difference in emotional experience between Dutch and Thai participants for the two types of cultural foods, but no genuine difference for the two types of universal foods. For the latter category, we can safely assume a ground truth affective experience of low valence (unpleasant) and high arousal for the molded compared to the regular food images, in both Dutch and Thai participants. A lack of effect of universal food category would therefore indicate that the measure is insensitive. No differences between the two nationalities are expected for implicit measures for universal food image categories while we expect differences on explicit measures due to culturally-dependent response bias. On the other hand, we expect response

differences between Dutch and Thai participants on both explicit and implicit measures for the cultural food image categories. Viewing food images happens in real life, and is practical in (psychophysiological) experiments in which a large number of trials is desirable. However, viewing food images is expected to elicit different responses than being confronted and tasting real food (de Wijk et al., 2012). To extend our study beyond viewing food images, we also examined implicit and explicit affective responses in Dutch and Thai participants when tasting typically Thai and typically Dutch drinks.

Our specific hypotheses are as follows:

1) Nationality affects explicit measures for universal food images (regular and molded food) due to a culturally dependent response bias. ERS is expected for Dutch participants, and MRS for Thai participants.

2) Nationality affects explicit measures for the cultural food images due to a genuine difference in affective experience caused by a difference in familiarity with the types of food (in addition to a culturally dependent response bias), resulting in an interaction between participant nationality and cultural food category.

3) Implicit physiological measures are affected by universal food image category (regular vs. molded food), but the effect is the same for both nationalities (i.e., no interaction between participant nationality and universal food image category).

4) Implicit physiological measures toward cultural food image categories (Dutch and Thai food) reflect genuine differences in affective experience between participant groups, resulting in an interaction between participant nationality and cultural food image category.

5) Similar to food images, explicit and implicit responses to tasting cultural drinks show an interaction effect between participant nationality and cultural drink category.

METHODS

Participants

42 Thai participants were recruited from Chulalongkorn University in Thailand and 45 Dutch participants were recruited from the participant pool of the research institute where the main part of the research was conducted (TNO Soesterberg, The Netherlands). The recruitment process excluded people with color vision deficiencies; food allergies; diets such as vegetarian, vegan or religion-related; an immigration background; an eating disorder diagnosed in the last 3 years. Also, people who had visited The Netherlands (for Thai participants) or Thailand (for Dutch participants) and who had lived abroad for more than 1 month could not participate. Participants were asked not to eat for 1 h before testing. The experimental protocol was approved by the TNO Institutional Review Board (Ethical Approval Ref: 2019-033) and was in accordance with the revised Helsinki Declaration (World Medical Association, 2013). All participants signed an informed consent sheet before the experiment started and received a reward to thank them for participating in the study after completing the experiment.

Materials
Food Images

Food images were selected from the Cross Cultural Food Image Database (CROCUFID) (Toet et al., 2019), which is a collection of food and non-food images, photographed on a standardized plate using a standardized photographing protocol (Charbonnier et al., 2016). From this database, we selected 60 "universal" food images which are expected to be familiar to participant of both nationalities (47 regular food images and 13 molded images); and 60 "cultural" food images (30 typically Dutch and 30 typically Thai food images). As shown in **Figure 1**, the national flag of the food's origin was presented on the right bottom of each image to ensure participants recognizing and interpreting the food in a similar way. Universal dishes were accompanied by an image of a globe.

Drinks

As a typically Dutch drink, a popular long seller yogurt drink, Fristi (Friesland Campina B.V. the Netherlands), was used. As a typically Thai drink we selected a Chrysanthemum tea drink (Vitasoy International Holdings Limited, Thailand). Both drinks were served in white plain cups.

Rating Scales and Implicit Behavioral Measure

The following rating scales were used to rate emotions evoked by viewing food images and tasting drinks, and to check for the familiarity of the participants with the stimuli:

EmojiGrid

An intuitive visual self-report tool that has been specifically developed for the assessment of food-evoked emotions and that has shown to be suitable for cross-cultural testing (Toet et al., 2018; Kaneko et al., 2019b). Participants report their emotion by clicking the appropriate location in the grid, where each location is associated with a valence and arousal score, ranging from 0 (lowest) to 100 (highest). The Emojigrid is depicted in **Figure 2**.

Hedonic Liking Scale

9-point scale with anchors for each point. The anchors are: (1) "dislike extremely," (2) "dislike very much," (3) "dislike moderately," (4) "dislike slightly," (5) "neither like nor dislike," (6) "like slightly," (7) "like moderately," (8) "like very much," and (9) "like extremely" (Lim, 2011).

Familiarity Scale

5-point scale with anchors for each point. Anchors of this five-point scale were labeled: (1) "I do not recognize it," (2) "I recognize it, but I have not tasted it," (3) "I have tasted it," (4) "I occasionally eat it," (5) "I regularly eat it" [adapted from Tuorila et al. (2001)]. This scale was used to check whether Thai and Dutch cultural food images were more familiar to participants of the matching nationality compared to the other nationality.

A scale was used during the tasting session to measure sip size of both cultural drinks by weighing the drink before and after the participant had taken a sip, following the procedure in a previous study (Kaneko et al., 2019a).

FIGURE 1 | Stimulus examples of **(A)** universal molded food images (melon and strawberries) and universal regular food images (cucumber and French fries), **(B)** Dutch food images (herring, wheat bread, liquorice candy, stroopwafels), and **(C)** Thai food images (dragon fruit, seaweed chips, som tam, grass jelly).

Physiological Recording Equipment (Electrocardiogram and Electrodermal Activity)

Electrocardiogram [ECG; for heart rate (HR)] and electrodermal activity (EDA; for phasic EDA) were recorded using an Active Two MkII system (Biosemi B.V., Amsterdam, the Netherlands), with a sampling frequency of 512 Hz. ECG electrodes were placed on the right clavicle and on the lowest floating left rib. EDA was measured by placing gelled electrodes on the fingertips of the index finger and the middle finger of the non-dominant hand. Two reference electrodes were attached to the temporal bone behind the ears.

Experimental Design and Procedure

After participants arrived at the laboratory (depending on the nationality, either located at Chulalongkorn University, Bangkok, Thailand; or TNO Soesterberg, the Netherlands), they were told that the experiment consisted of a "tasting session," a "viewing session," and a "rating familiarity session." The experimenter also explained that ECG and EDA sensors would be attached to measure HR and EDA during the experiment. Participants signed the informed consent form and were seated in a comfortable chair in front of an experimental presentation notebook. Then, the ECG, EDA, and reference electrodes were attached, and all signals were checked. HR and EDA were recorded during the tasting and viewing sessions. A schematic image of the study is shown in **Figure 2**. The total duration of the experiment was ~75 min.
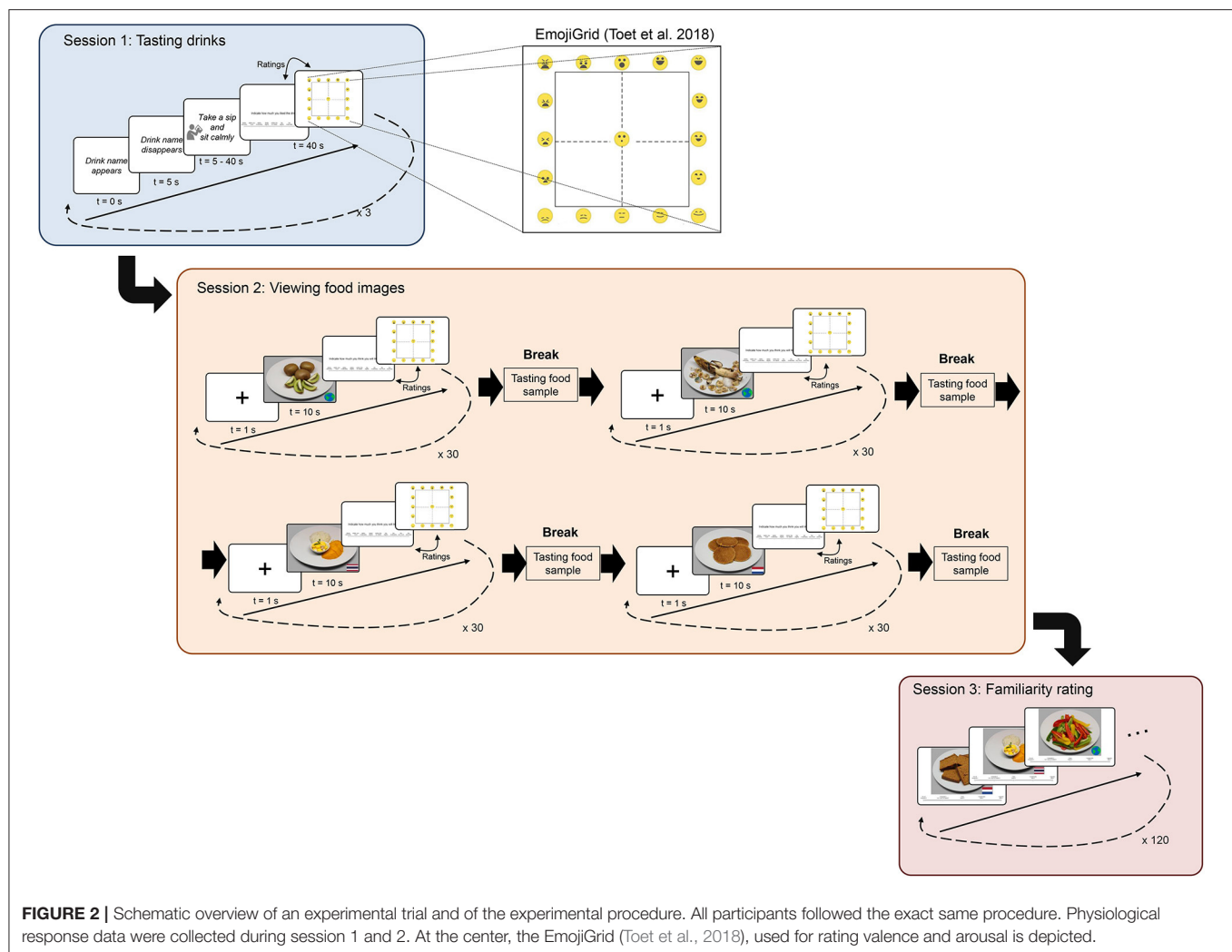
Tasting Session

The procedure of the tasting session followed that used in a previous study (Kaneko et al., 2019a). Before the tasting session started, the experimenter showed and explained how to take a sip and to put the cup down after the sip, and participants

performed a practice trial with water. After this there was time for additional practice or instructions when needed. The testing procedure started with the presentation of the name of the drink on the screen. This was the sign for the experimenter to place the appropriate drink in front of the participant. After 5 s, the name of the drink disappeared, which was the sign for the participant to take one sip. After taking the sip, the participant put the cup down, sat still and looked at a blank white screen. Thirty-five seconds after the name of the drink had disappeared from the screen, the EmojiGrid and hedonic scale appeared in successive, randomized order. Order was randomized so that the ratings from the two scales could be compared and would not be confounded by a possible order effect. After rating, the name of the next drink appeared on the screen. This procedure was repeated until three drinks had been served: first water for practice, followed by the Thai and Dutch drinks in counterbalanced order.

Viewing Session

After a short break, the instruction for the viewing session appeared on a screen, and participants had a chance to ask any questions to the experimenter. In this session participants viewed a total of 120 food images in two counterbalanced blocks (universal and cultural food image categories) each consisting of 60 randomized images. A fixation cross was presented for 1 s, after which a food image was presented for 10 s. After 10 s of viewing time, the EmojiGrid and hedonic scale appeared in successive, randomized order. Participants had unlimited time to provide their ratings using the mouse but were instructed to follow their initial impression. In an effort to increase participants' engagement with the images, they were told before starting the viewing session that they would be asked to taste

FIGURE 2 | Schematic overview of an experimental trial and of the experimental procedure. All participants followed the exact same procedure. Physiological response data were collected during session 1 and 2. At the center, the EmojiGrid (Toet et al., 2018), used for rating valence and arousal is depicted.

some of the depicted food images they rated during each of four breaks. This part of the instruction was formulated in such a way that the participants were led to believe that they could be asked to eat an unpleasant (molded) food item: "There are four short breaks in this session. During these breaks, we will serve you one of the foods depicted in the images that you just saw. We ask you to taste this food. You are permitted to refuse, but we hope you will taste it." Tasting breaks were introduced after each half block (30 images), and tasting samples were sliced banana pieces, peanut chocolate candies, a seaweed chip (typically Thai food item) and a small "stroopwafel" (typically Dutch food item).

Familiarity Rating Session

After finishing the viewing session, participants were instructed to rate their familiarity with all drinks and all food images they rated in the previous sessions. For drinks, the name of the drinks appeared successively on the screen in random order, accompanied by the familiarity scale. Actual tasting was omitted. Food images also appeared successively in randomized order, accompanied by the familiarity scale. Lastly, participants filled

out a short demographic questionnaire, asking about age, gender, height, and weight.

Physiological Data Processing

Data from three Thai participants were discarded due to failure of physiological recordings. Data processing was done using Matlab 2020a software (Mathworks, Natick, MA, USA).

Inter-beat intervals were extracted from ECG following Pan and Tompkins (1985) using a Matlab implementation from Sedghamiz (2014). By inversing the inter-beat interval, a heart rate semi-time series was obtained. Intervals exceeding the absolute threshold of 160 bpm or intervals deviating more than 20% from the previous interval were removed. The heart rate semi-time series was transformed to a regular timeseries at a 10 Hz sampling frequency using a piecewise cubic spline interpolation.

Raw EDA was down-sampled to 10 Hz. Continuous Decomposition Analysis as implemented in the Ledalab toolbox for Matlab was used to separate the tonic (slow) and phasic (fast)

TABLE 1 | Demographic variables for the Dutch and Thai participants.

Cultural group	N	Female	Male	Age	BMI
Dutch participants	45	28	17	21.2 (±1.7)	22.29 (±2.77)
Thai participants	42	24	18	20.6 (±1.5)	21.58 (±3.60)

components of the EDA (Benedek and Kaernbach, 2010). In further analysis, only the phasic component is considered.

For food images, the pre-processed continuous physiological data were divided in epochs that were time-locked to stimulus onset for each image and each participant. The epoch comprises data ranging from fixation-cross onset to 10 s after fixation-cross offset. Response traces were baseline corrected based on the average value of the 1 s that the fixation-cross was presented.

For drinks, the pre-processed continuous physiological data were divided in epochs ranging from 5 s after drink name onset to 35 s after drink name offset (which was the sign for the participant to pick up the cup and take a sip), for each drink and each participant. Response traces were baseline corrected based on the average value of the 5 s during drink name onset.

Outliers were detected using Matlab's "isoutlier" function. Image epochs for which the average physiological response value was more than five median absolute deviations away from the median value across images were removed. This resulted in a removal of 0.06% of the HR data and 8.6% of the phasic EDA data. No outlier detection was conducted for physiological response value for drinks.

Grand-average response traces were obtained for each participant and each of the food image categories [universal food image category (regular and molded), cultural food image category (Dutch and Thai), and cultural drink category (Dutch and Thai)] by averaging over all data traces corresponding to the food image categories of each participant. We then computed the mean HR and mean phasic EDA for each participant and each food image category over the 10 s following image onset and over 35 s following drink name offset.

Statistical Analysis

Statistical analysis was conducted with SPSS ver. 25 (IBM, USA). Two-way mixed ANOVAs were performed on each explicit and implicit measure in response to the universal food images, the cultural food images and the cultural drinks. In all of these analyses, the between-subject factor was participant nationality (Dutch or Thai), and the within-subject factor was respectively universal food image category (regular and molded); cultural food image category (Dutch and Thai); and cultural drink category (Dutch and Thai).

RESULTS

Demographics

Table 1 shows the demographic descriptives for Dutch and Thai participants. The two groups did not differ in age [$t_{(85)} = 1.662$, $p = 0.100$] and BMI [$t_{(85)} = 1.022$, $p = 0.310$] according to Welch's t-test.
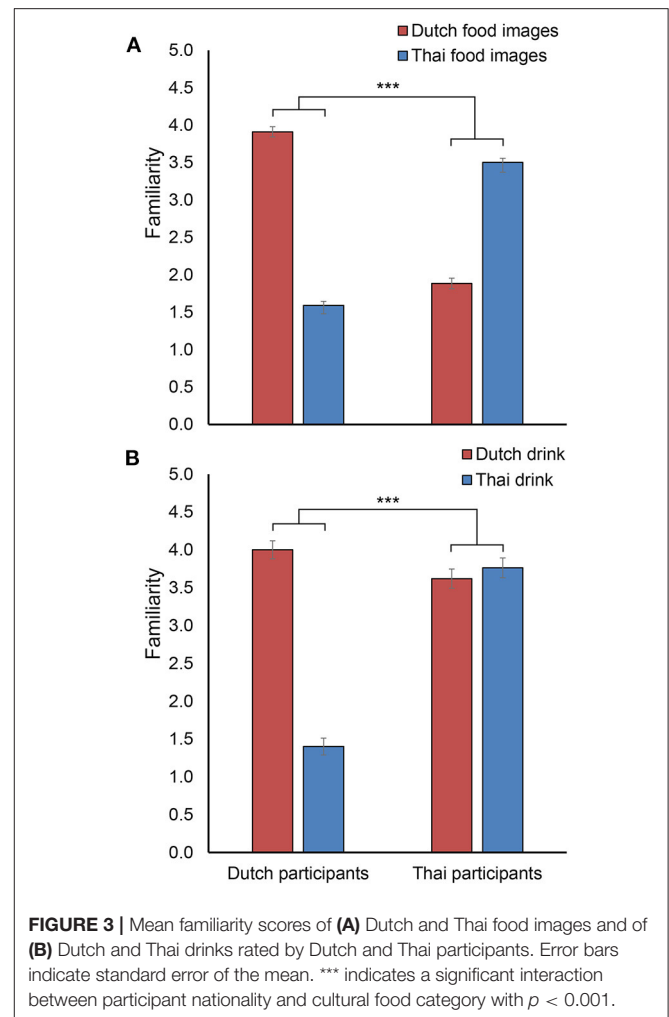
FIGURE 3 | Mean familiarity scores of **(A)** Dutch and Thai food images and of **(B)** Dutch and Thai drinks rated by Dutch and Thai participants. Error bars indicate standard error of the mean. *** indicates a significant interaction between participant nationality and cultural food category with $p < 0.001$.

Familiarity Scores

Familiarity ratings are presented in **Figures 3A,B**.

Figure 3A shows the expected pattern of Dutch participants rating Dutch food as more familiar than Thai food and Thai participants rating Thai food as more familiar than Dutch food. This pattern is statistically corroborated by a significant interaction between participant nationality and cultural food image category [$F_{(1, 85)} = 1299.52$, $p < 0.001$, $\eta_p^2 = 0.939$], and indicates that the selection of cultural food images was appropriate. The two-way mixed ANOVA showed no significant main effect of participant nationality on familiarity scores [$F_{(1, 85)} = 0.69$, $p = 0.407$], and a significant main effect of cultural food image category [$F_{(1, 85)} = 41.19$, $p < 0.001$, $\eta_p^2 = 0.326$] on familiarity scores, with Dutch food images receiving higher familiarity ratings overall. As a comparison, regular food images from the set of universal food images, received very similar familiarity ratings as food images from the own culture: they received an average score of 4.1 from Dutch participant and an average score of 3.4 from Thai participants.

Figure 3B and the statistical analysis showed the expected interaction between participant nationality and cultural drink category [$F_{(1, 85)} = 147.88$, $p < 0.001$, $\eta_{\mathrm{p}}^2 = 0.635$] even though familiarity scores for the Dutch and Thai drinks were similar for Thai participants, indicating that Thai participants reported to be quite familiar with yogurt drinks. The analysis showed a significant main effect of cultural drink category [$F_{(1, 85)} = 118.68$, $p < 0.001$, $\eta_{\mathrm{p}}^2 = 0.583$] on familiarity scores, where the Dutch drink received overall higher familiarity ratings than the Thai drink. In addition, there was a significant main effect of participant nationality [$F_{(1, 85)} = 57.19$, $p < 0.001$, $\eta_{\mathrm{p}}^2 = 0.402$], with Thai participants rating significantly higher familiarity scores than Dutch participants.

Explicit Measures for Universal and Cultural Food Categories (Hypotheses 1 and 2)

Universal Food Category

The mean Emojigrid valence and arousal scores for all universal food images (regular and molded) rated by both Dutch and Thai participants are plotted in **Figure 4**. The figure shows the typically found *U*-shaped distribution between valence and arousal scores (Toet et al., 2018; Kaneko et al., 2019b). As expected, scores for molded food images are situated on the left (low valence) while scores for regular food images are on the right (high valence). The expected culture-dependent response bias was observed in this figure: Thai participants use a smaller portion of the valence and arousal scale compared to Dutch participants. Dutch rated molded food images as more extreme in low valence, and regular food images more extreme in high valence than Thai participants; and Dutch participants rated both images as more arousing compared to Thai participants. These observations are confirmed by the following figures and analyses.

Figures 5A,B represent average valence and arousal scores for each participant nationality and universal food image category. The two-way mixed ANOVA on valence scores showed a significant interaction between participant nationality and universal food image category [$F_{(1, 85)} = 51.32$, $p < 0.001$, $\eta_{\mathrm{p}}^2 = 0.376$], indicating that Dutch participants rated valence more extremely (relatively higher valence rating for regular food images and relatively lower valence rating for molded food images compared to Thai participants). There was also a significant main effect of food image category [$F_{(1, 85)} = 2583.37$, $p < 0.001$, $\eta_{\mathrm{p}}^2 = 0.968$], indicating lower rated valence for molded than for regular food images, and a main effect of participant nationality [$F_{(1, 85)} = 9.10$, $p = 0.003$, $\eta_{\mathrm{p}}^2 = 0.097$] with a somewhat higher overall valence score for Dutch than Thai participants. For arousal, there was a significant main effect of participant nationality [$F_{(1, 85)} = 8.89$, $p = 0.004$, $\eta_{\mathrm{p}}^2 = 0.095$], indicating higher arousal scores for Dutch participants than for Thai participants. There was also an effect of food image category [$F_{(1, 85)} = 61.33$, $p < 0.001$, $\eta_{\mathrm{p}}^2 = 0.419$], showing higher arousal scores for molded food images than for regular food images. There was no significant interaction between participant nationality and universal food image category [$F_{(1, 85)} = 0.005$, $p = 0.942$].
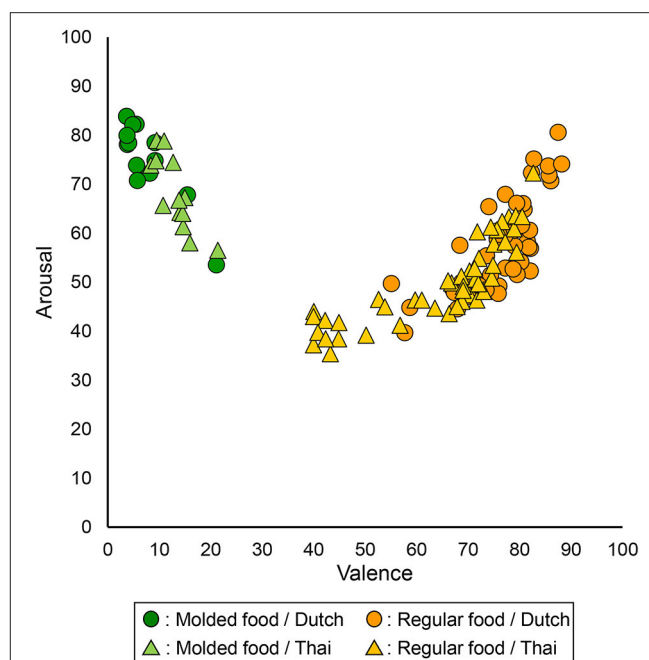

FIGURE 4 | Mean valence (x-axis) and arousal (y-axis) scores for all 60 universal (regular and molded) food images. Each data point represents the mean score from either Dutch or Thai participants for each food image.

Figure 5C shows hedonic liking scores averaged for each participant nationality and universal food image category. The pattern of results was identical to that of valence ratings, with a significant interaction between participant nationality and universal food image category [$F_{(1, 85)} = 63.78$, $p < 0.001$, $\eta_{\mathrm{p}}^2 = 0.429$], and main effects of both participant nationality [$F_{(1, 85)} = 11.45$, $p = 0.001$, $\eta_{\mathrm{p}}^2 = 0.119$] and universal food image category [$F_{(1, 85)} = 3152.01$, $p < 0.001$, $\eta_{\mathrm{p}}^2 = 0.974$].

Cultural Food Category

The mean EmojiGrid valence and arousal scores for Dutch and Thai cultural food image categories are shown in **Figures 6A,B**. The two-way mixed ANOVA showed the hypothesized interaction between participant nationality and cultural food image category on valence [$F_{(1, 85)} = 129.01$, $p < 0.001$, $\eta_{\mathrm{p}}^2 = 0.603$], indicating that Dutch participants rated higher valence for Dutch food images compared to Thai food images, whereas Thai participants showed the opposite pattern. Thus, participants rated food images that match their own nationality as more pleasant. There were also significant main effects of participant nationality [$F_{(1, 85)} = 4.71$, $p = 0.033$, $\eta_{\mathrm{p}}^2 = 0.053$] and cultural food image category [$F_{(1, 85)} = 22.82$, $p < 0.001$, $\eta_{\mathrm{p}}^2 = 0.212$] on valence scores, indicating overall higher scores for Dutch food images, and higher scores by Dutch participants. As for arousal scores, a significant interaction between participant nationality and cultural food image category [$F_{(1, 85)} = 8.23$, $p = 0.005$, $\eta_{\mathrm{p}}^2 = 0.088$] indicated that Dutch participants rated relatively higher arousal for Dutch food images than for Thai
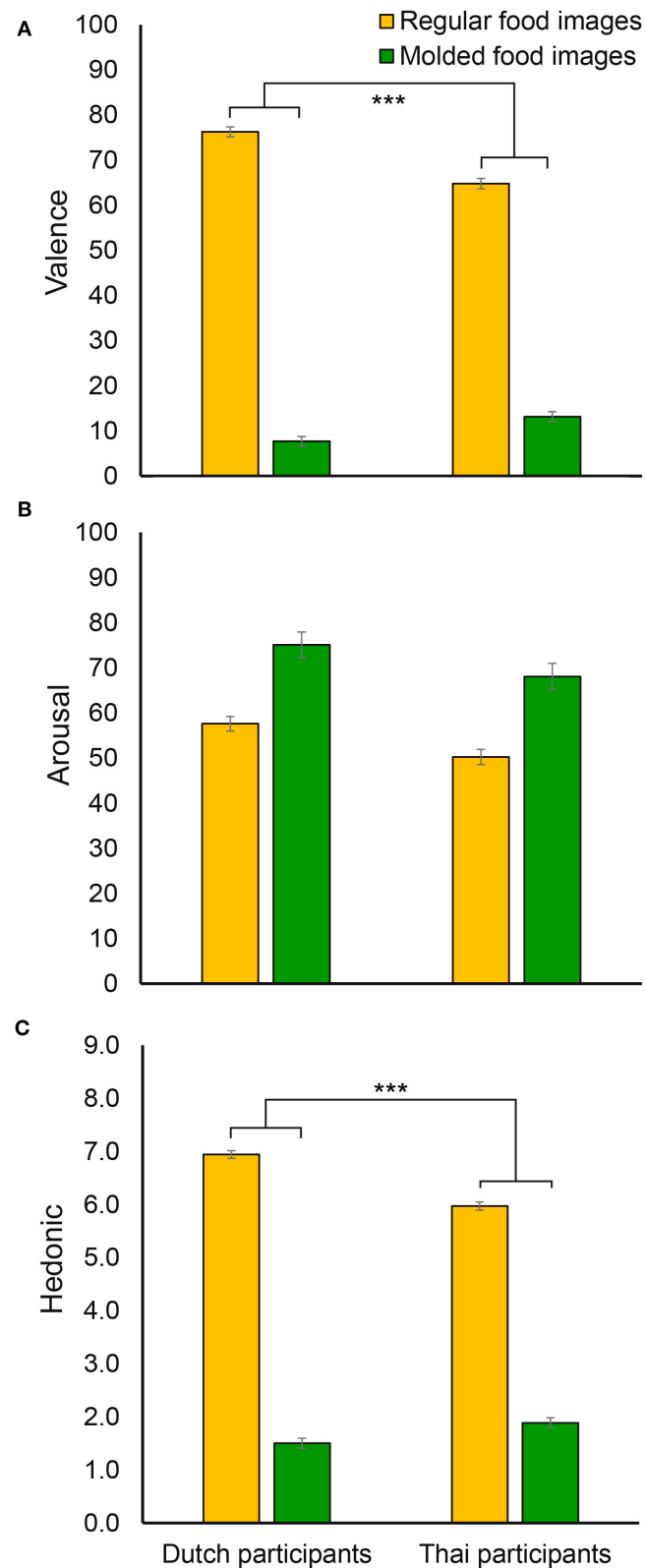
FIGURE 5 | The averaged valence **(A)**, arousal **(B)**, and hedonic **(C)** scores for each participant nationality (Dutch and Thai) and universal food image category (regular and molded). Error bars indicate standard error of the mean. *** indicates a significant interaction between participant nationality and universal food image category with $p < 0.001$.

FIGURE 6 | The averaged valence **(A)**, arousal **(B)**, and hedonic **(C)** scores for each participant nationality (Dutch and Thai) and cultural food image category (Dutch and Thai). Error bars indicate standard error of the mean. ** and *** indicate significant interactions between participant nationality and cultural food image category with $p < 0.01$ and $p < 0.001$, respectively.
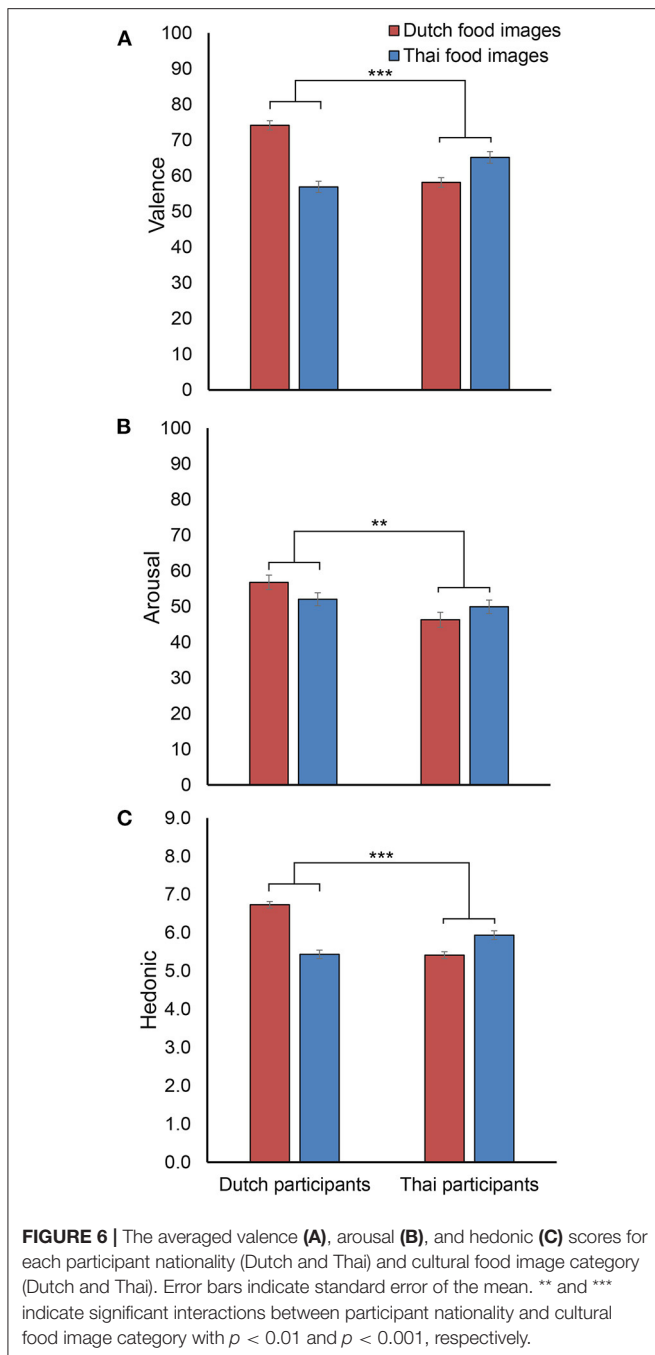
food images, while Thai participants showed the opposite pattern. A significant main effect of participant nationality [$F_{(1, 85)} = 7.06$, $p = 0.009$, $\eta_p^2 = 0.077$] was found with overall higher arousal scores for Dutch participants. There was no significant main effect of cultural food image category [$F_{(1, 85)} = 0.15$, $p = 0.704$].

The mean hedonic liking scores are shown in **Figure 6C**. It shows the same pattern as valence scores, with a significant interaction between participant nationality and cultural food image category [$F_{(1, 85)} = 119.71$, $p <0.001$, $\eta_p^2 = 0.585$],

and significant main effects of both participant nationality [$F_{(1, 85)} = 11.98$, $p = 0.001$, $\eta_p^2 = 0.124$] and cultural food image category [$F_{(1, 85)} = 21.75$, $p <0.001$, $\eta_p^2 = 0.204$]).

Implicit Measures for Universal and Cultural Food Categories (Hypothesis 3 and 4)

Universal Food Category

Figure 7A shows the averaged HR and the HR traces for each participant nationality and universal food image category. As expected, there was a significant main effect of universal food image category on HR [$F_{(1, 82)} = 116.42$, $p < 0.001$, $\eta_p^2 = 0.587$], where we observed a lower HR for molded compared to regular images. Also as expected, this effect was the same for participants of both nationalities, indicated by a lack of interaction between participant nationality and universal food image category [$F_{(1, 82)} = 0.909$, $p = 0.343$]. There was a main effect of participant nationality, [$F_{(1, 82)} = 7.14$, $p = 0.009$, $\eta_p^2 = 0.080$] with Thai participants showing higher HR than Dutch participants. **Figure 7B** shows the averaged phasic EDA and the phasic EDA traces for each participant nationality and each universal food image category. The two-way mixed ANOVA on phasic EDA did not show main effects of universal food image category [$F_{(1, 82)} = 0.166$, $p = 0.684$] or participant nationality [$F_{(1, 82)} = 2.856$, $p = 0.095$], and no interaction [$F_{(1, 82)} = 0.001$, $p = 0.977$].
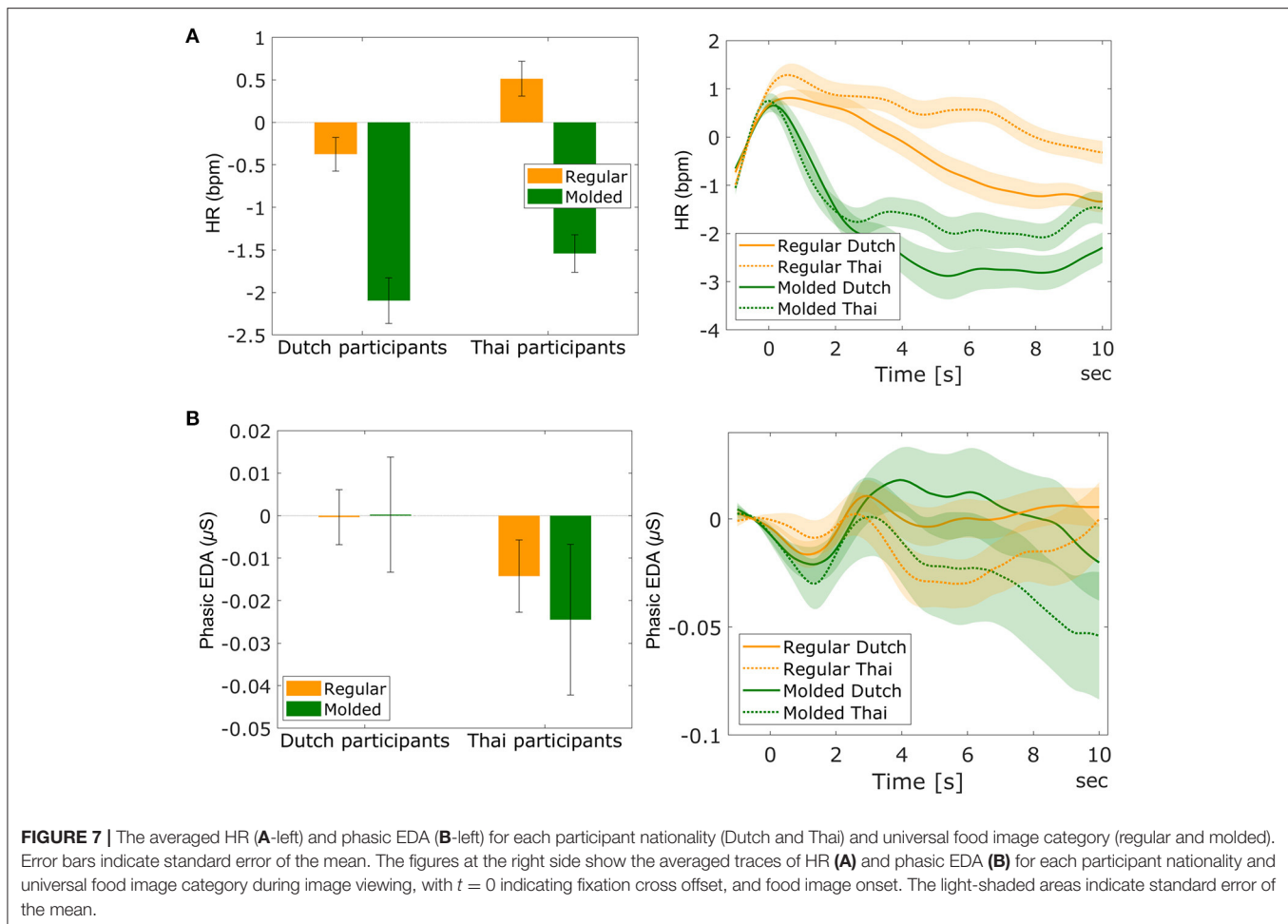
Cultural Food Category

Figure 8A shows the averaged HR for each participant nationality and each cultural food image category. As expected, a significant interaction between participant nationality and cultural food image category was found on HR [$F_{(1, 82)} = 8.50$, $p = 0.005$, $\eta_p^2 = 0.094$], with Dutch participants showing lower heart rate for Thai food images compared to Dutch food images, while the pattern was opposite for Thai participants. There was no significant main effect of cultural food image category [$F_{(1, 82)} = 2.93$, $p = 0.091$], but there was a significant main effect of participant nationality [$F_{(1, 82)} = 10.05$, $p = 0.002$, $\eta_p^2 = 0.109$] with Thai participants showing higher HR than Dutch participants. **Figure 8B** shows the averaged phasic EDA and the phasic EDA traces for each participant nationality and each cultural food image category. For phasic EDA, no interaction between nationality and cultural food image category was found [$F_{(1, 82)} = 3.02$, $p = 0.086$], and no significant main effect of cultural food image category [$F_{(1, 82)} = 2.74$, $p = 0.101$]. There was a significant main effect of participant nationality [$F_{(1, 82)} = 4.28$, $p = 0.042$, $\eta_p^2 = 0.050$] with Dutch participants showing higher EDA than Thai participants.

Explicit and Implicit Measures for Cultural Drinks (Hypothesis 5)

Explicit Measures

Figures 9A–C show respectively the average valence, arousal, and hedonic liking ratings for each participant nationality and each cultural drink. The two-way mixed ANOVAs did not show any interaction or main effects of cultural drink category and

FIGURE 7 | The averaged HR (**A**-left) and phasic EDA (**B**-left) for each participant nationality (Dutch and Thai) and universal food image category (regular and molded). Error bars indicate standard error of the mean. The figures at the right side show the averaged traces of HR (**A**) and phasic EDA (**B**) for each participant nationality and universal food image category during image viewing, with $t = 0$ indicating fixation cross offset, and food image onset. The light-shaded areas indicate standard error of the mean.

participant nationality on any of the three measures (all p-values are 0.094 or higher).

Implicit Measures

The average HR and phasic EDA for each participant nationality and each cultural drink are shown in **Figures 10A,B**. The two-way mixed ANOVA showed the hypothesized interaction between participant nationality and cultural drink category on HR [$F_{(1, 82)} = 5.79$, $p = 0.018$, $\eta_p^2 = 0.066$]. While both nationalities tend to show a higher HR for the Dutch drink, this is especially the case for Thai participants. Participant nationality [$F_{(1, 82)} = 5.72$, $p = 0.019$, $\eta_p^2 = 0.065$] and cultural drink category [$F_{(1, 82)} = 10.46$, $p = 0.002$, $\eta_p^2 = 0.113$] showed main effects on HR, with overall higher HR for the Dutch drink than the Thai drink, and higher HR for Thai participants than Dutch participants. For phasic EDA, there was no interaction between participant nationality and cultural drink category [$F_{(1, 82)} = 1.07$, $p = 0.303$], and no significant main effect of cultural drink category [$F_{(1, 82)} = 2.96$, $p = 0.089$]. Dutch participants showed overall higher phasic EDA than Thai participants (main effect of participant nationality: [$F_{(1, 82)} = 5.62$, $p = 0.020$, $\eta_p^2 = 0.064$].

Figure 11 shows the averaged behavioral measure of sip size. The expected interaction between participant nationality and cultural drink category was found [$F_{(1, 85)} = 15.24$, $p < 0.001$, $\eta_p^2 = 0.152$], indicating that Dutch participants took a larger sip of the Dutch drink than the Thai drink, while this was opposite for Thai participants. A significant main effect of participant nationality was found [$F_{(1, 85)} = 6.97$, $p = 0.010$, $\eta_p^2 = 0.076$] indicating that Thai participants took larger sips in general. There was no main effect of cultural drink category [$F_{(1, 85)} = 0.82$, $p = 0.367$].

DISCUSSION

The present study investigated the potential of implicit physiological measures to provide objective measures of affective food experience in contrast to explicit self-report measures, for Dutch and Thai participants. Explicit self-reports of these participants are expected to be influenced by differential cultural bias, therewith hampering comparison of food experience across cultures. Implicit physiological measures could potentially solve this problem.
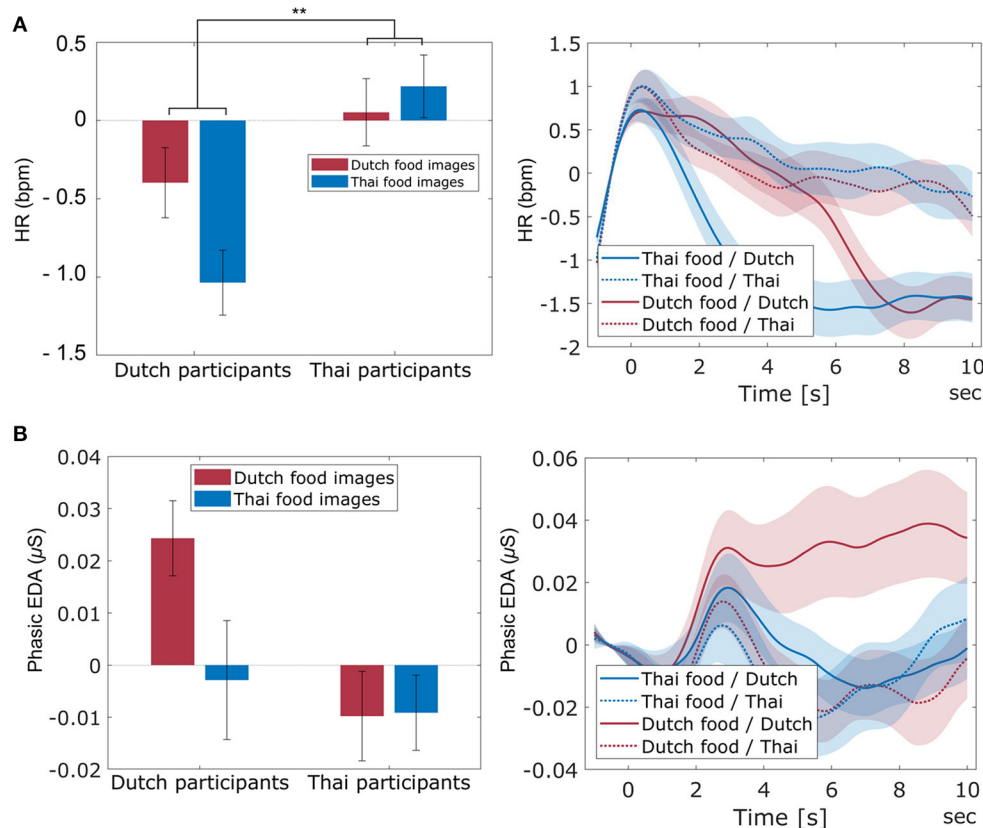
FIGURE 8 | The averaged HR (**A**-left) and phasic EDA (**B**-left) across participant nationality (Dutch and Thai) and cultural food image category (Dutch and Thai). Error bars indicate standard error of the mean. The figures at the right side indicate the averaged trace of HR **(A)** and phasic EDA **(B)** across participant nationality and cultural food image category during image viewing, with $t = 0$ indicating fixation cross offset, and food image onset. The light-shaded areas indicate standard error of the mean. ** indicates a significant interaction between participant nationality and cultural food image category with $p < 0.01$.
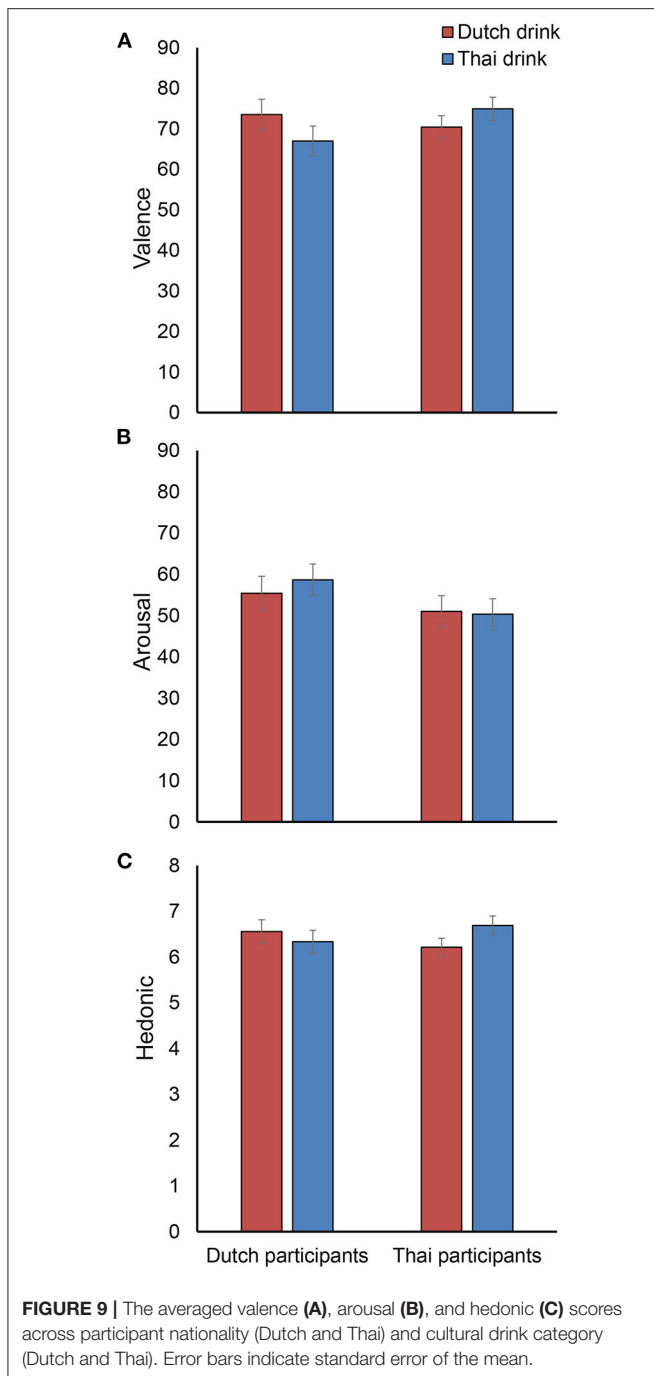
Explicit self-report responses toward universal food image categories (regular and molded food images) revealed the usage of an extreme response style by Dutch participants compared to Thai participants who used a middle response style, which is consistent with the literature on response style characteristics of cultural groups (Chen et al., 1995; Johnson et al., 2005; Kaneko et al., 2019b), and confirming hypothesis 1. The valence and liking scores were higher at both ends of the scale for Dutch participants than for Thai participants. While there was no such extreme response style at both ends of the scale for arousal (where in contrast to valence, no obvious neutral location exists), Dutch participants rated universal food images more extremely on high arousal compared to Thai participants.

Familiarity ratings of the cultural food images confirmed that for Thai participants, Thai food images were more familiar than Dutch food images, while the reverse was true for Dutch participants. This confirmed that our food image stimuli were properly selected. As hypothesized in hypothesis 2, significant interactions between participant nationality and cultural food image category for self-reported valence, arousal, and hedonic liking revealed that participants rated food images from their own

cultural food images as more pleasant and arousing than from the other culture. Higher liking scores for familiar foods is consistent with previous studies (Torrico et al., 2019).

As stated in hypothesis 3, HR responses were sensitive to affective food experience, as indicated by an effect of universal image type (lower HR for molded compared to regular images). The direction of this effect is consistent with literature on heart rate responses to high and low valence images (Bradley et al., 1990; Lang et al., 1993). As also stated in hypothesis 3, this effect was the same for both cultural groups, suggesting that despite the culturally-dependent difference in explicit ratings, affective food experience of universal food images (regular and molded) is the same across cultures. For cultural food images, we found the expected interaction effect between participant nationality and cultural food category on HR (hypothesis 4). Dutch participants had a lower HR in response to Thai food images compared to Dutch food images and *vice versa* for Thai participants. The direction of the effect is as expected, consistent with lower valence for unfamiliar food.
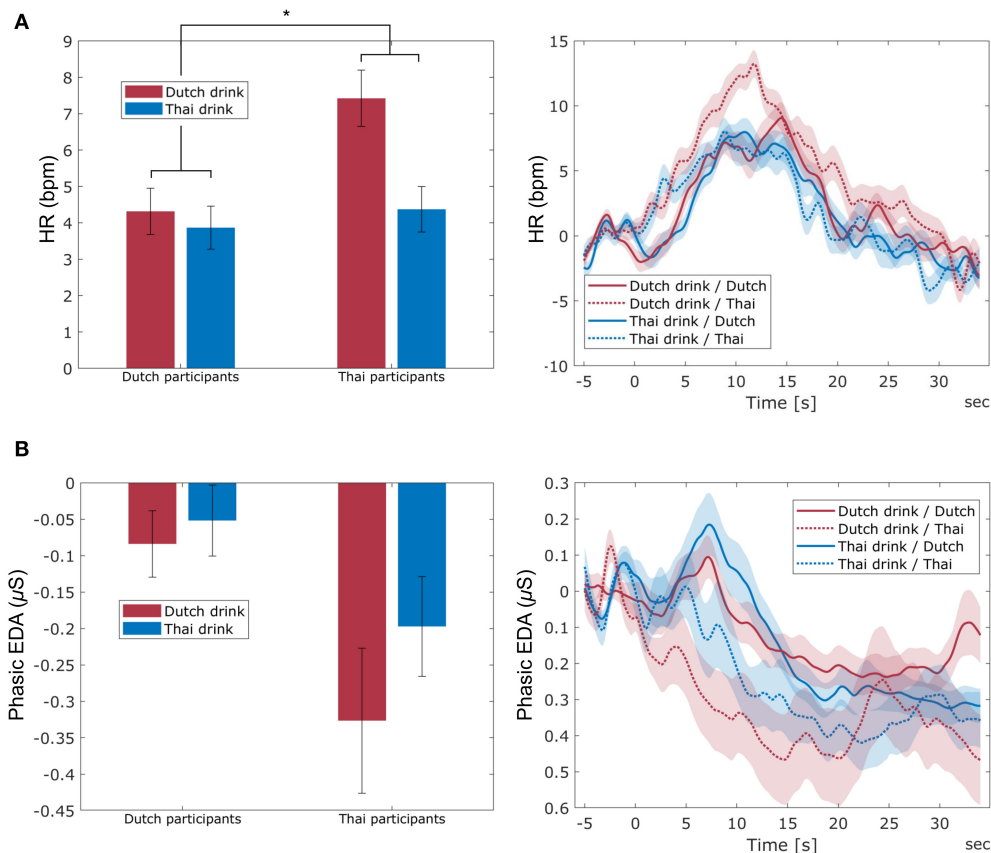
For cultural drinks, we found an interaction between participant nationality and cultural drink category on the

FIGURE 9 | The averaged valence **(A)**, arousal **(B)**, and hedonic **(C)** scores across participant nationality (Dutch and Thai) and cultural drink category (Dutch and Thai). Error bars indicate standard error of the mean.

implicit behavioral measure of sip size in the expected direction (hypothesis 5), indicating that participants of both nationalities take a larger sip of the drink from their own culture. HR also showed an interaction between participant nationality and cultural food category, where Thai participants showed a particularly strong response to the Dutch drink. These results corroborate previous results on sip size and HR (and other physiological measures) as sources of information on affective experience of taking a sip of a drink (Kaneko et al., 2019a),

where sip size was smaller and HR was higher for (low valence and) high arousal drinks. In the present study, visual, olfactory and taste properties of the drinks were held constant and only the associated emotion differed. In contrast to the sip size and HR results, we unexpectedly did not find an interaction effect between participant nationality and cultural food category in explicit self-reported affective scores. Also, while familiarity scores showed the intended interaction between participant nationality and cultural food category, and Dutch participants rated the Dutch drink as much more familiar than the Thai drink, Thai participants rated the two drinks almost equal in familiarity, which is in apparent conflict with their smaller sip size and higher HR response. Note that in contrast to the implicit and explicit responses to food images, in the case of taking a sip, participants' sensory perception fundamentally changes over time. It starts with an expectation primed by the announcement of the drink name, is followed by perceiving its visual and olfactory properties, and ends with the actual taste. We speculate that the seemingly conflicting results between implicit and explicit measures are caused by their difference in time. For measures that reflect the time of taking a sip and shortly thereafter (sip size and HR), we observe a pattern in line with the hypothesis. When rating the drinks, which happens about half a minute later, Thai participants may realize they are actually familiar with the yogurt drink, and participants of both nationalities may like the taste of both drinks.

In contrast to HR, phasic EDA was not a sensitive measure of food evoked emotion in the present study, as indicated by our finding that phasic EDA responses did not differ between regular and molded food images. In addition, no interaction between participant nationality and cultural food category was found for phasic EDA; neither for images nor for drinks. While in general, there is no straightforward relation between HR and valence, studies using emotional images as stimuli consistently show valence (rather than arousal) effects, where pleasant stimuli correlate with higher heart rate acceleration than unpleasant stimuli (Hare et al., 1970; Libby Jr et al., 1973; Winton et al., 1984; Greenwald et al., 1989; Bradley et al., 1990; Lang et al., 1993, 1998; Bradley and Lang, 2000; Anttonen and Surakka, 2005; Codispoti and De Cesarei, 2007; Sokhadze, 2007). We found this HR effect as well, both in universal and cultural food categories. For EDA, consistent positive associations with arousal have been found in a range of contexts, including emotional/neutral picture viewing tasks (Greenwald et al., 1989; Lang et al., 1993, 1998) and tasting drinks (Kaneko et al., 2019a; Brouwer et al., 2020). In the present study, phasic EDA showed no (interaction with) food category effects. We did not have a clear a priori expectation that arousal should differ between cultural food image categories (familiar and unfamiliar), but we had expected higher arousal for molded food images compared to regular food images. This expectation was supported by the ratings, but the difference in rated arousal between regular and molded food images was modest compared to the difference in valence. Thus, the lack of effect in EDA may have been caused by relatively small genuine difference in arousal between our stimulus categories; they differed more in valence which, at least for images, is better captured by HR. Consistent with our findings is a study by Anderson et al.
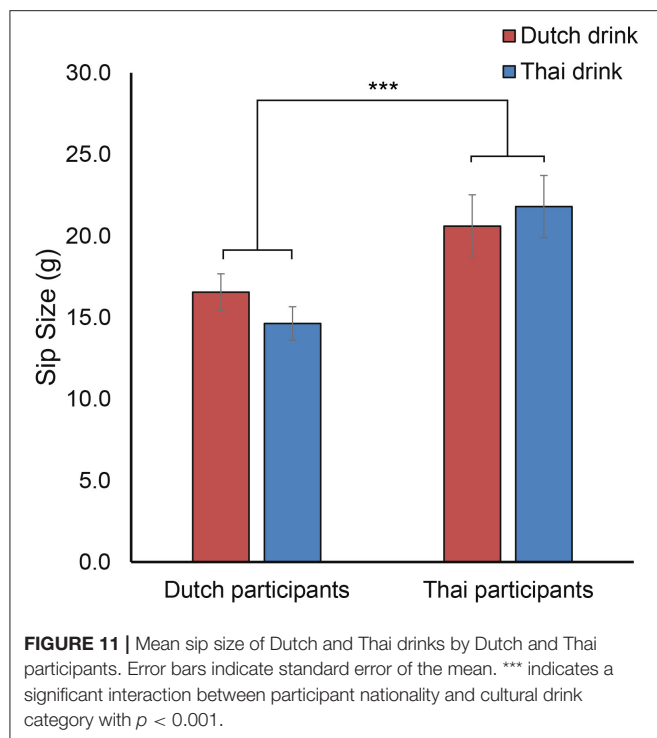
FIGURE 10 | The averaged HR (**A**-left) and phasic EDA (**B**-left) across participant nationality (Dutch and Thai) and cultural drink category (Dutch and Thai). Error bars indicate standard error of the mean. The figures at the right side indicate the averaged trace of HR (**A**) and phasic EDA (**B**) across participant nationality and cultural drink category. The name of the drink is presented from $t = -5$ to $t = 0$, after which participants took one sip. The light-shaded areas indicate standard error of the mean. * indicates a significant interaction between participant nationality and cultural drink category with $p < 0.05$.

(2019). This study showed similar skin conductance levels when viewing both rotten and sweet food images, while HR was decreased when viewing molded food images compared to sweet food images.

Our analyses showed trends indicating overall higher HR and lower phasic EDA for Thai participants than for Dutch participants. Note that our HR and phasic EDA variables reflect baselined responses. We examined whether these differences in responses to food stimuli might be related to differences in overall HR and phasic EDA, which e.g., could have been caused by a difference in climate between the Netherlands and Thailand (Madaniyazi et al., 2016). However, we did not find statistically significant differences between nationalities for unbaselined HR and EDA levels.

While implicit physiological measures have been praised as objective markers of affective experience, we do not know of studies that clearly show or suggest that explicit self-reported ratings are less accurate markers of food experience than physiological measures. With our study, we attempted to fill this gap, and our results suggest that implicit physiological measures could indeed be considered as better indicators of

food experience than explicit ratings. However, we should consider that we do not (cannot) know the absolute "true" emotion. Our findings could still be consistent with a genuine difference between Thai and Dutch participants with respect to the experience of viewing regular and molded food images, following the explicit ratings which are not "really" biased, but reflecting genuine differences. In this interpretation, HR would not be sensitive enough to capture this, even though HR could capture other effects that may have been stronger. Replication of our results with other types of ground truth stimuli, and relating emotions to behavior (e.g., food choice), would therefore be good to further establish the findings of the present study. For future studies, it would also be of interest to investigate if and how possible cross-cultural differences in food neophobia influence different measures of affective food experience. Finally, for adding another dimension of food experience, it would be valuable to include implicit measures of approach-avoidance motivation such as EEG alpha asymmetry (Harmon-Jones et al., 2010; Brouwer et al., 2017) or a behavioral approach-avoidance task (Solarz, 1960; Piqueras-Fiszman et al., 2014; Van Beers et al., 2020).

FIGURE 11 | Mean sip size of Dutch and Thai drinks by Dutch and Thai participants. Error bars indicate standard error of the mean. *** indicates a significant interaction between participant nationality and cultural drink category with $p < 0.001$.

CONCLUSION

In the present study, we examined implicit measures of food experience in a case that self-reported ratings cannot be taken at face value because of possible culturally dependent response bias. Our study confirmed the existence of such a cultural response bias and showed that only the implicit physiological measure of HR followed the prior expectations of genuine food experience. Different contexts, in this case, different types of food stimuli (images and drinks) resulted in different sensory, affective and dynamics of physiological responses. For both types of stimuli, we found the expected interaction between participant nationality and cultural food category on implicit measures (HR, and in case of tasting, sip size). This study indicates that physiological responses can be used to investigate differences in affective food experience between cultures. Especially when estimating possible acceptance of products and product promotion in cultures that strongly differ with respect

to expressing affect, such as Asian and western cultures, using self-reports alone may lead to incorrect conclusions.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by TNO Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Anderson, E. C., Wormwood, J., Barrett, L. F., and Quigley, K. S. (2019). Vegetarians' and omnivores' affective and physiological responses to images of food. *Food Qual. Prefer.* 71, 96–105. doi: 10.1016/j.foodqual.2018.06.008

Anttonen, J., and Surakka, V. (2005). "Emotions and heart rate while sitting on a chair," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, OR), 491–499.

Ares, G. (2018). Methodological issues in cross-cultural sensory and consumer research. *Food Qual. Prefer.* 64, 253–263. doi: 10.1016/j.foodqual.2016.10.007

Bach, A. J., Stewart, I. B., Minett, G. M., and Costello, J. T. (2015). Does the technique employed for skin temperature assessment alter outcomes? A systematic review. *Physiol. Meas.* 36:R27. doi: 10.1088/0967-3334/36/9/R27

Benedek, M., and Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *J. Neurosci. Methods* 190, 80–91. doi: 10.1016/j.jneumeth.2010.04.028

Bradley, M. M., Cuthbert, B. N., and Lang, P. J. (1990). Startle reflex modification: emotion or attention? *Psychophysiology* 27, 513–522. doi: 10.1111/j.1469-8986.1990.tb01966.x

Bradley, M. M., and Lang, P. J. (2000). Affective reactions to acoustic stimuli. *Psychophysiology* 37, 204–215. doi: 10.1111/1469-8986.3720204

Brouwer, A., Hogervorst, M., Grootjen, M., Van Erp, J., and Zandstra, E. (2017). Neurophysiological responses during cooking food associated with different emotions. *Food Qual. Prefer.* 62, 307–316. doi: 10.1016/j.foodqual.2017.03.005

Brouwer, A.-M., Van Den Broek, T., Hogervorst, M., Kaneko, D., Toet, A., Kallen, V., et al. (2020). Estimating affective taste experience using combined implicit behavioral and neurophysiological measures. *IEEE Trans. Affect. Comput.* doi: 10.1109/TAFFC.2020.3032236. [Epub ahead of print].

Charbonnier, L., Van Meer, F., Van Der Laan, L. N., Viergever, M. A., and Smeets, P. a. M. (2016). Standardized food images: a photographing protocol and image database. *Appetite* 96, 166–173. doi: 10.1016/j.appet.2015.08.041

Chen, C., Lee, S.-Y., and Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychol. Sci.* 6, 170–175. doi: 10.1111/j.1467-9280.1995.tb00327.x

Codispoti, M., and De Cesarei, A. (2007). Arousal and attention: picture size and emotional reactions. *Psychophysiology* 44, 680–686. doi: 10.1111/j.1469-8986.2007.00545.x

Dalenberg, J. R., Gutjar, S., Ter Horst, G. J., De Graaf, K., Renken, R. J., and Jager, G. (2014). Evoked emotions predict food choice. *PLoS ONE* 9:e115388. doi: 10.1371/journal.pone.0115388

Danner, L., Haindl, S., Joechl, M., and Duerrschmid, K. (2014). Facial expressions and autonomous nervous system responses elicited by tasting different juices. *Food Res. Int.* 64, 81–90. doi: 10.1016/j.foodres.2014.06.003

de Wijk, R. A., Kooijman, V., Verhoeven, R. H. G., Holthuysen, N. T. E., and De Graaf, C. (2012). Autonomic nervous system responses on and facial expressions to the sight, smell, and taste of liked and disliked foods. *Food Qual. Prefer.* 26, 196–203. doi: 10.1016/j.foodqual.2012.04.015

Dorado, R., Chaya, C., Tarrega, A., and Hort, J. (2016). The impact of using a written scenario when measuring emotional response to beer. *Food Qual. Prefer.* 50, 38–47. doi: 10.1016/j.foodqual.2016.01.004

Greenwald, A. G., Klinger, M. R., and Liu, T. J. (1989). Unconscious processing of dichoptically masked words. *Memory Cogn.* 17, 35–47. doi: 10.3758/BF03199555

Gutjar, S., De Graaf, C., Kooijman, V., De Wijk, R. A., Nys, A., Ter Horst, G. J., et al. (2015). The role of emotions in food choice and liking. *Food Res. Int.* 76, 216–223. doi: 10.1016/j.foodres.2014.12.022

Hare, R., Wood, K., Britain, S., and Shadman, J. (1970). Autonomic responses to affective visual stimulation. *Psychophysiology* 7, 408–417. doi: 10.1111/j.1469-8986.1970.tb01766.x

Harmon-Jones, E., Gable, P. A., and Peterson, C. K. (2010). The role of asymmetric frontal cortical activity in emotion-related phenomena: a review and update. *Biol. Psychol.* 84, 451–462. doi: 10.1016/j.biopsycho.2009.08.010

Harzing, A.-W. (2006). Response styles in cross-national survey research: a 26-country study. *Int. J. Cross Cult. Manage.* 6, 243–266. doi: 10.1177/1470595806066332

Hassan, M. A., Malik, A. S., Fofi, D., Saad, N., Karasfi, B., Ali, Y. S., et al. (2017). Heart rate estimation using facial video: a review. *Biomed. Signal Process. Control* 38, 346–360. doi: 10.1016/j.bspc.2017.07.004

Johnson, T., Kulesa, P., Cho, Y. I., and Shavitt, S. (2005). The relation between culture and response styles: evidence from 19 countries. *J. Cross Cult. Psychol.* 36, 264–277. doi: 10.1177/0022022104272905

Kaneko, D., Hogervorst, M., Toet, A., Van Erp, J. B., Kallen, V., and Brouwer, A.-M. (2019a). Explicit and implicit responses to tasting drinks associated with different tasting experiences. *Sensors* 19:4397. doi: 10.3390/s19204397

Kaneko, D., Toet, A., Brouwer, A.-M., Kallen, V., and Van Erp, J. B. (2018). Methods for evaluating emotions evoked by food experiences: a literature review. *Front. Psychol.* 9:911. doi: 10.3389/fpsyg.2018.00911

Kaneko, D., Toet, A., Ushiama, S., Brouwer, A., Kallen, V., and Van Erp, J. (2019b). "EmojiGrid: a 2D pictorial scale for cross-cultural emotion assessment of negatively and positivey valenced food: poster en abstract," in *13th Pangborn Sensory Science Symposium: Engage with the Future* (Edinburgh). doi: 10.1016/j.foodres.2018.09.049

Köster, E. P., and Mojet, J. (2015). From mood to food and from food to mood: a psychological perspective on the measurement of food-related emotions in consumer research. *Food Res. Int.* 76, 180–191. doi: 10.1016/j.foodres.2015.04.006

Kranjec, J., Beguš, S., Geršak, G., and Drnovšek, J. (2014). Non-contact heart rate and heart rate variability measurements: a review. *Biomed. Signal Process. Control* 13, 102–112. doi: 10.1016/j.bspc.2014.03.004

Lagast, S., Gellynck, X., Schouteten, J., De Herdt, V., and De Steur, H. (2017). Consumers' emotions elicited by food: a systematic review of explicit and implicit methods. *Trends Food Sci. Technol.* 69, 172–189. doi: 10.1016/j.tifs.2017.09.006

Lang, P. J., Bradley, M. M., Fitzsimmons, J. R., Cuthbert, B. N., Scott, J. D., Moulder, B., et al. (1998). Emotional arousal and activation of the visual cortex: an fMRI analysis. *Psychophysiology* 35, 199–210. doi: 10.1111/1469-8986.3520199

Lang, P. J., Greenwald, M. K., Bradley, M. M., and Hamm, A. O. (1993). Looking at pictures: affective, facial, visceral, and behavioral reactions. *Psychophysiology* 30, 261–273. doi: 10.1111/j.1469-8986.1993.tb03352.x

Lawless, H. T., and Heymann, H. (2010). *Sensory Evaluation of Food: Principles and Practices*. New York, NY: Springer Science and Business Media. doi: 10.1007/978-1-4419-6488-5_3

Libby Jr., W. L., Lacey, B. C., and Lacey, J. I. (1973). Pupillary and cardiac activity during visual attention. *Psychophysiology* 10, 270–294. doi: 10.1111/j.1469-8986.1973.tb00526.x

Lim, J. (2011). Hedonic scaling: a review of methods and theory. *Food Qual. Prefer.* 22, 733–747. doi: 10.1016/j.foodqual.2011.05.008

Madaniyazi, L., Zhou, Y., Li, S., Williams, G., Jaakkola, J. J., Liang, X., et al. (2016). Outdoor temperature, heart rate and blood pressure in Chinese adults: effect modification by individual characteristics. *Sci. Rep.* 6, 1–9. doi: 10.1038/srep21003

Meiselman, H. L. (2013). The future in sensory/consumer research: evolving to a better science. *Food Qual. Prefer.* 27, 208–214. doi: 10.1016/j.foodqual.2012.03.002

Meiselman, H. L. (2015). A review of the current state of emotion research in product development. *Food Res. Int.* 76, 192–199. doi: 10.1016/j.foodres.2015.04.015

Pan, J., and Tompkins, W. J. (1985). A real-time QRS detection algorithm. *IEEE Trans. Biomed. Eng.* 32, 230–236. doi: 10.1109/TBME.1985.325532

Piqueras-Fiszman, B., Kraus, A. A., and Spence, C. (2014). "Yummy" versus "Yucky"! Explicit and implicit approach–avoidance motivations towards appealing and disgusting foods. *Appetite* 78, 193–202. doi: 10.1016/j.appet.2014.03.029

Rozin, P. (1988). "Cultural approaches to human food preferences," in *Nutritional Modulation of Neural Function*, eds J. E. Morley, M. B. Sterman, and J. H. Walsh (Waltham, MA: Academic Press), 137–153. doi: 10.1016/B978-0-12-506455-2.50017-4

Rozin, P. (2006). The integration of biological, social, cultural and psychological influences on food choice. *Front. Nutr. Sci.* 3:19. doi: 10.1079/9780851990323.0019

Samant, S. S., Chapko, M. J., and Seo, H.-S. (2017). Predicting consumer liking and preference based on emotional responses and sensory perception: a study with basic taste solutions. *Food Res. Int.* 100, 325–334. doi: 10.1016/j.foodres.2017.07.021

Sedghamiz, H. (2014). *Matlab Implementation of Pan Tompkins ECG QRS Detector*. Code Available at the File Exchange Site of MathWorks. Available online at: https://de.mathworks.com/matlabcentral/fileexchange/45840-complete-pan-tompkins-implementation-ecg-qrs-detector

Silva, A. P., Jager, G., Van Bommel, R., Van Zyl, H., Voss, H.-P., Hogg, T., et al. (2016). Functional or emotional? How Dutch and Portuguese conceptualise beer, wine and non-alcoholic beer consumption. *Food Qual. Prefer.* 49, 54–65. doi: 10.1016/j.foodqual.2015.11.007

Sokhadze, E. M. (2007). Effects of music on the recovery of autonomic and electrocortical activity after stress induced by aversive visual stimuli. *Appl. Psychophysiol. Biofeedback* 32, 31–50. doi: 10.1007/s10484-007-9033-y

Solarz, A. K. (1960). Latency of instrumental responses as a function of compatibility with the meaning of eliciting verbal signs. *J. Exp. Psychol.* 59:239. doi: 10.1037/h0047274

Soto, J. A., Lee, E. A., and Roberts, N. A. (2016). Convergence in feeling, divergence in physiology: how culture influences the consequences of disgust suppression and amplification among European Americans and Asian Americans. *Psychophysiology* 53, 41–51. doi: 10.1111/psyp.12579

Soto, J. A., Levenson, R. W., and Ebling, R. (2005). Cultures of moderation and expression: emotional experience, behavior, and physiology in Chinese Americans and Mexican Americans. *Emotion* 5:154. doi: 10.1037/1528-3542.5.2.154

Toet, A., Kaneko, D., Kruijf, I. D., Ushiama, S., Schaik, M. V., Brouwer, A.-M., et al. (2019). CROCUFID: a cross-cultural food image database for research on food elicited affective responses. *Front. Psychol.* 10:58. doi: 10.3389/fpsyg.2019.00058

Toet, A., Kaneko, D., Ushiama, S., Hoving, S., De Kruijf, I., Brouwer, A.-M., et al. (2018). EmojiGrid: a 2D pictorial scale for the assessment of food elicited emotions. *Front. Psychol.* 9:2396. doi: 10.3389/fpsyg.2018.02396

Torrico, D. D., Fuentes, S., Viejo, C. G., Ashman, H., and Dunshea, F. R. (2019). Cross-cultural effects of food product familiarity on sensory acceptability and non-invasive physiological responses of consumers. *Food Res. Int.* 115, 439–450. doi: 10.1016/j.foodres.2018.10.054

Torrico, D. D., Fuentes, S., Viejo, C. G., Ashman, H., Gunaratne, N. M., Gunaratne, T. M., et al. (2018). Images and chocolate stimuli affect physiological and affective responses of consumers: a cross-cultural study. *Food Qual. Prefer.* 65, 60–71. doi: 10.1016/j.foodqual.2017.11.010

Tsai, J. L., Chentsova-Dutton, Y., Freire-Bebeau, L., and Przymus, D. E. (2002). Emotional expression and physiology in European Americans and Hmong Americans. *Emotion* 2:380. doi: 10.1037/1528-3542.2.4.380

Tsai, J. L., Levenson, R. W., and Carstensen, L. L. (2000). Autonomic, subjective, and expressive responses to emotional films in older and younger Chinese Americans and European Americans. *Psychol. Aging* 15:684. doi: 10.1037/0882-7974.15.4.684

Tuorila, H., Lähteenmäki, L., Pohjalainen, L., and Lotti, L. (2001). Food neophobia among the Finns and related responses to familiar and unfamiliar foods. *Food Qual. Prefer.* 12, 29–37. doi: 10.1016/S0950-3293(00)00025-2

Uchida, Y., and Kitayama, S. (2009). Happiness and unhappiness in east and west: themes and variations. *Emotion* 9:441. doi: 10.1037/a0015634

Van Beers, J. J., Kaneko, D., Stuldreher, I. V., Zech, H. G., and Brouwer, A.-M. (2020). "An accessible tool to measure implicit approach-avoidance tendencies towards food outside the lab," in *Companion Publication of the 2020 International Conference on Multimodal Interaction* (New York, NY), 307–311. doi: 10.1145/3395035.3425647

Van Zyl, H., and Meiselman, H. L. (2015). The roles of culture and language in designing emotion lists: comparing the same language in different English and Spanish speaking countries. *Food Qual. Prefer.* 41, 201–213. doi: 10.1016/j.foodqual.2014.12.003

Van Zyl, H., and Meiselman, H. L. (2016). An update on the roles of culture and language in designing emotion lists: English, Spanish and Portuguese. *Food Qual. Prefer.* 51, 72–76. doi: 10.1016/j.foodqual.2016.02.019

Winton, W. M., Putnam, L. E., and Krauss, R. M. (1984). Facial and autonomic manifestations of the dimensional structure of emotion. *J. Exp. Soc. Psychol.* 20, 195–216. doi: 10.1016/0022-1031(84)90047-7

World Medical Association (2013). World Medical Association declaration of Helsinki: ethical principles for medical research involving human subjects. *J. Am. Med. Assoc.* 310, 2191–2194. doi: 10.1001/jama.2013.281053

"Psyosphere": A GPS Data-Analysing Tool for the Behavioural Sciences

Benjamin Ziepert, Peter W. de Vries* and Elze Ufkes

Department of Psychology of Conflict, Risk, and Safety, University of Twente, Enschede, Netherlands

Positioning technologies, such as GPS are widespread in society but are used only sparingly in behavioural science research, e.g., because processing positioning technology data can be cumbersome. The current work attempts to unlock positioning technology potential for behavioural science studies by developing and testing a research tool to analyse GPS tracks. This tool—psyosphere—is published as open-source software, and aims to extract behaviours from GPSs data that are more germane to behavioural research. Two field experiments were conducted to test application of the research tool. During these experiments, participants played a smuggling game, thereby either smuggling tokens representing illicit material past border guards or not. Results suggested that participants varied widely in variables, such as course and speed variability and distance from team members in response to the presence of border guards. Subsequent analyses showed that some of these GPS-derived behavioural variables could be linked to self-reported mental states, such as fear. Although more work needs to be done, the current study demonstrates that psyosphere may enable researchers to conduct behavioural experiments with positioning technology, outside of a laboratory setting.

Keywords: GPS, positioning technologies, implicit measurement, spatial movement, walking, psychology

INTRODUCTION

Positioning technologies, such as Global Positioning Systems (GPS), Glonass, and Galileo can be used to determine the position on the globe and to record, for instance, the movement of planes, cars, and individuals (Hofmann-Wellenhof et al., 2007). Positioning technologies are now omnipresent in mobile devices, such as smart phones, tablets, and laptops, which makes them potentially interesting for the study of behaviour in naturalistic settings. It could, for instance, be used to identify people with early warnings signs for depression (Saeb et al., 2015; Palmius et al., 2017), partly or fully replace self-reported diaries in traffic research (Stopher et al., 2002; Wolf, 2006; Bohte and Maat, 2009; Schuessler and Axhausen, 2009), determine how populations behave after a disaster, such as an earthquake (Bengtsson et al., 2011), or to automatically detect active pickpockets in a shopping mall (Bouma et al., 2014). Surprisingly, behavioural scientists have so far used positioning technologies only sparingly.

Arguably, there are two reasons why positioning technologies have largely been neglected in behavioural research. First, the data are too complex to analyse with software that are traditionally used in the social sciences, such as IBM SPSS Statistics (SPSS). Second, only a limited number of studies has so far investigated the relationship between psychological variables and positioning

technology data. Consequently, little information is available about which psychological variables could be studied with positioning technologies.

Therefore, the aim of this study is to develop and test a tool that enables social scientists to conduct behavioural experiments with positioning technology. The tool will be tested with a "proof of principle" study, to illustrate that aspects of people's movements could be used to shed light on their states, and the study will explore possible relationships rather than testing specific hypotheses.

In current behavioural science research, the assessment of movement is often done via trained observers, interviewers, or self-reported diaries. Although these methods have been of great use (Goodchild and Janelle, 1984; Janelle et al., 1988; Doherty and Miller, 2000; Axhausen et al., 2002; Shoval et al., 2010), they do have their drawbacks. Shoval et al. (2010), for instance, point to the bias in participants' self-reported movements. For example, people frequently underreport small trips and trips that do not start or end at home. Moreover, participants driving a car tend to underestimate travel time, whereas public transportation users often overestimate it (Stopher, 1992; Ettema et al., 1996). Furthermore, participants may decide to omit information, such as when they view the information as socially undesirable. Finally, interviewers could fail to prompt recall (interviewer error), or the participants could simply forget the information over time (recall bias; Anderson, 1971; Golledge, 1997; Vazquez-Prokopec et al., 2009). These limitations can be circumvented by using positioning technologies, such as GPS (Bohte and Maat, 2009).

Especially in traffic research, positioning technologies, such as GPS have been argued to have several benefits over traditional methods of movement tracking (Stopher et al., 2002; Wolf et al., 2004; Bohte and Maat, 2009; Schuessler and Axhausen, 2009). Compared to self-reported diaries or interviews, for instance, GPS loggers are less effortful to work with, as they may substantially reduce information that otherwise needs to be self-reported by participants or needs to be asked by interviewers. As a result, GPS loggers may draw less heavily on interviewers and may therefore reduce the costs of conducting studies. In addition, they afford longer survey periods, as smartphone apps tracking movement in the background allow for longer data-collection periods compared to when the participants self-report their trips. Moreover, use of GPS may increase the quality of the data gathered, as travel times are recorded accurately regardless of the length of the data-collection period, whether they include one's home or not, and whether they are spent driving or walking. Finally, the sensors also have the benefit of recording additional data, such as speed and acceleration which can be used for additional analysis (Wolf et al., 2004).

Studies in other fields have also employed positioning technologies to replace or augment traditional methods of movement tracking, especially research with target groups that are unable to maintain a self-reported diary, such as the mentally impaired, children, and the elderly (Shoval et al., 2011). Measuring movement in such groups typically requires enlisting the help of caretakers or family members, registering their activities or filling in behavioural checklists in their stead (Shoval et al., 2011). This often turns out to be relatively expensive, burdensome and biased, even to the point where researchers may avoid conducting studies with these target groups altogether (Isaacson et al., 2016). In such contexts, replacing human observers with GPS loggers may prove useful (Shoval et al., 2008, 2010, 2011; Isaacson et al., 2016), and the required protocols can be followed by mentally impaired and elderly people (Isaacson et al., 2016).

Table 1 shows that whereas positioning technologies have been taken up by quite a few researchers, studies specifically investigating the link between positioning-technology data and psychological variables, especially mental states, are few and far between. This is surprising, given that several laboratory studies have already managed to establish such links with more conventional means. For instance, sad, depressed and frightened people tend to walk slower than others; additionally, joy and anger are linked to increased walking speed (Michalak et al., 2009; Barliya et al., 2012; Gross et al., 2012). Other research indicates that personality traits, such as agreeableness are also linked to increased walking speed (Satchell et al., 2017). Similarly, positive affect, extraversion, or openness to experiences have been associated with physical activity (Byrne and Byrne, 1993; Schwerdtfeger et al., 2010). Insofar GPS has been used in behavioural research, it has focused on constructing measures of daily activity (Wolf et al., 2013; Fillekes et al., 2019; also see Carlson et al., 2014; Jankowska et al., 2015; James et al., 2016), to detect risk-taking behaviour on the road, i.e., speeding (Bolderdijk et al., 2011), or to predict depression and social anxiety (Saeb et al., 2015; Huang et al., 2016; Palmius et al., 2017). However, links between location-based data and state variables have, so far, received little attention, if any at all.

Arguably, the apparent lack of interest among behavioural scientists, especially psychologists, can be explained by a limited awareness that positioning technologies may yield information that sheds light on psychological state variables. We therefore decided to develop and test a tool that extracts different aspects of behaviour from GPS logs, i.e., to enable behavioural scientists to analyse movement data without the need of additional special expertise. This tool was subsequently applied to field-experimental data to see whether these behavioural aspects can be linked to various state variables.

Out of the variety of positioning technologies we decided to focus on GPS for our tool. GPS can be used all over the globe and does not depend on local GSM, Wi-Fi, or other infrastructure. GPS is also omnipresent in smart phones or other devices, and dedicated GPS loggers are affordable. The data analysis software will work with longitude, latitude, and timestamp data points that are typical for GPS loggers. The movement data from other positioning technologies, such as GSM and Wi-Fi data can be converted to be used with the same software once it is converted to longitude and latitude.

The tool is an R package called "psyosphere" (Ziepert et al., 2018; see **Supplementary Appendix 1**). R is an open-source programming language and data-analysis tool that is becoming more widespread (Muenchen, 2012). The choice for R has several benefits: since R is used by psychologists and computer scientists it could improve cooperation of interdisciplinary teams, the software is free of charge and therefore more easily accessible

Measures	Research
Anxiety, depression, or lifestyle (e.g., positive affect or extraversion)	Determining relationship between active vs. sedentary lifestyle, social anxiety and depression, and number places visited with GPS (Wolf et al., 2013; Huang et al., 2016; Saeb et al., 2016).
Community specific routes description and visualisation	Measuring segregation in city communities with GPS (Davies et al., 2017; Whyatt et al., 2017).
Depression detection	Detecting depression from GPS movement data characteristics, such as location variance, home stay, or mobility between favourite locations (Saeb et al., 2015; Palmius et al., 2017).
Environmental exposure	Measuring daily environmental exposure with GPS (Phillips et al., 2001; Chaix et al., 2013).
Following and leadership detection	Detecting leadership and followership with movement patterns (e.g., co-moving) with Wi-Fi data (Kjargaard et al., 2013).
Information or disease spreading characteristics	Tracking the spreading of information in face-to-face networks with Bluetooth, RFID, and Wi-Fi (Madan et al., 2010; Isella et al., 2011a,b).
Physical activity and mobility	Measuring physical activity of children, the elderly, or other target groups with GPS (Elgethun et al., 2002; Fjørtoft et al., 2009; Maddison and Mhurchu, 2009; Krenn et al., 2011; Shoval et al., 2011; Carlson et al., 2014; Isaacson et al., 2015; Jankowska et al., 2015; James et al., 2016; Fillekes et al., 2019).
Pickpocket detection	Detecting pickpockets with movement characteristics (e.g., walking speed) measured with security cameras (Bouma et al., 2014).
Population movement characteristics	Identify population behaviour after a disaster with GSM (Bengtsson et al., 2011).
Risk seeking	Measuring speeding as a form of risk seeking with GPS (Bolderdijk et al., 2011).
Travel characteristics, such as travel mode, route choice, or speed	Studying travel behaviour, such as travel mode choice, route choice or speed with GPS (Murakami and Wagner, 1999; Draijer et al., 2000; Wolf, 2000, 2006; Stopher et al., 2002; Wolf et al., 2004; Bohte and Maat, 2009; Schuessler and Axhausen, 2009; Necula, 2015).
Virus transmission risk	Determining the spreading of disease with GPS (Vazquez-Prokopec et al., 2009, 2013).
Walking routes	Assessing tourist walking routes with GSM and GPS (Xia et al., 2008).

than, for instance, SPSS, there are pre-existing packages for spherical calculations and handling of GPS data (e.g., Kahle and Wickham, 2013; Hijmans et al., 2015; Loecher and Ropkins, 2015; Wickham, 2016), and since R is open-source software psyosphere can be improved upon by the research community. Furthermore, psyosphere allows researchers to analyse movement between persons (e.g., by calculating distances between persons while they are moving on different paths) and within-person differences (e.g., measuring movement changes over time). Additionally, it is possible to add custom measures to psyosphere.

Research outside of the laboratory has shown that pickpocketing corresponds with specific body movements (Bouma et al., 2014). Their algorithms to detect pickpockets were based on variations in walking speed, orientation change or

distance to other people, and were shown to have a sensitivity up to 95.6% with 0.5% false alarms. Similarly, the mental processes while smuggling, i.e., transporting an illegal package, may also lead to changes in behaviour that could be detected by the observers (Wijn et al., 2017).

The mental processes involved when transporting illegal substances can be linked to hostile intent. Wijn et al. (2017) define hostile intent "as an individual's intent to act in ways that imply or aim to inflict harm onto others" (p. 2), and although with smuggling the harm inflicted on others is perhaps more indirect and systemic—i.e., harm to the end users' health and society—they share essential characteristics. Indeed, people with hostile intent try to hide it when they expect that others will try to prevent their actions (Ekman et al., 1988; DePaulo et al., 2003; Koller et al., 2016; Wijn et al., 2017), and this should be no different for those who carry contraband. Furthermore, Wijn et al. (2017) suggest that the risk of being detected and apprehended—which would apply to smugglers as well—predisposes the psychological mind set of those harbouring such intentions towards increased anxiousness, self-focus, and vigilance. This, in turn, may influence their responses to environmental cues, specifically those that signal risk of exposure, causing a fear-related response pattern (e.g., fight, flight, or freeze) while transporting illegal substances. To illustrate how risk of exposure can influence body movement, a study on lie detection showed that participants became more rigid in their movement when instructed to tell a lie. Moreover, even when the participants were told that moving less would give them away, participants were unable to correct for the increased rigidity (Vrij et al., 1996; Van der Zee et al., 2019).

The current study focuses on establishing links between state variables and movement-descriptive variables in a smuggling game, requiring participants to transport legal or supposedly illegal material across a border area at the risk of being apprehended by border guards. As there is little in the way of theory to guide selection and construction of movement-specific variables in this context, the selection of behavioural variables—speed, speed variation, distance to peers, distance from a shortest route, and variation in the distance from the shortest route—was based on somewhat liberal interpretation of whatever literature was available.

The first variable, Speed, has been linked in past research to mental states and traits (Michalak et al., 2009; Bolderdijk et al., 2011; Barliya et al., 2012; Gross et al., 2012; Satchell et al., 2017), has been used successfully in detecting pickpockets (Bouma et al., 2014), and has also emerged as a strategy to avoid fear- and stress-inducing stimuli (Krieglmeyer et al., 2013). Moreover, environmental factors can influence walking speed. For instance, walking in nature decreases walking speed compared to urban settings, and walking speed decreased more when the setting was more natural (e.g., higher trees; Franěk and Režný, 2017). Interestingly, this effect could be linked to an environmental stressor, i.e., traffic noise (Franěk et al., 2018); apparently, stress caused by exposure to noise motivated people to minimise exposure to it by walking faster. Other work also shows that stimulus evaluations evoke approach-avoidance behaviours, e.g., suggesting that negatively evaluated stimuli cause people to

try and increase the distance between them and the stimulus (Krieglmeyer et al., 2013). Thus, smugglers fearing apprehension by border guards could for instance be expected to walk faster.

Distance to peers (Intra-Team Distance) is the mean distance from one person to the other persons within a group and past research has shown that people stay closer together when confronted with an outside threat (Schachter, 1959; Feshbach and Feshbach, 1963; Brady and Walker, 1978) and additionally, people keep a larger interpersonal distance when they are in uncertain situations with a threat to personal self-esteem (Schachter, 1959; Brady and Walker, 1978). Consequently, when transporting illegal substances and encountering a police officer, people could, for instance, be expected to walk closer together when they are afraid or walk further apart when they believe they are doing something illegal.

The variable Speed Variation is calculated as the standard deviation of the Speed between a start and finish of a series of coordinates. The distance from a shortest route (Route Deviation) is calculated as the distance between the measured points and the shortest possible route between two locations, and the variation in the distance from the shortest route (Variation Route Deviation) is the standard deviation of the Route Deviation for all data points between two locations. Subsequently, Route Deviation would increase if a person moves away further from the shortest route and Variation Route Deviation would increase if a person meanders while moving between two locations. To our knowledge, all three measures have not been used in past research. However, they could be argued to tie in either with fear-related response patterns, such as fight, flight, and freeze responses (Wijn et al., 2017), and rigidity (Vrij et al., 1996), or with strategies to avoid becoming close to a fear- or stress-inducing stimulus, i.e., the border guard (e.g., Krieglmeyer et al., 2013). Thus, a person transporting an illegal substance and encountering a guard may try to act normal by continuing on a straight path and keeping the same pace (cf. Vrij et al., 1996; Wijn et al., 2017), while avoidance tendencies could lead to behaviour changes, such as walking away from the guard, taking a longer route, changing the route more often or changing pace.

The proposed measures could also be used in other situations to study fear-related response patterns in combination with marked events. For instance, fear and avoidance responses can be measured when approaching meetings, classes, outgroup members or work environments, and may be used to analyse anxiety, depression, group association or workplace morale. Specifically, Intra-Team Distance, could be used to assess interpersonal relationships, fear, or feelings of guilt at a workplace environment. Moreover, Route Deviation and Variation Route Deviation could be used in wayfinding studies to assess if people are certain about their route taken, and Speed Variation could be used to analyse movement patterns during mass events, for instance, stop-and-go movement patterns signalling overcrowding in a specific area (Moussaïd et al., 2011).

To exemplify the possible use of psyosphere and the proposed measures, the current study encompasses two field experiments conducted as part of an undergraduate course for psychology students. During the experiments, the participants would wear GPS loggers and would transport cards either representing legal or supposedly illegal material across a border area with guards that could apprehend them and confiscate the cards. After each round, various state variables were measured. The two experiments serve as a "proof of principle," that given a certain context (i.e., hostile intent), aspects of people's movement could relate to their mental states. Thus, the experiments explore possible relationships rather than testing specific hypotheses pertaining to the behaviour and mental states of smugglers vs. non-smugglers.

MATERIALS AND METHODS

Participants

Two similar experiments were conducted as part of an undergraduate psychology course at the University of Twente. The first experiment took place in 2014 and the second in 2015. In the first experiment 69 students participated, who all received a GPS logger. Two were excluded from the analysis due to failure of GPS loggers and five others served as guards, and so 62 students (44 female and 18 male) remained as participants. The average age was 21.61 ($SD = 5.60$) and ranged from 18 to 37. Furthermore, 30 students were Dutch and 32 were German.

In the second experiment 91 students participated; the 80 available GPS loggers were randomly distributed among them. Of the 80 participants with loggers, three served as guards, and another three suffered from sensor failure. The remaining 74 students (51 female and 23 male) had an average age of 22.41 ($SD = 5.60$), ranging from 18 to 46; 38 were Dutch and 34 were German. The participants that acted as guards (five in Exp. 1, three in Exp. 2) were excluded from analysis to limit the scope of the current study.

Power analyses were not conducted in order to determine a required number of participants; instead, we were constrained by the number of students who had enrolled in the courses in 2014 and 2015, respectively, as well as the numbers of GPS loggers that were available. For logistical and practical reasons, additional data collection under comparable circumstances in order to meet pre-determined numbers of participants was virtually impossible.

Procedure

The participants signed up for the experiment during an introductory lecture for an undergraduate course. The experiment was explained to the participants and they received written instructions. Afterwards they signed an informed consent form. On the basis of randomness participants were either assigned to one of the smuggling teams or to act as one of the guards. In Exp. 1 group sizes ranged between three and six ($M = 4.77$, $SD = 0.83$); in Exp. 2 they ranged between two and six ($M = 4.63$, $SD = 1.36$). The participants received a GPS logger and were told to gather 3 h later in a small park on the university campus.

Tasks

The teams (smugglers) had the task to transport supposed legal and illegal material and the other participants, i.e., guards, were required to intercept those with illegal material. This "material"

consisted of a card the size of a playing card that either displayed cocaine ("illegal card") or flour ("legal card"). They received both legal and illegal cards in the starting area, and distributed these among each other, making sure that each participant carried one (either legal or illegal). The cards stated that the teams would win 10 points for each illegal card they transported and one point for each legal card. Guards would also score 10 points for intercepting in illegal card, but lose one point if they intercepted a legal card. The best team's members and best guard would each win a cinema voucher and a chocolate bar.

Before starting the teams had to write their GPS logger's number and the starting time on the card.

Area

After arriving at the park, participants were directed to their assigned locations. The teams would go to a starting point that was behind a mount and out of sight of the guards. The guards were waiting at the finish in an approximately 2-m-wide and 20-m-long strip that the teams were required to cross. The finish area was marked with barrier tape on the ground. A group of 17 tall trees were standing inside and around the finish area. A visual inspection of the data did not reveal signal distortions by the trees. The distance from the starting area to the finish area was 150 m. Roughly halfway between start and finish was a semi-circular bicycle path that was slightly elevated; as a result, the teams and guards could observe each other only after the former had walked onto the bicycle path. The guards were positioned in the centre of the semi-circle in about 75 m from the edge of the semi-circle. See **Figure 1** for an illustration of the area.

Round

The teams were instructed to walk from the starting area, across the park and through the finish area. The guards could confiscate the team members' card by tapping them on their shoulder. The team member would give their card to the guard and the guard would note a number on the card which was assigned to the guard. The team members had to avoid being checked by the guards. This could be done by, for instance, distracting them by sending the team members with legal cards first or walking with a wide distance among each other therefore it would be difficult for the guards to reach all team members before they crossed the finish area. The guards were not allowed to leave the finish area and had to wear safety vests to enable the team members to spot them easily. Each time after crossing the finish area the team members would drop the remaining cards they had into a box and fill in a questionnaire. After this they would walk back to the start position for another Round.

Experiment 1

Five participants were assigned to be guards and the other participants formed teams of three to six members, with an average of 4.77. Before the start of the first experiment the participants filled in an additional trait questionnaire. Further, participants were instructed not to run, and four rounds were conducted. Additionally, all teams were wearing a card with their team number on their chest. One team of five participants was asked to wear skin conductivity sensors to measure stress levels,

but these all failed. At the starting area each team received a set of cards, two of which were "illegal" cards; teams could decide themselves who would smuggle. Afterwards, between four and five teams with an average of 21.08 participants would start at the same time and the ratio between guard's to participants was 0.24 when the participants approached the finish area. After each Round the teams would fill in a Dutch version of the State questionnaire (see **Supplementary Appendix 2**).

Experiment 2

Experiment 2 differed slightly from 1. First, three (rather than five) participants were made guards; the others formed two- to six-member teams, with an average of 4.63. Furthermore, the instruction not to run was omitted, and the participants were not wearing any stress sensors or team numbers. At the starting area the teams could freely choose the ratio of illegal and legal cards and they were asked to write down which strategy they wanted to use. Additionally, the finish area in Experiment 2 was larger than in Experiment 1 and enabled the guards to walk more freely. At the end of each Round, the team members would write down their points and could see the total points of the other teams. Finally, they would only fill in an English version of the State questionnaire. Some questions were removed and a few were added in the State questionnaire; for instance, participants were asked to rank the extent to which they thought each team member acted as a leader, and to describe how they thought they could improve their strategy as a team in the next round (these are beyond the scope of this paper, however). The guards were frequently informed how many points they had obtained.

Measures
State Questionnaire

The mental states of the participants were measured with a questionnaire based on Stekkinger (2012) and Wijn et al. (2017). Some questions were amended for a better fit to the current context and two questions were added to measure self-observed behaviour changes (e.g., whether the participants changed their pace after seeing the guards). **Supplementary Appendix 2** contains all questions, moreover, **Supplementary Appendix Tables 10, 11** report the Cronbach's alpha or Pearson's R for the state questionnaire constructs. The reliabilities of all scales were above 0.78. All State questions used a 7-point Likert scale from 1 "Not at all" to 7 "Very much."

Two items checked to what extent participants experienced hostile intentions (Hostile Intent; Stekkinger, 2012; Wijn et al., 2017). Five items measured the participants alertness to threats from the guards (Alertness to Being Target of Guards; Galbraith et al., 2008; Stekkinger, 2012; Wijn et al., 2017). Five items checked the inhibitory and activation control (Cognitive Self-Regulation; Stekkinger, 2012; Wijn et al., 2017). Four items measured the self-focus of the participants (Situational Self Awareness; Govern and Marsch, 2001; Stekkinger, 2012; Wijn et al., 2017). Four items assessed the feelings of fright that the participants felt through the presence of the guards (Frightened by Presence of Guards; Stekkinger, 2012; Wijn et al., 2017). Five items checked the impulses that were suppressed by the participants (Suppressed Impulses to Change Movement;

FIGURE 1 | Experiment area with participant tracks and GPS polygons. The tracks of six team members in Experiment 2 are plotted in black (illegal card) and yellow (legal card). They started in polygon **(A)**, crossed the semi-circular bicycle path into the area where they would be visible to the guards (polygon **B**), and ended before the finish (polygon **C**).

Stekkinger, 2012). Three items measured the extent that participants questioned themselves (Contemplation of Hostile Intent; Stekkinger, 2012; Wijn et al., 2017). Finally, two items were added to the questionnaire that assessed the self-observed behaviour changes of participants (Awareness Movement Change in Presence of Guards; Stekkinger, 2012). For a detailed explanation of the mental processes and their function, see Wijn et al. (2017).

GPS Data

Every participant carried an i-gotU GT-600 GPS logger with a SiRF Star III GPS Chip and a L1, 1575.42 MHz frequency. The sensors were turned on at least 5 min before the start of the experiments and checked if they were recording. Every second,

the logger received time, latitude, longitude, and elevation and saved these in a data point. From the GPS data Speed, Speed Variation, Intra-Team Distance, Route Deviation and Variation Route Deviation were calculated in 1-s increments using the R package psyosphere. Speed was measured as the mean kilometres per hour between each data point. Speed Variation was calculated as the standard deviation of the kilometres per hour between each data point between start and finish. Intra-Team Distance is the mean distance from one team member to the other team members in metres. Route Deviation is the distance in metres between the data points and the shortest route from start to finish. Variation Route Deviation is the standard deviation of the Route Deviation for all data points from start to finish. The start was the point where a participant entered the starting area (see polygon

A in **Figure 1**) and the finish was the point where a participant crossed the finish line (see polygon C in **Figure 1**). Therefore, the exact start and finish point could differ for each participant.

Analysis
GPS Data Preparation
The data from the GPS loggers were exported and analysed with the R package psyosphere, developed for the current study (Ziepert et al., 2018). The software created a track for each Round of each participant, and plotted the tracks on a map retrieved from Google maps (Google LLC, 2018). Polygons could be defined to mark areas of specific interest (see **Figure 1** for map, tracks, and polygons).

A track began in the starting area that was determined by a polygon of GPS coordinates (polygon A) and ended when the participants crossed a GPS based finish line behind the finish area (polygon C). In between was the area where team members and guards could actually see each other (polygon B). Before the point of visibility, the teams generally followed a straight line and started to change their movement after seeing the guards. Therefore, the analysis included only the data from the point where teams became visible to the guards and vice versa, to where the members crossed the finish line. Within this line-of-sight area 31,113 coordinates were recorded in Experiment 1, for four rounds, and 17,172 in Experiment 2, for three rounds. These data were used by the R package to calculate the GPS variables mentioned above.

The R package detects two types of faulty data. First, an unrealistic speed can, for instance, be recorded due to temporary signal loss from the GPS satellites. Therefore, if the speed exceeded 40 km/h the data was excluded from analysis. This occurred 16 times (0.05%) in Experiment 1 and 8 times (0.05%) in Experiment 2. Second, if the time difference between coordinates exceeded 1 s then the Speed, Speed Variation and Distance between the coordinates were marked as missing values and also excluded from analysis. Three coordinates (0.01%) in Experiment 1 and 152 coordinates (0.89%) in Experiment 2 were excluded because of a time difference larger than one second. Additionally, two sensors in Experiment 1 and three sensors in Experiment 2 stopped recording all GPS satellite data and had to be excluded from the analysis.

State PCA
We analysed the State questions that were used in both experiments with a principal component analysis (PCA) and a confirmatory factor analysis (CFA). In total, six explorative PCAs were conducted, one for each of the three rounds in the two experiments. Afterwards, we compared the PCAs and counted how often items shared a component. A model with eight components emerged and was subsequently tested with a CFA.

Relationships Between State and GPS Variables
Descriptive statistics and correlations, for the State components and GPS variables, were calculated for each experiment separately. Finally, a multi-level analysis was conducted with the GPS variables as dependent variables, and the State components and rounds as the predictors. In total, 10 models were created, 5

for each experiment. The multi-level analysis tested for consistent changes per Round (e.g., increasing Intra-Team Distance per Round) and the impact of grouping in teams. Three random effect models did not converge, and two of these models were models with a maximum random-effects structure based on the experimental design. Moreover, according to Barr et al. (2013), a maximum random-effects model should be prioritised when conducting a multilevel analysis. Therefore, we choose for a maximum random-effects model with random slopes per Round and a static intercept per team and participant. To improve the model convergence rate, Barr et al. (2013) suggest to remove outliers, and therefore, data have been removed from the GPS variables, except Intra-Team Distance, when the data were outside of the Inter Quartile Range times 1.5 (Hoaglin, 2003). Fifteen outliers have been removed from Speed, nine from Speed Variation, five from Route Deviation, and fourteen from Variation Route Deviation. After removing the outliers, all models with a maximum random-effects structure converged. Intra-Team Distance was excluded from the outlier removal since this increased the model convergence rate.

Exclusions
Only the first three rounds of both experiments were used since the participants did not complete the State questionnaire after the fourth Round of Experiment 1. For educational purposes, a limited number of participants in Exp. 1 were fitted with skin conductivity to measure stress levels; these data could not be used due to sensor malfunctions. Questions pertaining to intended strategy changes in the next round, leadership and motivation were also outside of the scope of this paper and are not analysed. In Exp. 1 two GPS loggers failed, and in Exp. 2 three.

RESULTS

We first conducted principal component analyses (PCA), confirmatory factory analyses (CFA), and invariance model comparisons for 30 items of the state questionnaire for each experiment and for each of the first three rounds. The questionnaire can be found in **Supplementary Appendix 2**. The descriptive statistics, correlations tables, and the CFA results can be found in **Supplementary Appendix 3**; a section describing notable differences between the experiments can also be found in this **Supplementary Appendix**.

Regression Analysis
Per experiment and for each of the five GPS-based outcome variables regression analyses were conducted with the state constructs as predictors. These 10 regression models were tested for random effects per participant, team, and Round. Testing for random effects is necessary since the measurements for each participant are not independent but depend on the Round that is measured and the team a participant is in. For instance, a team with fast walking members could have motivated a slow walking member to walk faster. Barr et al. (2013) suggest to select a maximum random-effects model based on the experimental design instead of selecting a model based on the model fit. For

our current study, a maximum random-effects model includes random slopes per round and a static intercept per team and participant (see Eq. 1). Finally, we were interested in the effect of all state variables on the GPS outcome variable (e.g., speed), and consequently, the State variables were added as predictors to the regression model. See Eq. 2 for a simplified version of the fixed effects formula.

$$\sim Round \mid \frac{Team}{Participant} \qquad (1)$$

$$
\begin{aligned}
GPS = \ & Illegal\ Card\ Selection\ \beta + Self\ as\ Target\ \beta \\
& + Awareness\ Cognitive\ Behaviour\ Change\ \beta \\
& + Situational\ Self\ Awareness\ \beta + Fright\ \beta \\
& + Impulse\ \beta + Dubious\ Thoughts\ \beta \\
& + Awareness\ Physical\ Behaviour\ Change\ \beta \qquad (2)
\end{aligned}
$$

Model for Speed

We calculated Model 6.2 with Speed as the outcome variable and **Table 2** displays the results per estimate. As **Table 2** highlights, Awareness Movement Change in Presence of Guards was a significant and positive predictor for Speed in Experiment 1 and the same relationship was not significant in Experiment 2 ($b_1 = 0.08$, $p_1 < 0.001$, $b_2 = 0.05$, $p_2 = 0.241$). Thus, the more participants reported a speed or route change after seeing the guards, the faster they walked. Likely, they did so in an attempt to outpace or evade the guards.

Additionally, in Experiment 1 Suppressed Impulses to Change Movement is a significant and negative predictor for Speed though the same relationship was not significant in Experiment 2 ($b_1 = -0.09$, $p_1 = 0.008$, $b_2 = -0.07$, $p_2 = 0.195$). This means that people who suppressed their impulses also walked slower, and an explanation could be that participants walked slower in order not to attract the attention of the guards. An alternative explanation

could be that participants were uncertain which route would be the best to avoid the guards and therefore slowed their pace.

Model for Speed Variation

We calculated Model 6.2 with Speed Variation as the outcome variable and **Table 3** displays the results per estimate. The table shows that, Suppressed Impulses to Change Movement was a significant a positive predictor in Experiment 1 but not in Experiment 2 ($b_1 = 0.12$, $p_1 < 0.001$, $b_2 = 0.03$, $p_2 = 0.575$). This means that when the participants had suppressed impulses, then they varied their walking pace more. A simple explanation could be that participants failed in suppressing their impulses and therefore varied more. However, as **Table 2** shows, participants reduced their pace when they had suppressed impulses ($b_1 = -0.09$, $p_1 = 0.008$, $b_2 = -0.07$, $p_2 = 0.195$) and if participants had failed in suppressing their impulses, one would suspect that their overall pace would have increased and not decreased. Hence, an alternative explanation could be that Suppressed Impulses to Change Movement measures the uncertainty of the participants on how to avoid the guards rather than suppressed impulses. Accordingly, the uncertainty could have caused the participants to slowdown, in order to orient themselves, and then to follow the new path with an increased pace.

Furthermore, Round is a positive and significant predictor for Speed Variation in Experiment 1; the same relationship is not significant in Experiment 2, however ($b_1 = 0.08$, $p_1 < 0.031$, $b_2 = 0.05$, $p_2 = 0.575$). Consequently, with each consecutive Round the participants varied more in their pace, and the variation could have helped the participants to avoid the guards better.

Additionally, Alertness to Being Target of Guards is a positive and significant predictor for Speed Variation, though the same relationship was not significant in Experiment 2 ($b_1 = 0.05$, $p_1 = 0.008$, $b_2 = 0.07$, $p_2 = 0.139$). Namely, participants feeling targeted by the guards varied their speed more. A likely

TABLE 2 | Regression beta, SE, and *P*-values for speed as dependent variable.

Estimate	Experiment 1			Experiment 2		
	b	*SE*	*p*	*b*	*SE*	*p*
Round	0.05	0.04	0.232	0.11	0.07	0.116
Illegal Card Selection	−0.03	0.07	0.636	0.04	0.18	0.819
Alertness to Being Target of Guards	−0.03	0.02	0.166	−0.01	0.04	0.721
Cognitive Self-Regulation	−0.02	0.02	0.395	0.01	0.05	0.832
Situational Self Awareness	−0.01	0.03	0.687	−0.05	0.05	0.339
Frightened by Presence of Guards	0.03	0.03	0.290	−0.00	0.05	0.951
Suppressed Impulses to Change Movement	**−0.09**	0.03	**0.008**	−0.07	0.05	0.195
Contemplation of Hostile Intent	0.02	0.03	0.527	−0.03	0.05	0.537
Awareness Movement Change in Presence of Guards	**0.08**	0.02	**<0.001**	0.05	0.04	0.241

P-values less than 0.050 are in bold.

TABLE 3 | Regression beta, SE, and *P*-values for speed variation as dependent variable.

Estimate	Experiment 1			Experiment 2		
	b	*SE*	*p*	*b*	*SE*	*p*
Round	**0.08**	0.04	**0.031**	0.05	0.09	0.575
Illegal Card Selection	0.11	0.06	0.088	−0.16	0.22	0.449
Alertness to Being Target of Guards	**0.05**	0.02	**0.008**	0.07	0.04	0.139
Cognitive Self-Regulation	−0.03	0.02	0.217	0.07	0.06	0.285
Situational Self Awareness	0.04	0.02	0.117	−0.11	0.06	0.069
Frightened by Presence of Guards	−0.03	0.03	0.207	−0.01	0.06	0.885
Suppressed Impulses to Change Movement	**0.12**	0.03	**<0.001**	0.03	0.06	0.575
Contemplation of Hostile Intent	−0.05	0.03	0.065	−0.04	0.06	0.468
Awareness Movement Change in Presence of Guards	**-0.06**	0.02	**0.002**	0.00	0.05	0.933

P-values less than 0.050 are in bold.

	Experiment 1			Experiment 2		
Estimate	**b**	**SE**	**p**	**b**	**SE**	**p**
Round	**2.24**	0.73	**0.003**	0.85	0.46	0.067
Illegal Card Selection	0.17	0.58	0.770	0.08	0.70	0.906
Alertness to Being Target of Guards	0.18	0.18	0.323	0.16	0.14	0.262
Cognitive Self-Regulation	0.13	0.19	0.497	0.27	0.18	0.129
Situational Self Awareness	−0.32	0.20	0.119	−0.06	0.18	0.719
Frightened by Presence of Guards	**−0.62**	0.27	**0.023**	−0.20	0.17	0.242
Suppressed Impulses to Change Movement	0.19	0.27	0.479	−0.01	0.18	0.953
Contemplation of Hostile Intent	**0.52**	0.24	**0.034**	0.09	0.15	0.580
Awareness Movement Change in Presence of Guards	−0.07	0.16	0.665	−0.13	0.15	0.413

P-values less than 0.050 are in bold.

	Experiment 1			Experiment 2		
Estimate	**b**	**SE**	**p**	**b**	**SE**	**p**
Round	**2.14**	0.52	**<0.001**	0.18	0.53	0.738
Illegal Card Selection	−0.26	0.58	0.656	−1.89	1.32	0.155
Alertness to Being Target of Guards	−0.32	0.17	0.068	**0.62**	0.28	**0.031**
Cognitive Self-Regulation	−0.26	0.21	0.226	0.42	0.39	0.279
Situational Self Awareness	0.11	0.22	0.624	−0.10	0.37	0.778
Frightened by Presence of Guards	−0.27	0.28	0.335	−0.06	0.36	0.875
Suppressed Impulses to Change Movement	**0.63**	0.28	**0.028**	0.28	0.39	0.484
Contemplation of Hostile Intent	0.01	0.26	0.958	−0.52	0.35	0.138
Awareness Movement Change in Presence of Guards	−0.04	0.18	0.825	0.25	0.31	0.415

P-values less than 0.050 are in bold.

explanation is that participants tried to avoid the guards by changing their walking pace.

Finally, Awareness of Movement Change in Presence of Guards was a significant and negative predictor for Speed Variation in Experiment 1 but not in Experiment 2 ($b_1 = -0.06$, $p_1 = 0.002$, $b_2 = 0.00$, $p_2 = 0.993$). This means that participants' awareness that they changed their route or speed after seeing the guards, corresponded with less variation in their walking pace. A reason could be, that participants had chosen a route after seeing the guards that successfully avoided the guards and therefore the participants could keep their pace. Because of the lower guard ratio in Experiment 1, it was easier to avoid the guards than in Experiment 2.

Model for Intra-Team Distance

Model 6.2 was calculated with Intra-Team Distance as the outcome variable and **Table 4** displays the results per estimate. The table shows that Round was a positive and significant predictor for Intra-Team Distance in Experiment 1 and the same relationship was close to significant in Experiment 2 ($b_1 = 2.24$, $p_1 < 0.003$, $b_2 = 0.85$, $p_2 = 0.067$). This means that the distance to other team members increased with each Round, for instance, implying the emergence of a strategy to better avoid the guards.

Furthermore, Frightened by Presence of Guards was a significant and negative predictor for Intra-Team Distance in Experiment 1 while the same relationship was not significant in Experiment 2 ($b_1 = -0.62$, $p_1 = 0.023$, $b_2 = -0.20$, $p_2 = 0.242$). Therefore, when participants had feelings of fright because of the guards then they walked closer together, possibly to compensate for their fear. Alternatively, observing that they walked close together may also have made them realise they were more likely to be a target, which may in turn have inspired feelings of fright.

Additionally, Contemplation of Hostile Intent was a significant and positive predictor for Intra-Team Distance in Experiment 1 and the same relationship was not significant in Experiment 2 ($b_1 = 0.52$, $p_1 = 0.034$, $b_2 = 0.09$, $p_2 = 0.580$).

The more participants were questioning the legality of their actions or whether they have to hide something, the further they would walk apart from their fellow team members. It is possible, the participants had conflicting emotions about their hostile intentions and therefore did not want to affiliate with their team members.

Model for Route Deviation

Model 6.2 was conducted with Route Deviation as the outcome variable and **Table 5** displays the results per estimate. As the table shows, Alertness to Being Target of Guards is a significant and positive predictor for Route Deviation in Experiment 2 and a close to significant and negative predictor in Experiment 1 ($b_1 = -0.32$, $p_1 = 0.068$, $b_2 = 0.62$, $p_2 = 0.031$). Therefore, in Experiment 1, participants that perceived themselves as a target by the guards kept a shorter distance to the shortest route, and in Experiment 2, participants did the opposite. The difference could be explained by the change in the guard ratio. In Experiment 1, the guard had to be selective and participants who felt they were a target could try to act as normal as possible by deviating less from the shortest route. In Experiment 2, the guards could stop all participants if they were fast enough, and therefore, when participants would feel themselves a target would need to actively avoid the guards by outwalking them.

Additionally, Round is a significant and positive predictor for Route Deviation in Experiment 1, but the same relationship is not significant in Experiment 2 ($b_1 = 2.14$, $p_1 < 0.001$, $b_2 = 0.18$, $p_2 = 0.738$). This means that each Round participants walked further away from the shortest route. Possibly, they increasingly realised that a higher route deviation helped in avoiding the guards. Additionally, the higher guard ratio in Experiment 2 made it easier to pursue participants and therefore Route Deviation was greater in Experiment 2 from the start than in Experiment 1, and consequently, could not be increased as much as in Experiment 1.

Finally, Suppressed Impulses to Change Movement is a significant and positive predictor for Route Deviation in Experiment 1 and the same relationship was not significant in Experiment 2 ($b_1 = 0.63$, $p_1 = 0.028$, $b_2 = 0.28$, $p_2 = 0.484$). Therefore, participants who reported suppressed impulses also deviated more from the shortest route. An explanation could be that participants were uncertain about their route and therefore deviated more from the shortest route.

Model for Variation Route Deviation

We calculated Model 6.2 with Variation Route Deviation as the outcome variable and **Table 6** displays the results per estimate. As the table shows, Suppressed Impulses to Change Movement is a significant predictor for Variation Route Deviation in Experiment 1 and Experiment 2 ($b_1 = 0.34$, $p_1 = 0.033$, $b_2 = 0.37$, $p_2 = 0.017$). The more participants reported suppressed feelings, the more often they changed their routes. A reason could be that the participants were uncertain about the route to avoid the guards and therefore changed it more often.

Additionally, Round is a negative and significant predictor for Variation Route Deviation in Experiment 2 and the opposite relationship is not significant in Experiment 1 ($b_1 = 0.10$, $p_1 = 0.773$, $b_2 = -1.05$, $p_2 = 0.049$). Thus, those who selected an illegal card in Experiment 2 changed their route less often. An explanation is the small number of participants that carried a legal card in Experiment 2 and which were used by the teams to distract the guards by changing their route more often.

Furthermore, Illegal Card Selection is a negative and significant predictor for Variation Route Deviation in Experiment 2 but is not significant in Experiment 2 ($b_1 = 0.12$, $p_1 = 0.721$, $b_2 = -1.10$, $p_2 = 0.045$). Therefore, when participants chose an illegal card in Experiment 2 they changed their route less often. An explanation could be that participants with a legal card changed their route more often to attract the attention of the guards and that participant with an illegal card did the opposite. This strategy could have been more important in Experiment 2

since the area where the guards were allowed to walk was larger in Experiment 2. Consequently, when a participant got the attention of the guard, the guard had to walk further away from other participants.

Moreover, Alertness to Being Target of Guards is a positive and significant predictor for Variation Route Deviation in Experiment 2 but is not significant in Experiment 1 ($b_1 = -0.14$, $p_1 = 0.164$, $b_2 = 0.29$, $p_2 = 0.009$). This means that the more participants perceived themselves as a target by the guards, the more participants would change their route in Experiment 2 while the opposite happened in Experiment 1. The reason could be that in Experiment 2 every participant could be checked and when the participants perceived themselves as a target they actively avoided the guards by changing their route more often. In Experiment 1, in contrast, not all participants could be checked and the participants could try to act normally to reduce guard suspicion.

Finally, Awareness Movement Change in Presence of Guards is a positive and significant predictor for Variation Route Deviation in Experiment 1, although the same relationship is not significant in Experiment 2 ($b_1 = 0.23$, $p_1 = 0.024$, $b_2 = 0.08$, $p_2 = 0.529$). Thus, participants' awareness that they had changed their route or speed after seeing the guards corresponded with actual behaviour. A reason could be that the participants attempted to outmanoeuvre the guards by changing the route after seeing them.

Summary

In summary, the results show that the participants used strategies to avoid the guards. For instance, the participants changed their behaviour with each consecutive Round, by increasing the distance to team members, by accelerating and decelerating more often, by taking longer routes, and by changing the route more often. These changes may indicate a collective strategy by the participants to become better in avoiding the guards. Additionally, according to their descriptions in Experiment 2, teams made use of a distraction strategy: they chose to carry a legal card and distracted the guards, so as to improve the chances of their illegal-card carrying team members.

Participants were presumably uncertain about the best route to avoid the guards, and this may have become overt in more changes in direction, reduced pace, more changes in pace, and increased route lengths. Additionally, participants stayed closer to team members when they had feelings of fear and kept a greater distance if they had the feeling that they had to hide something. Furthermore, participants attempted to avoid guards by changing the pace more often when targeted, by increasing the pace after seeing the guards, and by changing the route more often after seeing the guards.

In Experiment 1 the participants had two illegal cards per team and in Experiment 2 the participants could choose a free ratio of legal and illegal cards. Therefore, the increased availability of illegal cards presumably reduced the relationship between the selection of an illegal card and feeling of hostile intent. Another difference between the experiments was the ratio of guards and participants. Specifically, in Experiment 2 were more guards

TABLE 6 | Regression beta, SE, and *P*-values for variation route deviation as dependent variable.

Estimate	Experiment 1			Experiment 2		
	b	**SE**	**p**	**b**	**SE**	**p**
Round	**0.82**	0.26	**0.002**	0.08	0.22	0.734
Illegal Card Selection	0.10	0.33	0.773	**-1.05**	0.53	**0.049**
Alertness to Being Target of Guards	–0.14	0.10	0.164	**0.29**	0.11	**0.009**
Cognitive Self-Regulation	–0.19	0.12	0.118	0.24	0.15	0.119
Situational Self Awareness	0.12	0.13	0.368	–0.19	0.15	0.188
Frightened by Presence of Guards	–0.20	0.16	0.207	–0.21	0.14	0.150
Suppressed Impulses to Change Movement	**0.34**	0.16	**0.033**	**0.37**	0.15	**0.017**
Contemplation of Hostile Intent	0.01	0.15	0.944	–0.23	0.14	0.093
Awareness Movement Change in Presence of Guards	**0.23**	0.10	**0.024**	0.08	0.12	0.529

P-values less than 0.050 are in bold.

per participant than in Experiment 1 and that made it more difficult for the participants to avoid the guards in Experiment 2. Consequently, when participants perceived themselves as target by the guards, the participants in Experiment 1 took a more direct path and made less changes to their direction in order not to attract further attention by the guards. Moreover, the participants in Experiment 2 did the opposite, in an attempt, to outmanoeuvre the guards and took a longer route and made more changes to their route. Finally, in Experiment 1, when participants saw the guards they reduced their speed in order not to attract any attention and a similar effect could not be found in Experiment 2.

DISCUSSION

The aim of the current study was to develop a research tool that enables behavioural scientists to more easily use positioning technologies, such as GPS, for psychological experiments. Additionally, two experiments were conducted as cases against which to test this new research tool, psyosphere.

Psyosphere

The R package psyosphere (Ziepert et al., 2018) analyses GPS data by transforming GPS tracks into descriptive variables, such as speed, direction or distance, that can be analysed with linear regression methods. It is optimised to handle multiple tracks simultaneously and to make comparisons between these tracks. This is done because comparisons between multiple participants with linear regression methods is a typical technique of conducting studies in behavioural science. To give a simplified example, the speed of multiple car drivers for a given route could be compared to investigate if speed warnings reduce risky driving behaviour. Furthermore, the package supports data preparation through cleaning up the data by marking coordinates with unrealistic speeds as missing values or by detecting measuring gaps. Additionally, sub-tracks can be selected by providing start and finish areas. The package also supports the visualisation of tracks. For this purpose, tracks and polygons can be plotted on maps, tracks can be coloured based on grouping variables, and tracks can be plotted per participant or groups of participants (e.g., teams; see **Figure 1**).

Psyosphere builds on existing R packages (e.g., Kahle and Wickham, 2013; Hijmans et al., 2015; Loecher and Ropkins, 2015; Wickham, 2016) and on recent work also aiming to unlock the potential of location-based data for psychologists (Geyer et al., 2019). Whereas, Geyer et al. (2019) predominantly focused on creating an Android app allowing accurate location logging and secure storage of data, and, in addition, offer assistance with subsequent data analysis, the psyosphere package focuses specifically on the latter. It complements this earlier work by increasing the possibilities for analyses and enhancing the potential of location data even further. By breaking up a sequence of timestamped coordinates into fine-grained facets of movement behaviours, psyosphere may be of interest for behavioural and psychological researchers.

Psychological Variables

To illustrate which type of variables could be studied with positioning technologies and psyosphere, an overview of variables that were used in past research was provided (see **Table 1**). Additionally, as a test case, we conducted two experiments to analyse the relationship between several psychological variables (i.e., feelings of hostile intent and several related states) and movement data measured with GPS loggers. The two experiments have shown relationships that have face validity, correspond with literature, or occur in both experiments. Due to the differences in experimental procedures, both studies cannot be seen as direct, exact replications, which may have caused some findings to occur in only one.

For instance, one finding was that participants who reported higher levels of fear also tended to walk closer together. This finding is in line with past research, demonstrating that people stay closer together when confronted with an outside threat (Schachter, 1959; Feshbach and Feshbach, 1963; Brady and Walker, 1978). In light of the correlational nature of our research, however, this finding is amenable to other interpretations as well.

Additionally, when participants were contemplating whether they were doing something illegal and whether they had to hide something, this corresponded with larger distances to their team members. Congruently, participants in uncertain situations with a threat to personal self-esteem have been shown to keep a larger interpersonal distance (Schachter, 1959; Brady and Walker, 1978). Although self-esteem was not measured here, the item to what extent participants believed that they were doing something illegal could be interpreted as self-esteem related.

Furthermore, participants developed evasive strategies over the three rounds to avoid the guards. In detail, the participants spread out more, took longer routes and changed their route and pace more often. Presumably, the evasive strategies gave the guards fewer opportunities to stop participants and check whether they had illegal cards. Similarly, the second experiment suggested that teams used distraction strategies to improve the overall team score. To distract the guards, one or two team members would carry a legal card, would walk ahead of the team members, with illegal cards, and would show an erratic movement, such as changing the route more often to attract the attention of the guards. In a somewhat similar pen-and-paper experiment, researchers asked participants to draw a route from a starting position to a designated target, without giving away their final destination (Jian et al., 2006). The experiments showed that participants would take a longer route with erratic movement, such as changing direction more often, to hide their intended target. The findings of Jian et al. (2006) are comparable with the evasion strategies observed in the current study and we argue that participants will use evasive strategies if they judge this to be normal, i.e., when other people around them also perform this behaviour as well. Furthermore, participants can use evasive movements that deviate from they believe to be normal movements to purposefully create suspicion.

Finally, when participants were presumably uncertain about their route, they also showed erratic movement, such as changing the route more often, taking longer routes, changing the

pace more often and overall walking slower. To test whether participants were actually uncertain, a future study could include related measures. An alternative explanation could be that participants felt regret about the route they chose because they got caught. It would therefore be beneficial to include regret-related measures in future studies as well.

Limitations

Arguably, the ratio between participants and guards differed between the two studies, and this may have influenced the relationship between the self-reported mental states and the measured GPS variables. Specifically, when the participants were carrying illegal cards, we assumed that they would try to hide this fact before the guards and would try to act normal. To act normal, the participants had to suppress fear-related responses, such as running away. Furthermore, the suppression of fear-related responses requires effort, and cues from the surroundings, such as encountering a guard or being targeted by a guard, could limit the ability of participants to act normal. Therefore, we measured whether the participants changed their movement when they encountered the guards. In the second experiment it was much more likely that participants would be stopped than in the first experiment. As a consequence, the guards were much less selective in stopping participants in the second experiment, and therefore, participants opted more for openly evading guards than trying to act normal. Thus, the higher guards-to-participants ratio may have been the reason that a smaller number of significant relationships between mental states and GPS variables was found in the second experiment compared to the first experiment. Nevertheless, we believe to have found some meaningful relationships, and we advise that future research should limit the number of guards to ensure that not all participants can be checked.

Another limitation of the current study is that we tested for 90 regression estimates, which renders the probability of finding statistically significant relationships merely by chance (Type I error inflation) rather high. In principle, it is possible to correct for this chance by reducing the significance level with, for instance, a Bonferroni correction (Holm, 1979), but it is a matter of long-standing methodological discussion exactly when and how this significance level should be adjusted (e.g., Fisher, 1956; Cabin and Mitchell, 2000). However, as we set out to merely explore our field-experimental data using the psyosphere tool to assess whether it can indeed be used to supplement more conventional means of data gathering, rather than testing specific hypotheses pertaining to the behaviour and mental states of smugglers vs. non-smugglers, we feel it is justified to simply accept a higher risk of false positives (for a discussion, see Wigboldus and Dotsch, 2016). For the current study we therefore chose not to correct the significance level, and, instead, focused on those findings that have some face validity, correspond with similar findings elsewhere, or are sufficiently robust to occur in both experiments.

The explorative nature of this exercise unfortunately sheds little light on the ecological validity of the findings presented here. Indeed, without evidence on the contrary, it would be safest to assume that interpretation of observed behaviour varies between contexts. Walking closer together, for instance, was associated with fear, but in other contexts, such as shopping areas or festivals, might be more likely as a sign of enjoyment in the company of friends. Similarly, changing route and pace was interpreted here as a deliberate strategy to avoid a guard, but in a way-finding context might just as well mean that one does not know where to go, of cannot make up one's mind about which path to choose.

The objective of this exercise, however, was not to find universal behavioural correlates of mental states per se, but to yield "proof of principle," i.e., that, given a certain context, aspects of people's movements could be used to shed light on their states. In this sense, the analyses provide here appear to have succeeded—be it in part. The possibility that these behavioural aspects may imply different emotions, cognitions, or intentions in different contexts is by no means detrimental to our conclusion. However, it does imply that researchers should exercise caution in interpreting such correlations. Extensive pilot testing of which aspects of observed behaviour are correlated with self-reported measures under controlled circumstances before going out into the field is key.

Future Research and Developments

On a more technical note, past research has shown that detecting movement patterns is dependent on the sensor accuracy (Kjargaard et al., 2013). An older version of the sensors that were used in the current study had an accuracy between 2.50 and 20 m (Vazquez-Prokopec et al., 2009). Research has shown that sensor accuracy can be greatly improved by combining multiple satellite systems, such as GPS, Glonass, Galileo, and BeiDou. These accuracy improvements will allow detection of movement patterns in more detail, making it easier to link them to cognitive processes.

Alternatively, also Wi-Fi and GSM signals can be used to determine the location. Smart phones, internet-of-things (IOT) devices, and specialised hardware can record the Wi-Fi and GSM signals and deduce locations (Kjargaard et al., 2013). It is also possible to track Wi-Fi and GSM-enabled devices from a GSM tower (Bengtsson et al., 2011) or with Wi-Fi routers (Sevtsuk, 2009) even if the devices are not connected to the network. Cameras (Burgess et al., 2014), Bluetooth (Madan et al., 2010), and RFID (Isella et al., 2011a) are yet other technologies offering data that, after conversion to (X, Y) coordinates are amenable to psyosphere analyses.

Thus, psyosphere may facilitate research in a number of domains, such as crowd control during events and behaviour of people in emergencies. Sime (1995), for instance, argued for the importance of psychological processes and phenomena during emergency evacuations. In two-thirds of the time needed to evacuate people in an emergency psychological processes play a pivotal role; only one third of this time is determined by parameters of a strictly technological nature. Sime (1995) argued that it takes time for people to discontinue their ongoing activities, and to seek contact with social others to reduce uncertainty. Whereas these findings were based on interviews with survivors of disasters a considerable time after they happened, collection of location data during emergency drills or

actual emergencies—i.e., in real life—and subsequent analyses with psyosphere may well help understand what happens when people need to vacate buildings. This better understanding can, in turn, inform more effective evacuation strategies.

Similarly, psyosphere may facilitate understanding of crowd dynamics. Location data extracted from camera footage or based on signals emanating from smart phones present in crowds may yield insights on crowd members' moods or the cohesion of groups within the crowd, as well as the way in which these change as a result of policing strategies or attempts at communication by organisers.

Finally, practitioners and scientists working in the field of environmental design may use the tool to study the influences of, for instance, sign-posting or building layout on way-finding; relatively subtle variations in behaviour may point to states as decisiveness, goal-directedness, and confusion. Similarly, the effect of nudges, for instance implemented in nightlife areas to reduce related noise and public urination (e.g., Bloeme et al., 2017), on visitors' behaviours and their states may be studied in more detail. Added into the bargain, using location data in such settings could eventually absolve researchers from parsing countless hours of video footage.

Once again, however, any such attempt at interpreting behaviour of individuals and groups requires tightly controlled experiments dedicated to establishing causal relationships between aspects of behaviour and state variables.

Furthermore, future research could extend the use of psyosphere by adding features, such as Kjargaard et al.'s (2013) time-lag method for detecting leadership and followership, or apply more complex methods, such as machine-learning classification. The data from studies, such as the current one might, for instance, be used to train an algorithm to establish links between aspects of movement or other behaviours and various psychological state and trait variables, such as having depression (Wolf et al., 2013; Huang et al., 2016; Saeb et al., 2016), or being a pickpocket (Bouma et al., 2014).

CONCLUSION

Hopefully, the findings presented here will encourage social scientists to use positioning technologies to study movement outside of a laboratory and in a real-world setting. Moreover, they show that psyosphere can prepare GPS data from psychological experiments for analysis with commonplace statistical methods, such as linear regression.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Behavioural, Management, and Social Sciences Ethics Committee of the University of Twente. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg. 2021.538529/full#supplementary-material

REFERENCES

Anderson, J. (1971). Space-time budgets and activity studies in urban geography and planning. *Environ. Plann. A* 3, 353–368. doi: 10.1068/a030353

Axhausen, K. W., Zimmermann, A., Schönfelder, S., Rindsfüser, G., and Haupt, T. (2002). Observing the rhythms of daily life: a six-week travel diary. *Transportation* 29, 95–124.

Barliya, A., Omlor, L., Giese, M. A., Berthoz, A., and Flash, T. (2012). Expression of emotion in the kinematics of locomotion. *Exp. Brain Res.* 225, 159–176. doi: 10.1007/s00221-012-3357-4

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68:1016. doi: 10.1016/j.jml.2012.11.001

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., and Von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. *PLoS Med.* 8:e1001083. 10.1371/journal.pmed.1001083

Bloeme, R., de Vries, P., Galetzka, M., and van Soomeren, P. (2017). "Persuasive technology against public nuisance – Public urination in the urban nightlife district," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10171, eds P. de Vries, H. Oinas-Kukkonen, L. Siemons, N. Beerlage-de Jong, and L. van Gemert-Pijnen (Cham: Springer), 187–198.

Bohte, W., and Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. *Transport. Res. Part C Emerg. Technol.* 17, 285–297. doi: 10.1016/j.trc.2008.11.004

Bolderdijk, J. W., Knockaert, J., Steg, E. M., and Verhoef, E. T. (2011). Effects of Pay-As-You-Drive vehicle insurance on young drivers' speed choice: results of a Dutch field experiment. *Accid. Anal. Prevent.* 43, 1181–1186. doi: 10.1016/j.aap.2010.12.032

Bouma, H., Baan, J., Burghouts, G. J., Eendebak, P. T., van Huis, J. R., Dijk, J., et al. (2014). "Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall," in *Proceedings of the Optics and Photonics*

for Counterterrorism, Crime Fighting, and Defence X; and Optical Materials and Biomaterials in Security and Defence Systems Technology, Vol. 9253, (Bellingham, WA: International Society for Optics and Photonics), 92530F.

Brady, A. T., and Walker, M. B. (1978). Interpersonal distance as a function of situationally induced anxiety. *Br. J. Clin. Psychol.* 17, 127–133. doi: 10.1111/j.2044-8260.1978.tb00254.x

Burgess, D., Owen, G., Rana, H., Zamboni, R., Kajzar, F., Szep, A. A., et al. (2014). Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall. *Paper Presented at the Optics and Photonics for Counterterrorism, Crime Fighting, and Defence X; and Optical Materials and Biomaterials in Security and Defence Systems Technology XI*, Amsterdam.

Byrne, A., and Byrne, D. (1993). The effect of exercise on depression, anxiety and other mood states: a review. *J. Psychos. Res.* 37, 565–574. doi: 10.1016/0022-3999(93)90050-p

Cabin, R. J., and Mitchell, R. J. (2000). To Bonferroni or not to Bonferroni: when and how are the questions. *Bull. Ecol. Soc. Am.* 81, 246–248.

Carlson, J. A., Jankowska, M. M., Meseck, K., Godbole, S., Natarajan, L., Raab, F., et al. (2014). Validity of PALMS GPS scoring of active and passive travel compared with SenseCam. *Med. Sci. Sports Exerc.* 47, 662–667. doi: 10.1249/MSS.0000000000000446

Chaix, B., Meline, J., Duncan, S., Merrien, C., Karusisi, N., Perchoux, C., et al. (2013). GPS tracking in neighborhood and health studies: a step forward for environmental exposure assessment, a step backward for causal inference? *Health Place* 21, 46–51. doi: 10.1016/j.healthplace.2013.01.003

Davies, G., Huck, J., Whyatt, J., Dixon, J., Hocking, B., Jarman, N., et al. (2017). "Belfast Mobility: extracting route information from GPS Tracks," in *Proceedings of the GIS Research UK 25th Annual Conference: GISRUK*, Manchester.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. (2003). Cues to deception. *Psychol. Bull.* 129:74.

Doherty, S. T., and Miller, E. J. (2000). A computerized household activity scheduling survey. *Transportation* 27, 75–97.

Draijer, G., Kalfs, N., and Perdok, J. (2000). Global positioning system as data collection method for travel research. *Transport. Res. Record J. Transport. Res. Board* 1719, 147–153. doi: 10.3141/1719-19

Ekman, P., Friesen, W. V., and O'sullivan, M. (1988). Smiles when lying. *J. Pers. Soc. Psychol.* 54:414. doi: 10.1037/0022-3514.54.3.414

Elgethun, K., Fenske, R. A., Yost, M. G., and Palcisko, G. J. (2002). Time-location analysis for exposure assessment studies of children using a novel global positioning system instrument. *Environ. Health Perspect.* 111, 115–122. doi: 10.1289/ehp.5350

Ettema, D., Timmermans, H., and van Veghel, L. (1996). *Effects of Data Collection Methods in Travel and Activity Research*. Eindhoven: European Institute of Retailing and Services Studies (EIRASS).

Feshbach, S., and Feshbach, N. (1963). Influence of the stimulus object upon the complimentary and supplementary projection of fear. *J. Abnorm. Soc. Psychol.* 66:498. doi: 10.1037/h0047016

Fillekes, M. P., Giannouli, E., Kim, E. K., Zijlstra, W., and Weibel, R. (2019). Towards a comprehensive set of GPS-based indicators reflecting the multidimensional nature of daily mobility for applications in health and aging research. *Int. J. Health Geograph.* 18:17. doi: 10.1186/s12942-019-0181-0

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Mananchira: Hafner Publishing Co.

Fjørtoft, I., Kristoffersen, B., and Sageie, J. (2009). Children in schoolyards: Tracking movement patterns and physical activity in schoolyards using global positioning system and heart rate monitoring. *Landsc. Urb. Plann.* 93, 210–217. doi: 10.1016/j.landurbplan.2009.07.008

Franěk, M., and Režný, L. (2017). The effect of priming with photographs of environmental settings on walking speed in an outdoor environment. *Front. Psychol.* 8:73.

Franěk, M., Režný, L., Šefara, D., and Cabal, J. (2018). Effect of traffic noise and relaxations sounds on pedestrian walking speed. *Int. J. Environ. Res. Public Health* 15:752. doi: 10.3390/ijerph15040752

Galbraith, N., Manktelow, K., and Morris, N. (2008). Subclinical delusional ideation and a self—reference bias in everyday reasoning. *Br. J. Psychol.* 99, 29–44. doi: 10.1348/000712607x204317

Geyer, K., Ellis, D. A., and Piwek, L. (2019). A simple location-tracking app for psychological research. *Behav. Res. Methods* 51, 2840–2846. doi: 10.3758/s13428-018-1164-y

Golledge, R. G. (1997). *Spatial Behavior: A Geographic Perspective*. New York, NY: Guilford Press.

Goodchild, M. F., and Janelle, D. G. (1984). The city around the clock: Space—time patterns of urban ecological structure. *Environ. Plann. A* 16, 807–820. doi: 10.1068/a160807

Google LLC (2018). *Google Maps*. Mountain View, CA: Google LLC.

Govern, J. M., and Marsch, L. A. (2001). Development and validation of the situational self-awareness scale. *Conscious. Cogn.* 10, 366–378. doi: 10.1006/ccog.2001.0506

Gross, M. M., Crane, E. A., and Fredrickson, B. L. (2012). Effort-shape and kinematic assessment of bodily expression of emotion during gait. *Hum. Mov. Sci.* 31, 202–221. doi: 10.1016/j.humov.2011.05.001

Hijmans, R. J., Williams, E., and Vennes, C. (2015). *Geosphere: Spherical Trigonometry*. R Package version 1.5-7. Available online at: http://CRAN.R-project.org/package=geosphere (accessed April 28, 2021).

Hirschfeld, G., and Von Brachel, R. (2014). Improving Multiple-Group confirmatory factor analysis in R–A tutorial in measurement invariance with continuous and ordinal indicators. *Pract. Assess. Res. Eval.* 19:7.

Hoaglin, D. C. (2003). John W. Tukey and data analysis. *Statist. Sci.* 18, 311–318.

Hofmann-Wellenhof, B., Lichtenegger, H., and Wasle, E. (2007). *GNSS–Global Navigation Satellite Systems: GPS, GLONASS, Galileo, and More*. Berlin: Springer Science & Business Media.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* 79, 65–70.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* 30, 179–185. doi: 10.1007/bf02289447

Huang, Y., Xiong, H., Leach, K., Zhang, Y., Chow, P., Fua, K., et al. (2016). "Assessing social anxiety using GPS trajectories and point-of-interest data," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, New York, NY, 898–903.

Isaacson, M., D'Ambrosio, L., Samanta, T., and Coughlin, J. (2015). Life-stage and mobility: an exploratory GPS study of mobility in multigenerational families, Ahmedabad, India. *J. Aging Soc. Policy* 27, 348–363. doi: 10.1080/08959420.2015.1058123

Isaacson, M., Shoval, N., Wahl, H.-W., Oswald, F., and Auslander, G. (2016). Compliance and data quality in GPS-based studies. *Transportation* 43, 25–36. doi: 10.1007/s11116-014-9560-3

Isella, L., Romano, M., Barrat, A., Cattuto, C., Colizza, V., Van den Broeck, W., et al. (2011a). Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors. *PLoS One* 6:e17144. doi: 10.1371/journal.pone.0017144

Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J.-F., and Van den Broeck, W. (2011b). What's in a crowd? Analysis of face-to-face behavioral networks. *J. Theoret. Biol.* 271, 166–180. doi: 10.1016/j.jtbi.2010.11.033

James, P., Jankowska, M., Marx, C., Hart, J. E., Berrigan, D., Kerr, J., et al. (2016). "Spatial Energetics": integrating data from GPS, accelerometry, and GIS to address obesity and inactivity. *Am. J. Prevent. Med.* 51, 792–800. doi: 10.1016/j.amepre.2016.06.006

Janelle, D. G., Goodchild, M. F., and Klinkenberg, B. (1988). Space-time diaries and travel characteristics for different levels of respondent aggregation. *Environ. Plann. A* 20, 891–906. doi: 10.1068/a200891

Jankowska, M. M., Schipperijn, J., and Kerr, J. (2015). A framework for using GPS data in physical activity and sedentary behavior studies. *Exerc. Sport Sci. Rev.* 43, 48–56. doi: 10.1249/JES.0000000000000035

Jian, J. Y., Matsuka, T., and Nickerson, J. V. (2006). "Recognizing deception in trajectories," in *Proceedings of the Cognitive Science Society*, Hoboken, NJ, 1563–1568.

Kahle, D., and Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *R J.* 5:544.

Kaiser, H. F., and Rice, J. (1974). Little jiffy, mark IV. *Educ. Psychol. Measur.* 34, 111–117. doi: 10.1177/001316447403400115

Kjargaard, M. B., Blunck, H., Wüstenberg, M., Grønbask, K., Wirz, M., Roggen, D., et al. (2013). "Time-lag method for detecting following and leadership behavior of pedestrians from mobile sensing data," in *Proceedings of the 2013*

IEEE International Conference on Pervasive Computing and Communications (PerCom), (New Tork, NY: IEEE), 56–64.

Koller, C. I., Wetter, O. E., and Hofer, F. (2016). 'Who's the Thief?' The influence of knowledge and experience on early detection of criminal intentions. *Appl. Cogn. Psychol.* 30, 178–187. 10.1002/acp.3175

Krenn, P. J., Titze, S., Oja, P., Jones, A., and Ogilvie, D. (2011). Use of global positioning systems to study physical activity and the environment: a systematic review. *Am. J. Prevent. Med.* 41, 508–515. doi: 10.1016/j.amepre.2011.06.046

Krieglmeyer, R., De Houwer, J., and Deutsch, R. (2013). On the nature of automatically triggered approach–avoidance behavior. *Emot. Rev.* 5, 280–284. doi: 10.1177/1754073913477501

Loecher, M., and Ropkins, K. (2015). RgoogleMaps and loa: unleashing R graphics power on map tiles. *J. Statist. Softw.* 63:2488.

Madan, A., Moturu, S. T., Lazer, D., and Pentland, A. (2010). Social sensing: obesity, unhealthy eating and exercise in face-to-face networks. *Wirel. Health* 2010, 104–110.

Maddison, R., and Mhurchu, C. N. (2009). Global positioning system: a new opportunity in physical activity measurement. *Int. J. Behav. Nutr. Phys. Activ.* 6:73. doi: 10.1186/1479-5868-6-73

Michalak, J., Troje, N. F., Fischer, J., Vollmar, P., Heidenreich, T., and Schulte, D. (2009). Embodiment of sadness and depression—gait patterns associated with dysphoric mood. *Psychosom. Med.* 71, 580–587. doi: 10.1097/psy. 0b013e3181a2515c

Moussaïd, M., Helbing, D., and Theraulaz, G. (2011). How simple rules determine pedestrian behavior and crowd disasters. *Proc. Natl. Acad. Sci. U.S.A.* 108, 6884–6888. doi: 10.1073/pnas.1016507108

Muenchen, R. A. (2012). *The Popularity of Data Analysis Software*. Available online at: http://r4stats.com/popularity (accessed April 28, 2021).

Murakami, E., and Wagner, D. P. (1999). Can using global positioning system (GPS) improve trip reporting? *Transport. Res. Part C Emerg. Technol.* 7, 149–165. doi: 10.1016/s0968-090x(99)00017-0

Necula, E. (2015). Analyzing traffic patterns on street segments based on GPS data using R. *Transport. Res. Proc.* 10, 276–285. doi: 10.1016/j.trpro.2015.09.077

Palmius, N., Tsanas, A., Saunders, K. E. A., Bilderbeck, A. C., Geddes, J. R., Goodwin, G. M., et al. (2017). Detecting bipolar depression from geographic location data. *IEEE Transact. Biomed. Eng.* 64, 1761–1771. doi: 10.1109/TBME. 2016.2611862

Phillips, M. L., Hall, T. A., Esmen, N. A., Lynch, R., and Johnson, D. L. (2001). Use of global positioning system technology to track subject's location during environmental exposure sampling. *J. Exp. Anal. Environ. Epidemiol.* 11, 207–215. doi: 10.1038/sj.jea.7500161

Saeb, S., Lattie, E. G., Schueller, S. M., Kording, K. P., and Mohr, D. C. (2016). The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ* 4:e2537. doi: 10.7717/peerj.2537

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., et al. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J. Med. Int. Res.* 17:e175. doi: 10.2196/jmir. 4273

Satchell, L., Morris, P., Mills, C., O'Reilly, L., Marshman, P., and Akehurst, L. (2017). Evidence of big five and aggressive personalities in gait biomechanics. *J. Nonverb. Behav.* 41, 35–44. doi: 10.1007/s10919-016-0240-1

Schachter, S. (1959). *The Psychology of Affiliation: Experimental Studies of the Sources of Gregariousness*. Palo Alto, CA: Stanford University Press.

Schuessler, N., and Axhausen, K. (2009). Processing raw data from global positioning systems without additional information. *Transport. Res. Rec. J. Transport. Res. Board* 2105, 28–36. doi: 10.3141/2105-04

Schwerdtfeger, A., Eberhardt, R., Chmitorz, A., and Schaller, E. (2010). Momentary affect predicts bodily movement in daily life: an ambulatory monitoring study. *J. Sport Exerc. Psychol.* 32, 674–693. doi: 10.1123/jsep.32.5.674

Sevtsuk, A. (2009). *Mapping the MIT Campus in Real Time Using WiFi. Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*. Hershey, PA: IGI Global, 326–338.

Shoval, N., Auslander, G. K., Freytag, T., Landau, R., Oswald, F., Seidl, U., et al. (2008). The use of advanced tracking technologies for the analysis of mobility in Alzheimer's disease and related cognitive diseases. *BMC Geriatr.* 8:1.

Shoval, N., Auslander, G., Cohen-Shalom, K., Isaacson, M., Landau, R., and Heinik, J. (2010). What can we learn about the mobility of the elderly in the GPS era? *J. Transp. Geogr.* 18, 603–612. doi: 10.1016/j.jtrangeo.2010.03.012

Shoval, N., Wahl, H.-W., Auslander, G., Isaacson, M., Oswald, F., Edry, T., et al. (2011). Use of the global positioning system to measure the out-of-home mobility of older adults with differing cognitive functioning. *Ageing Soc.* 31, 849–869. doi: 10.1017/s0144686x10001455

Sime, J. D. (1995). Crowd psychology and engineering. *Saf. Sci.* 21, 1–14. doi: 10.1016/0925-7535(96)81011-3

Stekkinger, M. R. (2012). *Can Hostile Intent be Detected by Means of Signaling?* Master's thesis. Enschede: University of Twente.

Stopher, P. R. (1992). Use of an activity-based diary to collect household travel data. *Transportation* 19, 159–176. doi: 10.1007/bf02132836

Stopher, P. R., Bullock, P., and Horst, F. (2002). Exploring the use of passive GPS devices to measure travel. *Appl. Adv. Technol. Transport.* 2002, 959–967.

Van der Zee, S., Poppe, R., Taylor, P. J., and Anderson, R. (2019). To freeze or not to freeze: a culture-sensitive motion capture approach to detecting deceit. *PLoS One* 14:e0215000. doi: 10.1371/journal.pone.0215000

Vazquez-Prokopec, G. M., Bisanzio, D., Stoddard, S. T., Paz-Soldan, V., Morrison, A. C., Elder, J. P., et al. (2013). Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. *PLoS One* 8:e58802. doi: 10.1371/journal.pone.0058802

Vazquez-Prokopec, G. M., Stoddard, S. T., Paz-Soldan, V., Morrison, A. C., Elder, J. P., Kochel, T. J., et al. (2009). Usefulness of commercially available GPS data-loggers for tracking human movement and exposure to dengue virus. *Int. J. Health Geograph.* 8:1. doi: 10.1186/1476-072X-8-68

Vrij, A., Semin, G. R., and Bull, R. (1996). Insight into behavior displayed during deception. *Hum. Commun. Res.* 22, 544–562. doi: 10.1111/j.1468-2958.1996. tb00378.x

Whyatt, J. D., Huck, J. J., Davies, G., Dixon, J., Hocking, B., Jarman, N., et al. (2017). *Belfast Mobility Project: Integrating PGIS and GPS to Understand Patterns of Segregation*. Nottingham: GISRUK.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Berlin: Springer.

Wigboldus, D. H., and Dotsch, R. (2016). Encourage playing with data and discourage questionable reporting practices. *Psychometrika* 81, 27–32. doi: 10. 1007/s11336-015-9445-1

Wijn, R., Kleij, R., Kallen, V., Stekkinger, M., and de Vries, P. W. (2017). Telling friend from foe: environmental cues improve detection accuracy of individuals with hostile intentions. *Legal Criminol. Psychol.* 22, 378–399. doi: 10.1111/lcrp. 12107

Wolf, J. (2006). "Applications of new technologies in travel surveys," in *Travel Survey Methods: Quality and Future Directions*, eds P. Stopher and C. Stecher (Bingley: Emerald Publishing), 531–544. doi: 10.1108/9780080464015-029

Wolf, J. L. (2000). *Using GPS Data Loggers To Replace Travel Diaries In The Collection of Travel Data*. Doctoral dissertation. Atlanta, GA: School of Civil and Environmental Engineering, Georgia Institute of Technology.

Wolf, J., Schönfelder, S., Samaga, U., Oliveira, M., and Axhausen, K. (2004). Eighty weeks of global positioning system traces: approaches to enriching trip information. *Transport. Res. Rec. J. Transport. Res. Board* 1870, 46–54. doi: 10.3141/1870-06

Wolf, P. S. A., Figueredo, A. J., and Jacobs, W. J. (2013). Global positioning system technology (GPS) for psychological research: a test of convergent and nomological validity. *Front. Psychol.* 4:315. doi: 10.3389/fpsyg.2013.00315

Xia, J. C., Arrowsmith, C., Jackson, M., and Cartwright, W. (2008). The wayfinding process relationships between decision-making and landmark utility. *Tour. Manag.* 29, 445–457. doi: 10.1016/j.tourman.2007.05.010

Ziepert, B., Ufkes, E. G., and de Vries, P. W. (2018). *psyosphere: Analyse GPS Data*. Available online at: https://cran.r-project.org/package=psyosphere (accessed April 28, 2021).

Validation of New Methods of Using Simulated Whole-Body Movements as Implicit Indicators of Sound and Odor Preferences

Donato Cereghetti [1,2], Pauline Faye [2], Laetitia Gros [2,3], Lucas Mahé [2,4], Emmanuelle Diaz [2,5], Isabelle Cayeux [1,2], Théo Heritier [2,4,6] and Rémy Versace [2,4]*

[1] Firmenich SA, Geneva, Switzerland, [2] Le Sensolier, Paris, France, [3] Orange Labs, Lannion, France, [4] EMC Laboratory, Institute of Psychology, Lyon2 University, Lyon, France, [5] PSA Groupe, Vélizy-Villacoublay, France, [6] Silliker SAS, Mérieux NutriSciences, Cergy-Pontoise, France

Would you get close to a stinky perfume bottle or to a loudspeaker producing noise? In this paper, we present two procedures that allowed us to assess the ability of auditory and olfactory cues to elicit automatic approach/avoidance reactions toward their sources. The procedures resulted from an adaptation of the Visual Approach/Avoidance by the Self Task (VAAST; Rougier et al., 2018), a task having the peculiarity of simulating approach/avoidance reactions by using visual feedback coming from the whole-body movements. In the auditory VAAST (Experiment 1), participants were instructed to move forward or backward from a loudspeaker that produced spoken words differentiated by their level of distortion and thus by their hedonic value. In the olfactory VAAST (Experiment 2), participants were asked to move forward or backward from a perfume bottle that delivered pleasant and unpleasant odors. We expected, consistent with the approach/avoidance compatibility effect, shorter latencies for approaching positive stimuli and avoiding negative stimuli. In both experiments, we found an effect of the quality of the emotional stimulus on forward actions of participants, with undistorted words and pleasant odors inducing faster forward movements compared with that for distorted words and unpleasant odors. Notably, our results further suggest that the VAAST can successfully be used with implicit instructions, i.e., without requiring participants to explicitly process the valence of the emotional stimulus (in Experiment 1) or even the emotional stimulus itself (in Experiment 2). The sensitivity of our procedures is analyzed and its potential in cross-modal and (contextualized) consumer research discussed.

Keywords: consumer, preferences, motivation, approach/avoidance, implicit measures, sounds, odors

INTRODUCTION

How can the hedonic value of a product be measured alone, without being simultaneously contaminated by something else? Most often, consumer judgments are collected by using direct measures (e.g., rating scales, questionnaires, or semi-directive interviews). Although such methods proved their ability to discriminate products, they are also subject to several biases. First, consumer judgments are likely to be affected by a social desirability bias (Edwards, 1953) and may rely on

individual introspective skills (Nisbett and Wilson, 1977). In addition, since the expression of a judgment requires the translation of a perceptual feeling into an explicit response, it is difficult to dissociate, in consumer judgment, the variability due to the perceptual feeling from the variability due to its explicit translation. Finally, many biases related to the way individuals use scales when quantifying their judgment have been reported, among which we mention contraction biases, centering biases, and logarithmic biases (for a review, see Poulton, 1989). To avoid such biases, it is necessary to use indirect measures that are likely to reflect the hedonic value of the product.

The hedonic value refers to the pleasure provided by the confrontation with a stimulus; it is therefore intrinsically linked to the emotions elicited in the individual by the stimulus. Etymologically, the term "emotion" refers to a setting in motion, a change from an initial state. Such changes are manifested at different levels: at the physiological level (brain electrophysiological responses, cardiovascular, respiratory, skin conductance responses, etc.) and at the behavioral level (motor reactions, approach/avoidance reactions, etc.). It is for this reason that many studies have used these changes as implicit indicators of the emotion felt; the underlying idea is that the more intense the emotion, the more important these changes must be.

From an adaptive point of view, it is indeed obvious that one of the primary and most important behavioral responses toward a stimulus is whether it should be approached or avoided. Approach/avoidance procedures were developed to investigate this primary behavioral response; the basic assumption is that positive stimuli elicit approach reactions, whereas negative stimuli elicit avoidance reactions. Approach/avoidance tendencies are typically explored in tasks that require arm and/or hand movements in response to emotional stimuli. In these tasks, reaction times (RTs) are usually recorded, and participants are expected to produce faster responses to approach positive stimuli and to avoid negative stimuli compared with the reverse, which is known as the approach/avoidance compatibility effect. Approach/avoidance tendencies have been investigated by using a large range of different apparatuses and procedures, including the lever or joystick task (e.g., Chen and Bargh, 1999), the modified keyboard task (e.g., Alexopoulos and Ric, 2007), the button stand task (e.g., Rotteveel and Phaf, 2004), and the zoom-feedback joystick task (e.g., Rinck and Becker, 2007). Approach/avoidance tendencies have, moreover, been explored by using different types of stimuli, including words (e.g., Solarz, 1960; Markman and Brendl, 2005; Alexopoulos and Ric, 2007), pseudo-words (e.g., Carr et al., 2016), pictures (e.g., Rinck and Becker, 2007; Saraiva et al., 2013), and faces (e.g., Paulus and Wentura, 2016).

The approach/avoidance compatibility effect was first identified in a study performed by Chen and Bargh (1999), in which participants were required to perform arm flexions and extensions by pulling/pushing a lever in response to emotional words. Their results showed that participants provided faster responses to positive words when pulling a lever toward them and to negative words when pushing the lever away (as compared with the reverse; Experiment 1) and that this occurred even when the task did not require participants to explicitly evaluate the

valence of the emotional stimulus (Experiment 2). From these findings, Chen and Bargh (1999) postulated the existence of an automatic, unconscious link between the stimulus evaluation and specific motor responses, with positive stimuli activating arm flexion (approach reaction; i.e., move the stimuli toward the self) and negative stimuli activating arm extension (avoidance reaction; i.e., move the stimuli away from the self). Although a number of studies identified the compatibility effect described in Chen and Bargh (1999), others failed to reproduce it or found the opposite effect, challenging *de facto* the hypothesis that a direct, hard-wired link between evaluation and behavior may exist (Wentura et al., 2000; Markman and Brendl, 2005; Lavender and Hommel, 2007; Rinck and Becker, 2007; Eder and Rothermund, 2008; Paladino and Castelli, 2008; Seibt et al., 2008; Krieglmeyer et al., 2010; Rotteveel et al., 2015). Indeed, arm flexion can be interpreted either as an approach (bringing a positive stimulus, e.g., a pizza, closer to the self) or as avoidance (withdrawing the hand from a negative stimulus, e.g., a spider), and arm extension can be interpreted either as approach (reaching a positive stimulus with the hand, e.g., a pizza) or as avoidance (pushing a negative stimulus, e.g., a spider, away from the self); this depends on contextual factors, on individual's goals (e.g., Bamford and Ward, 2008), and on the frame of reference (self-related vs. object-related; e.g., Seibt et al., 2008).

Not only are results across studies relatively inconsistent, but the compatibility effect sizes reported in the literature are also generally low to moderate (Phaf et al., 2014), suggesting the need to develop new procedures. In that spirit, and based on an embodied approach to cognition (e.g., Niedenthal, 2007; van Dantzig et al., 2009; Versace et al., 2014), Rougier et al. (2018) suggested that a sensorimotor task that simulates the visual information coming from the whole-body movements should produce stronger and more replicable effects. They then developed a new procedure, the Visual Approach/Avoidance by the Self Task (VAAST). This procedure no longer relies on sensorimotor indices limited to arm and hand movements, but rather on sensorimotor indices provided by a realistic visual flow that simulates the whole-body movements in relation to the emotional stimulus. Rougier et al. (2018) evaluated the sensitivity of this measure in a series of six experiments, in which participants had to approach or avoid emotional words as quickly and accurately as possible by using a button box. Notably, in each trial, participants were required to press the corresponding response key four times consecutively in order to complete a single forward or backward movement. After each key press, the whole visual scene (i.e., the surrounding environment and the target word) was zoomed in or out by 10%, giving the visual impression of walking forward or backward to the target as a consequence of their response. In their experiments, Rougier et al. (2018) showed that the VAAST can produce large and replicable compatibility effects.

Despite these promising results, the VAAST should be considered a relatively new procedure, and there are many opportunities for further developments and improvements. We identified three of them. First, previous research has focused only on the visual domain. We believe that this procedure has great potential for cross-modal research and thus deserve to

be further investigated by considering other sensory modalities. Second, in the original VAAST (Rougier et al., 2018), participants were asked to move forward or backward from words, which should be considered a non-ecological task. We believe that the VAAST procedure has the potential to be further applied in more (simulated) ecological contexts by using virtual 3D objects located in the visual scene as target stimuli instead of words. Third, although the VAAST has been largely validated by using explicit instructions, the use of implicit instructions still remains underinvestigated. Indeed, in five of the six experiments presented in Rougier et al. (2018), the task required participants to explicitly process the valence of the emotional stimuli (i.e., participants were asked to categorize words as being positive or negative by pressing two response keys). Only in Experiment 5 were participants not required to attentively attend to the stimulus valence. In this experiment, participants were asked to determine whether a sequence of letters (e.g., "nlkjdsOaq") contained a capital letter or not, while the emotional words were primed for a very short duration (30 ms) with a pre-mask (the sequence "WXWXWXWXW" displayed for 50 ms) and a post-mask (the sequence "@W@W@W@W" displayed for 50 ms).

The purpose of our study was therefore threefold. First, we aimed to extend the VAAST to other sensory modalities by using spoken words (section Experiment 1) and odors (section Experiment 2) as emotional stimuli. That is, we aimed to measure the ability of auditory and olfactory cues to elicit automatic reactions toward their source (i.e., a loudspeaker in Experiment 1 and a perfume bottle in Experiment 2). Notably, the sources were situated in a congruent surrounding environment (i.e., a living room in Experiment 1 and a bathroom in Experiment 2), thus ensuring the ecological validity of the task. Note that in both studies, participants were explicitly informed that auditory and olfactory cues "came from" the loudspeaker and from the perfume bottle, respectively, thus reinforcing the link between the cues and their source. Second, we aimed to provide further evidence concerning the ability of VAAST to measure approach/avoidance reactions for stimuli that differ in their hedonic value. In Experiment 1, spoken words were presented at different levels of distortion (i.e., undistorted, slightly distorted, and moderately distorted); we assume that the more distorted the signal, the more unpleasant the stimulus. In Experiment 2, we used odors that strongly differ in their liking value. We expected, consistent with the approach/avoidance compatibility effect, positive cues to promote an approach reaction and negative cues to promote an avoidance reaction. That is, we expected faster RTs in approaching positive stimuli and avoiding negative stimuli than the reverse. Third, we aimed to show that approach/avoidance compatibility effects can be observed without requiring participants to explicitly process the valence of the stimuli, even, in Experiment 2, without having to explicitly process the stimuli that induce emotions. In Experiment 1, participants had to indicate whether the spoken word represented a living being or an inanimate object, regardless of the level of distortion. In Experiment 2, participants had to judge whatever the perfume bottle was bent toward to the right or to the left; thus, they no longer had to explicitly deal with the stimuli that induce emotions.

EXPERIMENT 1

Materials and Methods

Participants

Twenty-six students from the University of Lyon (13 females, 11 males, two no answers) voluntarily participated for course credits. The average age of the panel was 21.8 years ($SD = 3.0$). The experiment was conducted in accordance with the ethical principles stated in the Declaration of Helsinki. All participants provided written informed consent and reported having a normal or corrected vision and normal hearing. A participant was excluded because of technical issues. The analysis was thus performed on the remaining 25 participants. To the best of our knowledge, this is the first study aiming to adapt the VAAST to auditory stimuli. A priori determination of sample size was therefore not possible. Consequently, we conducted a sensitivity power analysis using G*Power 3 (Faul et al., 2007) to identify the smallest effect size that our sample would be able to detect in paired t-tests performed in RT analysis (section A-VAAST: Reaction Times). The analysis revealed that our sample size ($n = 25$) could reliably detect the effect sizes of Cohen's $d_z = 0.51$ (one-tailed), assuming $\alpha = 0.05$ and power ($1 - \beta$) = 0.80.

Materials

Visual Scene

A virtual visual scene was designed for the Auditory VAAST (A-VAAST) with 3D Blender. The visual scene consisted of a living room presented in the first-person view (**Figure 1**). A loudspeaker (i.e., the source of auditory cues) was placed in the center of the visual scene, just above a media cabinet.

Auditory Stimuli

Ten neutral words were used in the A-VAAST. Five words represented living beings ("pélican," "renard," "cycliste," "tortue," "fourmi"; i.e., "pelican," "fox," "cyclist," "turtle," "ant") and the other five words represented inanimate objects ("tabouret," "bouée," "marmite," "volant," "cloche"; i.e., "stool," "buoy," "stockpot," "steering wheel," "bell"). The words were selected on the basis of a published normative study (Bonin et al., 2003) where participants ($n = 97$) were asked to evaluate the

FIGURE 1 | Visual scene of the Auditory Visual Approach/Avoidance by the Self Task (A-VAAST).

concreteness, imageability, subjective frequency, and emotional valence of 866 French words with a 5-point scale. In particular, we selected ten neutral (valence: $M = 3.1$, $SD = 0.1$) and medium-frequent words (frequency: $M = 2.7$, $SD = 0.6$) while controlling the number of syllables across the two-word categories. In addition, each word started with a different phoneme to avoid the implicit statistical learning of participants. The words uttered by a female speaker were recorded with a Zoom H2 high-definition microphone and normalized in level with an active speech level equalization at -26 dBov, using the implementation of the ITU-T Recommendation P.56 (2011) given in the ITU-T Software Tool Library (ITU-T Recommendation G.191, 2019; see program "sv56demo"). Spoken words were then degraded in quality with noise by using the modulated noise reference unit (ITU-T Recommendation P810, 1996) at a level Q of 22 (slightly distorted) and 15 dB (moderately distorted), with Q being the ratio of speech power to modulated noise power.

Procedure

Participants were seated on a chair in front of a computer screen (with a refresh rate of 60 Hz) and were wearing headphones (Sennheiser HD206). Each participant's head position was fixed with the aid of a chin rest such that the distance between the eyes and the screen was 50 cm. Participants were told that they would be immersed in a virtual environment, i.e., a living room equipped with a loudspeaker emitting spoken words. Participants were then instructed to move forward or backward from the loudspeaker as quickly and accurately as possible, depending on the category of the spoken word (i.e., living being, inanimate object). The correspondence between word category and the approach/avoidance action was counterbalanced across participants: Half of the participants had to approach the loudspeaker when the word belonged to the living being category and avoid it when the word belonged to the inanimate object category, whereas the remaining half had to approach the loudspeaker when the word belonged to the inanimate object category and avoid it when the word belonged to the living being category.

At beginning of each trial, the visual scene was presented full screen and participants had to press the "start" key (key 5 of the numeric keypad) to start the trial. Five hundred milliseconds later, a fixation point appeared in the center of the loudspeaker for 500 ms, after which a spoken word was presented on the headphones at a given level of sound distortion (undistorted, slightly distorted, or moderately distorted). Depending on the spoken word category, participants had to press key 8 of the numeric keypad three times consecutively to approach the loudspeaker and key 2 of the numeric keypad to move away from it. After each correct key-press response, the whole visual scene was zoomed in or out by 10%, giving the visual impression of walking forward or backward as a consequence of the approach/avoidance action of participants (Rougier et al., 2018). Following an incorrect response, a red capital "X" was presented for 100 ms, together with an error beep sound, and participants were required to provide the correct response to trigger the visual flow. The trial terminated after three key presses in the same direction (i.e., a complete forward or

backward movement) and was followed by a black background that masked the visual scene for 1,000 ms, after which the next trial started. Although participants had to press the relevant key three times consecutively to approach or avoid the loudspeaker, we considered accuracy and RT only for the first action of the participants (see Rougier et al., 2018).

Designed with OpenSesame (Mathôt et al., 2012), the A-VAAST consisted of two phases, namely, practice and test. The training phase was composed of 30 trials in which each word was presented in a randomized order at each level of distortion (i.e., undistorted, slightly distorted, and moderately distorted). In the testing phase, the same 30-item cycle was repeated four times, resulting in a total of 120 trials presented in a randomized order. Participants were thus exposed to 20 trials per each experimental condition level (action × sound quality), thus resulting in 60 "move forward" and 60 "move backward" trials.

Results

Experimental data and analysis scripts are available on the OpenScience Framework platform at https://osf.io/jkr7m/

A-VAAST: Accuracy

Accuracy of the participants on the A-VAAST across the six conditions is summarized in **Table 1**. We performed a repeated-measures ANOVA with the factors sound quality (undistorted, slightly distorted, and moderately distorted) and action ("move forward," "move backward") as within-subject factors for the response accuracy of participants. According to the results, accuracy was not influenced by sound quality [$F_{(1.74, 41.71)} = 1.16$, $p = 0.32$, partial eta-squared $= 0.05$, 90% CI (0.00, 0.15)], or by action [$F_{(1, 24)} = 0.01$, $p = 0.91$, $\eta_p^2 < 0.001$, 90% CI (0.00, 0.04)], or by critical sound quality–action interaction [$F_{(1.89, 45.30)} = 0.73$, $p = 0.48$, $\eta_p^2 = 0.03$, 90% CI (0.00, 0.11)]. The very high percentage of correct responses suggests that the degradation did not affect understanding of the words.

A-VAAST: Reaction Times

We first excluded incorrect trials (2.3%), as well as trials with RTs faster than 300 ms (no trials removed) and slower than 1,500 ms (5.0%). For the remaining correct trials, RTs falling outside 2.5 SD from the mean of each participant computed for each experimental condition level (action × sound quality) were also considered outliers and excluded from further analysis (0.9%). We then performed a repeated-measures ANOVA with the factors sound quality (undistorted, slightly distorted,
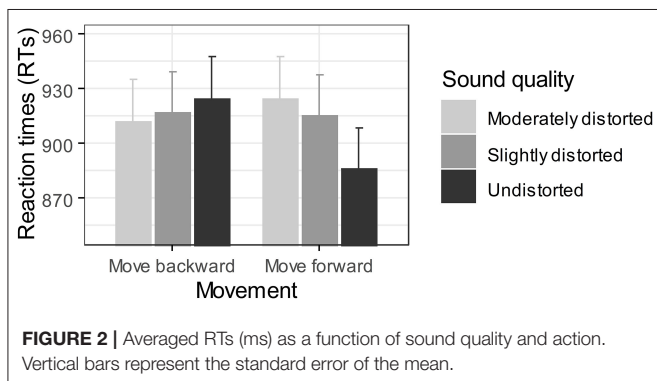
TABLE 1 | Auditory VAAST: accuracy of participants.

Action	Sound Quality	Accuracy ($M \pm SD$)
Move backward	Undistorted	98.0 ± 4.6%
Move backward	Slightly distorted	98.0 ± 3.5%
Move backward	Moderately distorted	97.0 ± 6.0%
Move forward	Undistorted	97.4 ± 4.8%
Move forward	Slightly distorted	98.2 ± 4.3%
Move forward	Moderately distorted	97.8 ± 5.4%

and moderately distorted) and action ("move forward," "move backward") as within-subject factors for RTs of the remaining participants. Analyses neither revealed a significant main effect of the sound quality [$F_{(1.64, 39.39)} = 1.15$, $p = 0.32$, $\eta_p^2 = 0.05$, 90% CI (0.00, 0.16)] nor of the type of action [$F_{(1, 24)} = 0.50$, $p = 0.49$, $\eta_p^2 = 0.02$, 90% CI (0.00, 0.17)]. However, as expected, the critical interaction between the factors sound quality and action was significant [$F_{(1.79, 43.05)} = 5.88$, $p = 0.007$, $\eta_p^2 = 0.20$, 90% CI (0.04, 0.34)] (**Figure 2**), suggesting that the effect of sound distortion on RTs depended on the action type of participants. To characterize this interaction, we performed two repeated-measures ANOVAs to evaluate the effect of sound quality on RTs separately for the two types of action ("move forward" and "move backward"). The sound quality effect was significant on the "move forward" [$F_{(1.79, 43.01)} = 7.02$, $p = 0.003$, $\eta_p^2 = 0.23$, 90% CI (0.06, 0.37)] but not on the "move backward" [$F_{(1.64, 39.33)} = 0.45$, $p = 0.60$, $\eta_p^2 = 0.02$, 90% CI (0.00, 0.10)] condition. Further, one-tailed Bonferroni corrected paired t-tests revealed that participants moved forward faster for undistorted words (886 ± 115 ms) than they did for either moderately distorted words (925 ± 117 ms) [$t(24) = 4.32$, $p < 0.001$, $d_z = 0.86$, 95% CI (0.40, 1.32)] or slightly distorted words (915 ± 113 ms) [$t(24) = 2.69$, $p = 0.02$, $d_z = 0.54$, 95% CI (0.11, 0.95)]. No difference was observed between moderately and slightly distorted words [$t(24) = 0.77$, $p = 0.67$, $d_z = 0.15$, 95% CI (−0.24, 0.55)].

Discussion Experiment 1

In this experiment, participants were instructed to move forward and backward from a loudspeaker as quickly as possible, depending on the spoken word category (i.e., living being, inanimate object). Words were presented at different levels of distortion (i.e., undistorted, slightly distorted, and moderately distorted), which allowed us to measure the impact of sound quality on the approach/avoidance reactions of participants. As expected, we observed a critical interaction between sound quality and action on RTs of participants (i.e., the approach/avoidance compatibility effect). Further analysis revealed that sound quality only influenced forward (not backward) movements. In particular, undistorted words induced faster forward movements compared with those of moderately and slightly distorted words, suggesting that this effect was principally driven by undistorted words.

FIGURE 2 | Averaged RTs (ms) as a function of sound quality and action. Vertical bars represent the standard error of the mean.

EXPERIMENT 2

Materials and Methods
Participants

A total of 26 panelists (18 females and eight males; French and/or Swiss citizens) recruited from all departments of Firmenich SA participated in this experiment. The average age of the panel was 45.1 years ($SD = 10.7$). It was conducted in accordance with the ethical principles stated in the Declaration of Helsinki issued by the World Medical Association. All participants provided written informed consent and were free to withdraw at any time without giving any reason. At the end of the panel, participants received snacks as a gesture of appreciation. To the best of our knowledge, this is the first study aiming to adapt the VAAST to olfactory stimuli. A priori determination of sample size was therefore not possible. Consequently, we conducted a sensitivity power analysis using G*Power 3 (Faul et al., 2007) to identify the smallest effect size that our sample would be able to detect in paired t-tests performed in RT analysis (section O-VAAST: Reaction times). The analysis revealed that our sample size ($n = 26$) could reliably detect effect sizes of Cohen's $d_z = 0.50$ (one-tailed), assuming $\alpha = 0.05$ and power $(1 - \beta) = 0.80$.

Materials
Visual Scene

A virtual visual scene was designed for the Olfactory VAAST (O-VAAST) with 3D Blender. The visual scene consisted of a bathroom presented in the first-person view (**Figure 3A**). In the center of the visual scene, a shelf was displayed, above which was positioned a perfume bottle (i.e., the source of olfactory cues). The perfume bottle could be presented in the vertical position (neutral stimulus, **Figure 3B**) or tilted to the left or right (visual targets of the O-VAAST; **Figures 3C,D**, respectively).

Olfactory Stimuli

We performed a preliminary study to select the odorants for the O-VAAST and identify suitable concentrations. Using visual analog scales (ranging from 0 to 100), 26 internal panelists (mean age = 41.9, $SD = 12.6$ years; 17 females and nine males; French and/or Swiss citizens) were asked to evaluate the perceived pleasantness (from "very unpleasant" to "very pleasant"), intensity (from "not perceived" to "very strong"), and familiarity (from "not familiar at all" to "very familiar") of pleasant (Perfume 1, Perfume 2) and unpleasant [butyric acid, synthetic body odor (SBO)] odors diluted at different concentrations in dipropylene glycol (DIPG) or isopropyl myristate (IPM). Perfume 1 was a commercial deodorant for women described by Firmenich perfumers as "floral," "green," "geranium," "citrus," and "aromatic." Perfume 2 was a commercial deodorant for men described by Firmenich perfumers as "floral," "fruity," "citrus," "woody," "musky," and "ambery." From the results (see **Table 2**), we selected two odors that maximized differences in liking and minimized differences in both intensity and familiarity. Perfume 1 diluted at 10% in DIPG (presented below as "Perfume") was then selected as the pleasant odor, while SBO diluted at 1% in IPM (presented below as "SBO") was selected as the unpleasant

FIGURE 3 | (A) Visual scene of the Olfactory Visual Approach/Avoidance by the Self Task (O-VAAST). **(B)** Perfume bottle in the vertical position (neutral stimulus). **(C,D)** Perfume bottle tilted to the left or right (visual targets of the task).

TABLE 2 | Pretested set of odors.

Product	Solvent	Liking ($M \pm SD$)	Intensity ($M \pm SD$)	Familiarity ($M \pm SD$)
Perfume 1, 1%	DIPG	52.2 ± 23.4	54.4 ± 17.8	59.4 ± 16.2
Perfume 1, 10%	DIPG	73.3 ± 18.3	66.4 ± 16.1	67.9 ± 15.9
Perfume 2, 1%	DIPG	46.2 ± 23.5	53.5 ± 20.3	58.7 ± 17.9
Perfume 2, 10%	DIPG	67.1 ± 20.4	68.2 ± 15.2	70.2 ± 16.7
Butyric acid, 0.5%	DIPG	7.1 ± 9.2	75.0 ± 18.4	72.7 ± 16.9
Butyric acid, 1%	DIPG	7.6 ± 12.3	81.9 ± 15.9	73.5 ± 18.6
SBO 0.001%	IPM	46.6 ± 18.8	36.5 ± 20.6	55.6 ± 20.5
SBO 0.005%	IPM	47.0 ± 16.6	32.0 ± 23.9	50.2 ± 17.7
SBO 0.1%	IPM	33.8 ± 22.2	51.2 ± 94.1	60.7 ± 21.8
SBO 1%	IPM	19.3 ± 21.7	62.9 ± 19.6	74.9 ± 18.0

odor. In the O-VAAST, DIPG was also used as the neutral odorless condition.

Olfactory Display System

The selected odors were then placed inside three glass vials and arranged in a custom-built computer-controlled olfactory display system (see Ischer et al., 2014). During the interstimulus interval, air valves were opened, thus delivering clean air to the nose of participants. During the delivery of odor, air valves were automatically closed and odor valves opened. As typically performed in our implicit olfactory procedures (Lemercier-Talbot et al., 2019; Cereghetti et al., 2020), the interstimulus interval and the odorant flow rate were both fixed at $2\ L \times min^{-1}$, thus delivering a constant flow to the nose of participants. Stainless steel tips were used as the final delivery piece.

Procedure

The procedure consisted of two successive phases: the odor evaluation phase and the O-VAAST.

Odor Evaluation Phase

In the odor evaluation phase, participants were instructed that they would be provided with odors to evaluate. The odors were administered in a random order for 6 s. After each olfactory stimulation, participants assessed pleasantness (from "very unpleasant" to "very pleasant"), intensity (from "not perceived" to "very strong"), and familiarity (from "not familiar at all" to "very familiar") by using visual analogic scales that ranged from 0 to 100 (see Delplanque et al., 2008, for more details).

O-VAAST

Developed in the MATLAB environment (ver. R2014b, The MathWorks, Inc., Natick, Massachusetts, United States) using the Psychtoolbox library (version 3.0.11), the O-VAAST consisted of two successive phases, namely, practice and test. In the practice phase, participants were told that they would be immersed in a virtual environment, i.e., a bathroom equipped with a perfume bottle. Participants were asked to move forward or backward from the perfume bottle as quickly and accurately as possible, depending on its inclination (i.e., left or right). The correspondence between the perfume bottle inclination and the approach/avoidance action was counterbalanced across participants: Half of the participants had to approach the left-tilted perfume bottles and avoid the right-tilted ones, whereas the remaining half had to approach the right-tilted perfume bottles and avoid the left-tilted ones.

In practice trials, the visual scene was presented full screen with the perfume bottle tilted to the left or right (visual target). Depending on its inclination, participants had to press key 8

of the numeric keypad three times successively to approach the perfume bottle and key 2 of the numeric keypad to move away from it. After each correct key-press response, the whole visual scene was zoomed in or out by 10%, giving the visual impression of walking forward or backward as a consequence of the approach/avoidance action of participants (Rougier et al., 2018). Following a wrong response, a red capital "X" was displayed for 500 ms and participants were required to provide the correct response to trigger the visual flow. Successive trials were separated by a fixation cross presented in the center of a blank screen for 250 ms. The practice phase was composed of four "move forward" and four "move backward" trials presented in a randomized order.

The test phase was similar, and the difference is that three olfactory stimuli (Perfume, SBO, and DIPG) were delivered as primes prior to the approach/avoidance actions of participants. In that respect, participants were informed that they would smell the odors coming from the perfume bottle throughout the task. The testing phase consisted of 24 odor blocks. In each block, the delivery of an olfactory cue was followed by four visual targets, i.e., four tilted perfume bottles presented consecutively (T1, T2, T3, and T4). At the beginning of each block, the visual scene was presented full screen with the perfume bottle in the vertical position (i.e., the neutral visual scene); participants were then instructed to press the "start" key (key 5 of the numeric keypad). This action triggered a 3-s countdown timer (3, 2, 1) displayed in the center of the vertical perfume bottle, after which an asterisk (*) was presented as the sniffing signal for 2 s. The odor was then delivered, and participants were instructed to inhale at this moment. In reality, the odor valve was opened 1 s before the onset of the sniffing signal, ensuring the presence of the odorant in the final delivery pieces of the olfactometer during participants' inhalation. The offset of the sniffing signal matched the closing of the odorant valve, which was thus opened for a total of 3 s. After olfactory priming, the vertical perfume bottle was replaced by T1. Participants thus had to approach/avoid the perfume bottle by following the inclination-action rules that they trained for during the practice phase. The correct response was then followed by a fixation cross presented in the center of a blank screen for 250 ms, after which the neutral visual scene was presented again. Participants thus had to again press the "start" key, this time to trigger the next trial within the block. A fixation cross was then presented in the center of the vertical perfume bottle for 500 ms, after which T2 was presented. This sequence also applied to T3 and T4; the difference is that after the response of participants to T4, the duration of the blank screen was increased from 250 ms to 4 s to prevent odor adaptation and habituation. Although participants had to press the relevant key three times consecutively to approach the perfume bottle or to move away from it, we considered accuracy and RT only for the first action of participants (see Rougier et al., 2018).

Across the 24 blocks, each participant was exposed to the perfume eight times, to the SBO eight times, and to the DIPG eight times, and was asked to respond in total to 96 tilted perfume bottles (48 "move forward" and 48 "move backward" trials). The order of presentation of odor and perfume bottle inclination was randomized for each participant. In particular, randomization of perfume bottle inclination was controlled at an individual level, ensuring that each inclination (left-tilted, right-tilted) was presented four times for each olfactory cue (Perfume, SBO, DIPG) × trial (T1, T2, T3, T4) level. Within a block, the four perfume bottles could be tilted to the same or to different directions, making it impossible to anticipate T3 and/or T4 responses from previous responses.

Results

Experimental data and analysis scripts are available on the OpenScience Framework platform at https://osf.io/jkr7m/

Odor Evaluation Phase

Three repeated-measures ANOVAs were performed to evaluate the differences in perceived intensity, familiarity, and liking between the three odors. A significant effect of odor was observed on perceived liking $[F_{(1.95,48.83)} = 36.03, p < 0.0001, \eta_p^2 = 0.59, 90\%$ CI $(0.42, 0.68)]$, intensity $[F_{(1.93,48.32)} = 37.10, p < 0.0001, \eta_p^2 = 0.60, 90\%$ CI $(0.43, 0.68)]$, and familiarity $[F_{(1.84,46.05)} = 12.86, p < 0.0001, \eta_p^2 = 0.34, 90\%$ CI $(0.15, 0.47)]$ (**Figure 4**). Two-by-two comparisons were then conducted by using a series of paired t-tests with Bonferroni correction. Not surprisingly, Perfume was perceived as being more pleasant than DIPG $[t(25) = 6.05, p \leq 0.0001, d_z = 1.19, 95\%$ CI $(0.67, 1.68)]$ and SBO $[t(25) = 8.20, p \leq 0.0001, d_z = 1.61, 95\%$ CI $(1.02, 2.19)]$. As expected, the SBO was also rated as being less pleasant than the DIPG $[t(25) = -2.60, p = 0.047, d_z = -0.51, 95\%$ CI $(-0.91, -0.10)]$. The two odorants were also perceived as being more intense [Perfume: $t(25) = 7.98, p \leq 0.0001, d_z = 1.57, 95\%$ CI $(0.98, 2.14)$; SBO: $t(25) = 7.20, p \leq 0.0001, d_z = 1.41, 95<\%$ CI $(0.86, 1.95)]$ and more familiar [Perfume: $t(25) = 4.18, p < 0.001, d_z = 0.82, 95\%$ CI $(0.37, 1.26)$; SBO: $t(25) = 4.02, p < 0.001, d_z = 0.79, 95\%$ CI $(0.34, 1.22)]$ than the DIPG. In contrast, we did not find any significant difference between Perfume and SBO in intensity $[t(25) = 1.49, p = 0.44, d_z = 0.29, 95\%$ CI $(-0.10, 0.68)]$ or familiarity $[t(25) = 0.37, p = 0.71, d_z = 0.07, 95\%$ CI $(-0.31, 0.46)]$. Critically, these results allowed us to exclude that a potential impact of the odors in the O-VAAST, analyzed below, could be attributed to differences in intensity or familiarity (i.e., differences other than valence).
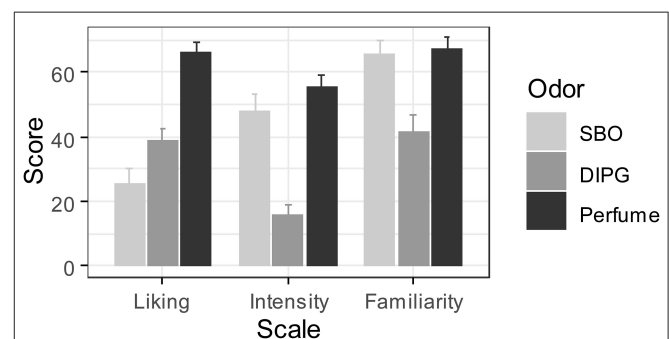
FIGURE 4 | Results on liking, intensity, and familiarity scales on Perfume, SBO, and DIPG.

Action	Olfactory cue	Accuracy ($M \pm SD$)
Move backward	DIPG	97.8 ± 3.9%
Move backward	Perfume	97.6 ± 4.4%
Move backward	SBO	98.1 ± 3.4%
Move forward	DIPG	95.4 ± 6.3%
Move forward	Perfume	97.8 ± 3.9%
Move forward	SBO	96.6 ± 5.1%

O-VAAST: Accuracy

Accuracy of participants on the O-VAAST across the six conditions is summarized in **Table 3**. We performed a repeated-measures ANOVA with the factors olfactory cue (Perfume, SBO, DIPG) and action ("move forward," "move backward") as within-subject factors for the response accuracy of participants. A significant effect of action [$F(1, 25) = 4.67$, $p = 0.04$, $\eta_p^2 = 0.16$, 90% CI (0.00, 0.36)] reflected better accuracy for "move backward" trials ($M = 97.8\%$, $SD = 2.9\%$) than for "move forward" trials ($M = 96.6\%$, $SD = 3.1\%$). Nevertheless, accuracy was influenced by neither the olfactory cue [$F_{(1.91, 47.84)} = 0.88$, $p = 0.42$, $\eta_p^2 = 0.03$, 90% CI (0.00, 0.12)] nor the cue–action interaction [$F_{(1.95, 48.76)} = 1.29$, $p = 0.28$, $\eta_p^2 = 0.05$, 90% CI (0.00, 0.15)].

O-VAAST: Reaction Times

Similar to the previous study, we first excluded incorrect trials (2.8%), as well as trials with RTs faster than 300 ms (1.1%) and slower than 1,500 ms (2.2%). For the remaining correct trials, RTs falling outside 2.5 SD from the mean of each participant computed at each target position level (T1, T2, T3, and T4) were also considered outliers and excluded from further analysis (2.2%). We then performed a repeated-measures ANOVA with the factors olfactory cue (Perfume, SBO, DIPG) and action ("move forward," "move backward") as within-subject factors for valid RTs. Analyses neither revealed a significant main effect of the olfactory cue [$F_{(1.95, 48.87)} = 1.62$, $p = 0.21$, $\eta_p^2 = 0.06$, 90% CI (0.00, 0.17)] nor of the type of action [$F_{(1, 25)} = 0.06$, $p = 0.81$, $\eta_p^2 = 0.002$, 90% CI (0.00, 0.09)]. However, notably, we found a critical significant interaction between action and olfactory cue [$F_{(1.84, 46.10)} = 3.80$, $p = 0.03$, $\eta_p^2 = 0.13$, 90% CI [0.01, 0.26]] (**Figure 5**), suggesting that the effect of odor pleasantness on RTs depended on the action type of participants. To characterize this interaction, we performed two repeated-measures ANOVAs to evaluate the cue effect (Perfume, SBO, DIPG) across the two move actions ("move forward" and "move backward"). The cue effect was significant on the "move forward" [$F_{(1.93, 48.22)} = 5.22$, $p = 0.01$, $\eta_p^2 = 0.17$, 90% CI (0.03, 0.30)] but not on the "move backward" [$F_{(1.98, 49.60)} = 0.31$, $p = 0.73$, $\eta_p^2 = 0.01$, 90% CI (0.00, 0.07)] condition. Further, one-tailed Bonferroni-corrected paired t-tests revealed that participants moved forward more slowly after being primed with the SBO ($M = 661$, $SD = 136$ ms) than after being primed by either the DIPG ($M = 626$, $SD = 111$ ms) [$t(25) = 2.86$, $p = 0.01$, $d_z = 0.56$, 95% CI (0.14, 0.97)] or the Perfume ($M = 636$, $SD = 119$ ms) [$t(25) = 2.42$, $p = 0.03$, $d_z =$
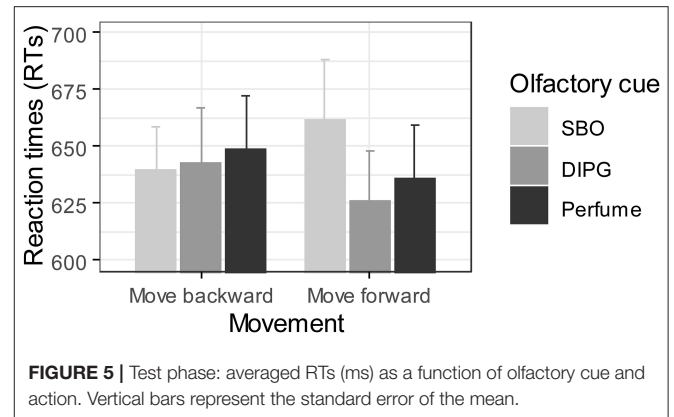
FIGURE 5 | Test phase: averaged RTs (ms) as a function of olfactory cue and action. Vertical bars represent the standard error of the mean.

0.47, 95% CI (0.06, 0.88)]. No difference was observed between DIPG and Perfume [$t(25) = 0.87$, $p = 1.00$, $d_z = 0.17$, 95% CI (−0.22, 0.56)].

Discussion Experiment 2

In this experiment, participants were asked to move forward or backward from the perfume bottle as quickly as possible, depending on its inclination (i.e., left or right). Three olfactory stimuli (Perfume, SBO, and DIPG) were delivered as primes, which allowed us to measure the impact of odor pleasantness on approach/avoidance reactions of participants. As expected, we observed a critical interaction between odor pleasantness and action on RTs of participants (i.e., the approach/avoidance compatibility effect). Further analysis revealed that odor pleasantness only influenced forward (not backward) movements. In particular, unpleasant odors (SBO) induced slower forward movements compared with those of neutral-to-pleasant odors (DIPG, Perfume), suggesting that this effect was principally driven by unpleasant odors.

GENERAL DISCUSSION

The main purpose of the present study was to extend the VAAST (Rougier et al., 2018) to the auditory and olfactory modalities. More specifically, we aimed to (1) validate new procedures to measure the ability of sounds and odors to elicit automatic approach/reactions toward their relative source, and (2) provide further evidence concerning the ability of VAAST to discriminate emotional stimuli that differ in their hedonic value, using approach/avoidance compatibility effects as implicit measures of hedonicity. In Experiment 1, participants were asked to move forward or backward from a loudspeaker that emitted spoken words presented at different levels of distortion (i.e., undistorted, slightly distorted, and moderately distorted); our assumption is that the more distorted the signal, the more unpleasant the stimulus. In Experiment 2, participants were asked to move forward or backward from a perfume bottle that delivered pleasant and unpleasant odors. We predicted that, consistent with the approach/avoidance compatibility effect, faster RTs would be required to approach positive stimuli and to avoid negative stimuli than was the case for the reverse.

In both experiments, we observed an expected interaction between the direction of the movement induced by the responses of participants (forward or backward) and the quality of the emotional stimulus. More precisely, undistorted words and neutral-to-pleasant odors (DIPG, Perfume) induced faster forward movements toward their source than did (moderately and slightly) distorted words and unpleasant odors (SBO). However, although the direction of the results was congruent with our hypothesis, the quality of the stimulation did not impact the backward movements of participants. This asymmetry of the compatibility effect could be related to the nature of the stimuli used in our experiments: Neither the emotional stimuli (auditory and olfactory cues) nor the visual scenes (including sources and the surrounding environments) were likely to be appraised as dangerous events that may induce fear and elicit avoidance reactions. That is, the methodology presented here is not intended to be used with strong emotional stimuli, especially negative stimuli, but rather with positive stimuli that slightly differ in their hedonic value. Consequently, to further develop the method, one could consider (1) dropping the avoidance condition and (2) contrasting the approach movement to a condition without movement.

Notably, our results confirmed that the VAAST is able to measure compatibility effects with a relatively small sample size, providing further evidence of its robustness and validity. According to Rougier et al. (2018), 35 subjects are enough to identify compatibility effects in the VAAST. Moreover, since (1) we used stronger emotional stimuli (in particular odors) and since (2) we implemented a behavioral task with higher ecological validity, we expected our tasks to be more sensitive. Indeed, the effect sizes observed in significant t-tests performed in RT analysis (Experiment 1: $d_z = 0.86$ and $d_z = 0.54$; Experiment 2: $d_z = 0.56$ and $d_z = 0.47$) were higher, or quite close, to the minimum effect size that can be detected with our sample size (Experiment 1: $d_z = 0.51$; Experiment 2: $d_z = 0.50$), assuming $\alpha = 0.05$ and power $(1 - \beta) = 0.80$.

Most studies conducted so far on the VAAST used explicit instructions (i.e., procedures in which participants are explicitly asked to process the valence of the emotional stimulus, see Rougier et al., 2018). Importantly, our study also supported the evidence that relatively large approach (but not avoidance) compatibility effects ($\eta_p^2 = 0.20$ in Experiment 1; $\eta_p^2 = 0.13$ in Experiment 2) can be observed when using implicit instructions, i.e., without requiring participants to explicitly attend the valence of the emotional stimulus (Phaf et al., 2014). In Experiment 1, participants were instructed to respond to the word category, regardless of the level of distortion at which these stimuli were presented. That is, they were required to explicitly process a feature of emotional stimuli other than valence. In Experiment 2, participants were instructed to respond to the perfume bottle inclination, regardless of the olfactory cues delivered by the olfactometer. Experiment 2 thus went further by demonstrating that compatibility effects can be found without requiring participants to explicitly process the stimuli that induce emotion, the latter being task irrelevant. In that sense, the VAAST may be a privileged procedure to measure approach/avoidance reactions by using implicit instructions, especially when multiple sensory modalities are involved. In this case, target stimuli and emotional stimuli could be presented in different sensory modalities (as in Experiment 2), thus providing the best conditions for applying implicit instructions. Moreover, the VAAST appears to be, by its nature, a procedure that fits particularly well in cross-modal research: One could imagine its use with very diverse emotional stimuli, including visual (as in the original version of the VAAST), auditory, and olfactory stimuli (as in the present study), but potentially also gustatory and haptic/tactile stimuli. In this case, the challenge for the investigator will be to identify a visual context (toward which the approach/avoidance reactions are directed) that is meaningfully linked to the emotional stimuli.

The choice of the task is also crucial because it defines the cognitive processes involved during its completion, potentially affecting the sensitivity of the measure, i.e., the approach/avoidance compatibility effect sizes. In the present study, RTs were higher in Experiment 1 ($M = 913$, $SD = 104$ ms) than in Experiment 2 ($M = 643$, $SD = 111$ ms). This difference is mostly due to the nature of cognitive processes involved in the two tasks. In Experiment 1, the task required semantic processing of spoken words. This was not the case in the Experiment 2, in which participants had to discriminate an elementary feature of visual targets (i.e., their inclination). Furthermore, as shown in different studies based on a dual-task paradigm (Gagné et al., 2017) and on cognitively overloaded auditory tasks (Gros et al., 2008), audio degradation could induce an increase in listening effort, resulting in an increase in RTs. However, in that case, lengthening of RTs would appear for both backward and forward movements, which was not the case in our experiment. We can thus exclude the possibility that the levels of degradation used in our (non-cognitively overloaded) auditory VAAST induced an increase in listening effort. Obviously, if we had used higher levels of degradation, differences in RTs would not specifically reflect the hedonicity of the listening experience, but rather the difficulty of understanding, or perhaps the intelligibility loss. In this case, either a lengthening of RTs would appear for highly distorted sounds regardless of the movement, or faster approach movements would be observed for highly distorted sounds only, with participants getting closer to the loudspeaker to better understand them. An experiment is in progress with our team to verify the outcome.

The VAAST is likely to have higher ecological validity compared with that of explicit measures, and even compared with that of other measures. First, the VAAST captures cognitive processes that "naturally" occur when emotional stimuli are encountered. In the so-called ecological situations, that is in "real-life" contexts, individuals rarely express their perceptual judgments explicitly. Perceptual feelings are nevertheless likely to influence decisions and behaviors of individuals in their daily life. On the basis of the idea that approach/avoidance behaviors belong to a critical adaptive process (Tooby and Cosmides, 1990), Bargh (1997) argues that encountered stimuli are automatically evaluated by the cognitive system on a positive/negative dimension. This evaluation leads to the activation of behavioral predispositions (Chen and Bargh, 1999) that finally elicit approach/avoidance reactions. These predispositions, which are activated in ecological situations, are also activated during the VAAST, making it a procedure able to capture cognitive processes occurring in "real-life" contexts. Second, by providing visual

feedback mimicking whole-body movements in relation to the emotional stimulus, the VAAST clearly simulates what happens in "real life" when such a stimulus is approached or avoided. Importantly, although in the original VAAST (Rougier et al., 2018) participants had to approach/avoid words located in a non-congruent environment (a corridor in Experiments 1, 2, 3, and 5; a street in Experiments 4 and 6), in our procedures, participants had to approach/avoid 3D virtual objects (i.e., a loudspeaker and a perfume bottle) located in a congruent environment (i.e., a living room and a bathroom), thus increasing the ecological validity of our procedures. We hope that future research will further explore the potential of the VAAST to measure behavioral tendencies in ecological situations by using 3D virtual objects as target stimuli instead of emotional words.

The VAAST could be a very promising procedure to capture cognitive processes that are "naturally" involved during product experience. Indeed, one could argue that the perception of a product (e.g., loudspeaker, perfume), its related cues (e.g., auditory, olfactory), and its conditioned stimuli (e.g., brand, packaging) could lead to the activation of behavioral tendencies that promote product purchase or consumption. The visual scene, including sources (e.g., loudspeaker, perfume bottle) and the surrounding environments (e.g., living room, bathroom), can potentially be manipulated by the investigator to test the effect of specific contextual factors on reactions of consumers toward products, thus making the VAAST a contextual-rich procedure that, we are certain, has much potential in consumer research.

CONCLUSION

Here, we presented an adapted version of the VAAST (Rougier et al., 2018) that allowed us to measure the ability of auditory and olfactory cues to trigger automatic approach/avoidance reactions toward their source. In two experiments, we showed that the VAAST can successfully be adapted to modalities other than visual.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: https://osf.io/jkr7m/.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DC, PF, LG, LM, ED, IC, and RV designed the studies. LM implemented the A-VAAST. DC implemented the O-VAAST. LM ran the first experiment. DC and LM ran the second experiment. DC and LM analyzed and interpreted the data of both studies. DC, TH, and RV wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

REFERENCES

Alexopoulos, T., and Ric, F. (2007). The evaluation-behavior link: direct and beyond valence. *J. Exp. Soc. Psychol.* 43, 1010–1016. doi: 10.1016/j.jesp.2006.10.017

Bamford, S., and Ward, R. (2008). Predispositions to approach and avoid are contextually sensitive and goal dependent. *Emotion* 8, 174–183. doi: 10.1037/1528-3542.8.2.174

Bargh, J. A. (1997). "The automaticity of everyday life," in *Advances in Social Cognition*, ed R. S. Wyer (Mahwah: Erlbaum), 1–61.

Bonin, P., Méot, A., Aubert, L., Malardier, N., Niedenthal, P., and Capelle-Toczek, M.-C. (2003). Normes de concrétude, de valeur d'imagerie, de fréquence subjective et de valence émotionnelle pour 866 mots [Concreteness, imageability, subjective frequency and emotionality ratings for 866 words]. *L'Année Psychologique* 103, 655–694. doi: 10.3406/psy.2003.29658

Carr, E. W., Rotteveel, M., and Winkielman, P. (2016). Easy moves: perceptual fluency facilitates approach-related action. *Emotion* 16, 540–552. doi: 10.1037/emo0000146

Cereghetti, D., Chill,à, C., Porcherot, C., Sander, D., Cayeux, I., and Delplanque, S. (2020). LikeWant: a new methodology to measure implicit wanting for flavors and fragrances. *Food Qual. Prefer.* 80:103829. doi: 10.1016/j.foodqual.2019.103829

Chen, M., and Bargh, J. A. (1999). Consequences of automatic evaluation: immediate behavioral predispositions to approach or avoid the stimulus. *Pers. Soc. Psychol. Bull.* 25, 215–224. doi: 10.1177/0146167299025002007

Delplanque, S., Grandjean, D., Chrea, C., Aymard, L., Cayeux, I., Le Calve, B., et al. (2008). Emotional processing of odors: evidence for a nonlinear relation between pleasantness and familiarity evaluations. *Chem. Senses* 33, 469–479. doi: 10.1093/chemse/bjn014

Eder, A. B., and Rothermund, K. (2008). When do motor behaviors (mis)match affective stimuli? An evaluative coding view of approach and avoidance reactions. *J. Exp. Psychol.* 137, 262–281. doi: 10.1037/0096-3445.137.2.262

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *J. Appl. Psychol.* 37, 90–93. doi: 10.1037/h0058073

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39 175–191. doi: 10.3758/BF03193146

Gagné, J. P., Besser, J., and Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm. *Trends Hear.* 21:2331216516687287. doi: 10.1177/2331216516687287

Gros, L., Durin, V., and Chateau, N. (2008). Redrawing the link between customer satisfaction and speech quality. *Acta Acustica* 94, 32–42. doi: 10.3813/AAA.918005

Ischer, M., Baron, N., Mermoud, C., Cayeux, I., Porcherot, C., Sander, D., et al. (2014). How incorporation of scents could enhance immersive virtual experiences. *Front. Psychol.* 5:736. doi: 10.3389/fpsyg.2014.00736

ITU-T Recommendation G.191 (2019). *Software Tools for Speech and Audio Coding Standardization*. Available online at: https://www.itu.int (accessed July 31, 2020).

ITU-T Recommendation P.56 (2011). *Objective Measurement of Active Speech Level*. Available online at: https://www.itu.int (accessed July 31, 2020).

ITU-T Recommendation P810 (1996). *Modulated Noise Reference Unit (MNRU)*. Available online at: https://www.itu.int. (accessed July 31, 2020).

Krieglmeyer, R., Deutsch, R., De Houwer, J., and De Raedt, R. (2010). Being Moved: Valence activates approach-avoidance behavior independently of evaluation and approach-avoidance intentions. *Psychol. Sci.* 21, 607–613. doi: 10.1177/0956797610365131

Lavender, T., and Hommel, B. (2007). Affect and action: Towards an event-coding account. *Cogn. Emotion* 21, 1270–1296. doi: 10.1080/02699930701438152

Lemercier-Talbot, A., Coppin, G., Cereghetti, D., Porcherot, C., Cayeux, I., and Delplanque, S. (2019). Measuring automatic associations between relaxing/energizing feelings and odors. *Food Qual. Prefer.* 77, 21–31. doi: 10.1016/j.foodqual.2019.04.010

Markman, A. B., and Brendl, C. M. (2005). Constraining theories of embodied cognition. *Psychol. Sci.* 16, 6–10. doi: 10.1111/j.0956-7976.2005.00772.x

Mathôt, S., Schreij, D., and Theeuwes, J. (2012). OpenSesame: an open-source, graphical experiment builder for the social sciences. *Behav. Res. Methods* 44, 314–324. doi: 10.3758/s13428-011-0168-7

Niedenthal, P. M. (2007). Embodying emotion. *Science* 316, 1002–1005. doi: 10.1126/science.1136930

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychol. Rev.* 84, 231–259. doi: 10.1037/0033-295X.84.3.231

Paladino, M.-P., and Castelli, L. (2008). On the immediate consequences of intergroup categorization: activation of approach and avoidance motor behavior toward ingroup and outgroup members. *Pers. Soc. Psychol. Bull.* 34, 755–768. doi: 10.1177/0146167208315155

Paulus, A., and Wentura, D. (2016). It depends: approach and avoidance reactions to emotional expressions are influenced by the contrast emotions presented in the task. *J. Exp. Psychol.* 42, 197–212. doi: 10.1037/xhp0000130

Phaf, R. H., Mohr, S. E., Rotteveel, M., and Wicherts, J. M. (2014). Approach, avoidance, and affect: a meta-analysis of approach-avoidance tendencies in manual reaction time tasks. *Front. Psychol.* 5:378. doi: 10.3389/fpsyg.2014.00378

Poulton, E. C. (1989). *Bias in Quantifying Judgments*. Hillsdale: Erlbaum.

Rinck, M., and Becker, E. S. (2007). Approach and avoidance in fear of spiders. *J. Behav. Ther. Exp. Psychiatry* 38, 105–120. doi: 10.1016/j.jbtep.2006.10.001

Rotteveel, M., Gierholz, A., Koch, G., van Aalst, C., Pinto, Y., Matzke, D., et al. (2015). On the automatic link between affect and tendencies to approach and avoid: Chen and Bargh (1999) revisited. *Front. Psychol.* 6:335. doi: 10.3389/fpsyg.2015.00335

Rotteveel, M., and Phaf, R. H. (2004). Automatic affective evaluation does not automatically predispose for arm flexion and extension. *Emotion* 4, 156–172. doi: 10.1037/1528-3542.4.2.156

Rougier, M., Muller, D., Ric, F., Alexopoulos, T., Batailler, C., Smeding, A., et al. (2018). A new look at sensorimotor aspects in approach/avoidance tendencies: the role of visual whole-body movement information. *J. Exp. Soc. Psychol.* 76, 42–53. doi: 10.1016/j.jesp.2017.12.004

Saraiva, A. C., Schüür, F., and Bestmann, S. (2013). Emotional valence and contextual affordances flexibly shape approach-avoidance movements. *Front. Psychol.* 4:933. doi: 10.3389/fpsyg.2013.00933

Seibt, B., Neumann, R., Nussinson, R., and Strack, F. (2008). Movement direction or change in distance? Self- and object-related approach–avoidance motions. *J. Exp. Soc. Psychol.* 44, 713–720. doi: 10.1016/j.jesp.2007.04.013

Solarz, A. K. (1960). Latency of instrumental responses as a function of compatibility with the meaning of eliciting verbal signs. *J. Exp. Psychol.* 59, 239–245. doi: 10.1037/h0047274

Tooby, J., and Cosmides, L. (1990). The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethol. Sociobiol.* 11, 375–424. doi: 10.1016/0162-3095(90)90017-Z

van Dantzig, S., Zeelenberg, R., and Pecher, D. (2009). Unconstraining theories of embodied cognition. *J. Exp. Soc. Psychol.* 45, 345–351. doi: 10.1016/j.jesp.2008.11.001

Versace, R., Vallet, G. T., Riou, B., Lesourd, M., Labeye, É., and Brunel, L. (2014). Act-in: an integrated view of memory mechanisms. *J. Cogn. Psychol.* 26, 280–306. doi: 10.1080/20445911.2014.892113

Wentura, D., Rothermund, K., and Bak, P. (2000). Automatic vigilance: the attention grabbing power of approach-and avoidance-related social information. *J. Pers. Soc. Psychol.* 78, 1024–1037. doi: 10.1037/0022-3514.78.6.1024

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Differences in Mechanical Parameters of Keyboard Switches Modulate Motor Preparation: A Wearable EEG Study

Hiroki Watanabe[1], Kae Nakajima[1], Shunsuke Takagi[2], Ryo Mizuyama[2], Mayumi Saito[2], Koichi Furusawa[2], Kunio Nakatani[2], Yusuke Yokota[1], Hirotaka Kataoka[2], Hiroshi Nakajima[2] and Yasushi Naruse[1]*

[1] Center for Information and Neural Networks, National Institute of Information and Communications Technology, and Osaka University, Kobe, Japan, [2] OMRON Corporation, Kyoto, Japan

The mechanical parameters of keyboard switches affect the psychological sense of pressing. The effects of different mechanical parameters on psychological sense have been quantified using questionnaires, but these subjective evaluations are unable to fully clarify the modulation of information processing in the brain due to these differences. This study aimed to elucidate the ability of electroencephalography (EEG) measurements to detect the modulation of subconscious information processing according to mechanical parameter values. To this end, we prepared five mechanical switches with linearly increasing values of pretravel (PT: the distance from the free position until the operating position). We hypothesized that the differences in PTs would subconsciously affect the motor preparation prior to pressing switches because switches with PTs that deviated from those commonly used were predicted to increase the users' attention level when pressing. Differences in motor preparation were quantified using the mean amplitudes of the late contingent negative variation (CNV). We recorded EEGs of 25 gamers during a reaction task for fast switch pressing after a response cue preceded by a pre-cue for response preparation; we also measured the reaction time feedback on each switch pressing trial. Participants performed five sessions (60 trials per session) in total. For the analysis, trials were divided into first (session 1, 2, and 3) and second half sessions (session 4 and 5). In the latter session, CNV amplitudes were significantly higher for the switch with the highest PT than for that with a medium PT, which is closest to that commonly used in commercial mechanical switches. On the other hand, the questionnaire did not detect any significant differences between PTs in their subjective rankings of the psychological effects of switch pressing. These results suggest that differences in PTs modulate motor preparation to press switches, and that EEG measurements may provide a novel objective evaluation of the mechanical parameters of keyboard switches.

Keywords: contingent negative variation, motor preparation, electroencephalography, event-related potentials, neuromarketing

1. INTRODUCTION

The quantification of the psychological feelings of products plays a crucial role in evaluating the product design process to ensure the development of consumer-oriented products (Nagamachi, 1995). To date, questionnaire surveys such as the rating scale method (raters select an applicable level on a 5- or 7-scale for each question item), paired comparisons (Thurstone, 1927; Scheffé, 1952), or semantic differential method (Osgood et al., 1957) have been utilized in evaluation of the design for quantification purposes. Although these subjective evaluations have the advantage of gathering a large amount of data in a relatively short period, raters' biases may hinder reliable measurements. For example, social desirability bias tends to lead raters to choose the answers that are favorably recognized by others (Furnham, 1986; King and Bruner, 2000). Further, subjective evaluation generally fails to quantify raters' exact feelings on the products, as emotional processes in the brain may be induced subconsciously (Kiss and Eimer, 2008; Pantazatos et al., 2012), and subconscious processing affects the consumption behavior of consumers (Winkielman et al., 2005).

A potential approach for more accurate and reliable evaluation of products is the incorporation of brain activity measurements into the product design process. Neuromarketing or consumer neuroscience has received substantial attention as a field focusing on the objective quantification of consumers' feelings, preferences, and cognitive processing of products or product advertising based on brain activity for product marketing (Ariely and Berns, 2010; Morin, 2011; Spence, 2019). Extensive research has demonstrated that machine-learning technology enables the estimation of emotional states using brain responses (Wang et al., 2014; Aldayel et al., 2020; Liu et al., 2020). In addition, measurements of blood-oxygen-level dependent (BOLD) signals using functional magnetic resonance imaging (fMRI) or electroencephalogram (EEG) enabled marketers to clarify the brain activity patterns underscoring consumers' willingness to pay for products (Plassmann et al., 2007; Ramsøy et al., 2018). Furthermore, for objective evaluation of products, Guo et al. (2020) indicated that the measurement of N400, an event-related potential (ERP) associated with semantic processing, could be used to identify adjectives that well describe products. For neuromarketing research, there are several ways to measure brain activity. MRI scanners are not suitable for real-world environments research due to their large size, but EEG is compact and has relatively low running costs, making it suitable for real-world neuromarketing research (Bazzani et al., 2020) such as an evaluation of a beverage machine in an office environment (Sargent et al., 2020).

Given the relevance of brain state information for the objective evaluation of products in neuromarketing, its measurement may be useful for evaluating the mechanical parameters of industrial products. In particular, we focused on incorporating brain states into the evaluation of the mechanical parameters of mechanical keyboard switches. Mechanical switches include various mechanical parameters such as pre-travel (PT), which corresponds to the distance traveled by the switch when moving from a position where external force is not applied (i.e., the free position) to another when it is pressed (i.e., the operating position); the operating force, which is the force required to move the switch from the free position to the operating position; and total travel, the travel distance from the free position to the switch limit position (**Supplementary Figure 1**). Recently, various types of mechanical switches with different mechanical parameters have been attracting attention in the pursuit of performance in e-sports. Among those parameters, PT is particularly difficult to evaluate because PT differences do not change any physical feeling while pressing switches (i.e., the difference is just a distance), and thus, users cannot explicitly notice PT differences just by switch pressing. Thus, the method to evaluate the optimal PT value in switch design is to receive feedback on the reaction times (RTs) when pressing switches with different PT values. However, there is a possibility that inappropriate PT values require users to increase their attention to switch operations. For a long PT, the RT would be longer than necessary, and the switch considered difficult to operate for quick responses. Given that RTs are correlated with sustained attention in reaction tasks (Buck, 1966), a longer PT might modulate the allocation of attentional resources to switch operation for quicker responses than switches with normal PT values. Also, switches with extremely short PTs tend to produce unexpectedly fast RTs, and such unpredictability may increase the allocation of attention to switch operations. Based on the above, we predicted that switches with largely different PT values from commonly used ones would require more attentional resources during motor preparation due to unpredictable responses. If so, adding the estimation of the attention allocated to switch operation to switch evaluation criteria would be beneficial to prevent the increase in attentional resources during the gameplay. Therefore, the present study aimed to investigate whether pressing switches with deviated PT values induces an increase in attentional resources in switch operations, and whether such an increase can be detected using EEG measurements.

If, as expected, differing PT values from commonly used ones increase the allocation of attentional resources during switch operation, differences in PT values would be expected to be reflected in the amplitude of contingent negative variation (CNV) during motor preparation before switch pressing. This ERP component is characterized by a sustained negative deflection during motor preparation before a response cue (the target cue) preceded by a pre-cue for response preparation (the pre-cue) (Walter et al., 1964). The CNV consists of early and late components, with the late one reflecting the participants' motor preparation to a target cue (Gaillard, 1977; Leuthold et al., 2004). Previous research has demonstrated that an increase in attention during motor preparation enhances late CNV amplitude (McCallum and Walter, 1968; Tecce, 1972) and CNV has been used to investigate attentional mechanisms in motor response tasks (Liebrand et al., 2017). If the difference in PT affects the user's attention to the switch operation, then the difference in PT values would modulate the late CNV amplitude. However, it remains unclear whether this measure of brain activity allows the assessment of an increase of the

attentional resources allocated to switch operation induced by PT differences.

To assess the validity of including the attentional resources allocated to switch operation into the evaluation criteria for mechanical switches, we investigated whether different PTs increase the attentional resources during motor preparation for switch pressing and thus modulate late CNV amplitude. To this end, we prepared five experimental switches with different PTs linearly increasing from Switch 1 (shortest PT) to Switch 5 (longest PT), and Switch 3 corresponding to the ordinary PT range in commercially available switches, while maintaining other mechanical parameters constant. If differences in PT modulate the attentional resources allocated to switch operation and the late CNV is a valid indicator of this change, then we predicted that the mean late CNV amplitude would be enhanced when pressing a switch with a largely different PT from normal switches (i.e., Switch 1 or Switch 5; switches with shorter or longer PT; see section 2.2 for details). In order to induce the late CNV, subjects performed a reaction task in which they had to press these switches as quickly as possible, preceded by a pre-cue to indicate which switch to press among all five switches. Due to the purpose of the current study, data were collected from subjects with experience using gaming keyboards. Since it is not possible to notice differences in PT values just by the sensation of switch pressing, auditory and visual feedback of RTs was provided so that participants implicitly noticed the gap between predicted and actual RTs with session progression. To consider such implicit learning throughout sessions, we divided the trials into a total of five sessions with first half (session 1, 2, and 3) and second half sessions (session 4 and 5) for data analysis.

We also examined whether differences in PT values are reflected in other EEG responses via implicit learning. First, we expected that the parietal pre-cue-locked P3 would also be modulated according to the switch type since a direct relationship between PT values and RT speed could result in a reward prediction that shorter PT switches tend to score better on the task. Given that task-relevant, reward-predictive cues increase P3 amplitude (Krebs et al., 2013; Schevernels et al., 2014, 2016; Carsten et al., 2021), it is likely that differences in subjective reward prediction across PT values modulate the amplitude. We also analyzed feedback-related P3, which is distributed in the parietal region and reflects outcome evaluation and reward processing (Yeung and Sanfey, 2004; Hajcak et al., 2005, 2007; Wu and Zhou, 2009), as this component is known to be enhanced when subjects put more effort in difficult tasks (Ma et al., 2014; Schevernels et al., 2016). If subjects change their effort due to PT differences, the amplitude of feedback-related P3 might also be modulated. Second, the gap between the actual and predicted RTs when pressing a switch with deviating PTs might direct participants' attention to the feedback and modulate feedback-related ERPs. Since top-down attention enhances evoked auditory ERPs such as N1 (Hillyard et al., 1973; Näätänen and Picton, 1987), we analyzed the N1 response evoked by the feedback sound. Finally, we measured frontal theta power before switch pressing as an EEG index to quantify the mental workload (Sammer et al., 2007; So et al., 2017). Since the change in the power has applied as an index of mental workload in real-world neuroergonomics research (e.g., Aricó et al., 2016), we also measured this spectral power during motor preparation in order to contrast our proposed CNV index with the index often used in the research field. After the EEG measurements, to determine whether a questionnaire survey could also detect differences in PT values, subjects were asked to respond to a subjective evaluation of each switch.
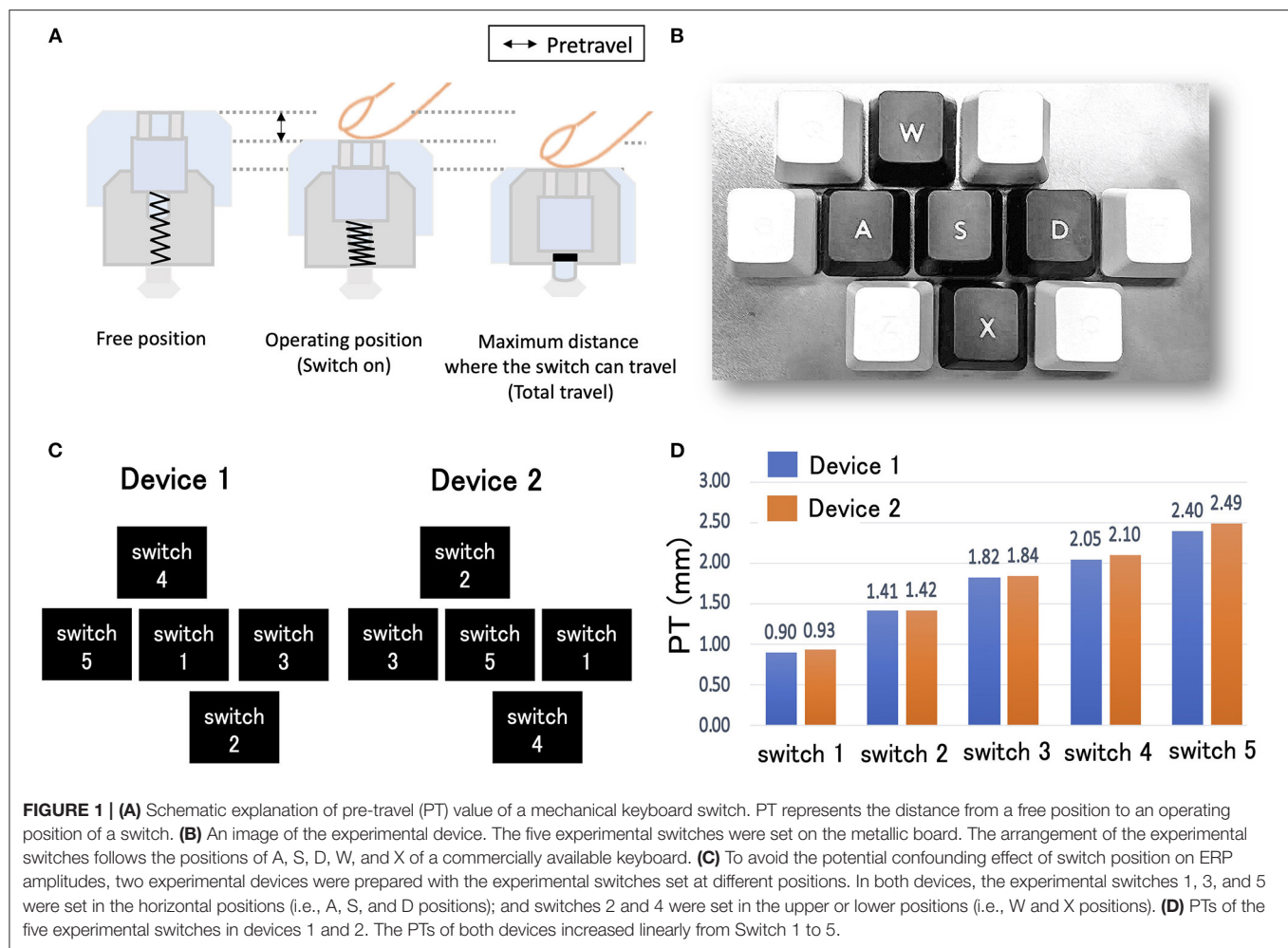
2. METHODS

2.1. Participants

To obtain data from participants with sufficient computer gaming experience using a gaming keyboard, we recruited participants who met both of the following conditions: an individual who (1) has purchased a gaming keyboard before, and (2) has played any of First Person Shooter (FPS), Third Person Shooter, Multiplayer Online Battle Arena, Multiplayer Online Role-Playing Game, and Massively Multiplayer Online Role-Playing Game over 50 h.

In total, 25 right-handed males participated in the current study (age range: 20–35, mean age = 25.8, SD = 5.0). Participants had normal or corrected-to-normal vision and reported no medical history of mental disorders, attention deficit disorder, or neurological disorders. The current study was approved by the Ethical Committee for Human and Animal Research of the National Institute of Information and Communications Technology. Participants agreed to participate in the current study and provided written informed consent before the experiment.

The total number of participants was determined based on a pilot study ($N = 9$) to analyze the differences in the mean amplitude of late CNV between switches with short, medium, and long PT values. The results of the pilot study showed that the effect size (Cohen's d) of the difference in late CNV between switches with short and medium PT values and between switches with long and medium PT values was 0.68 and 0.93, respectively. Power analysis at an alpha level of 0.05, statistical power of 0.8, and an effect size of 0.68 using G∗Power 3.1 (Faul et al., 2007) showed that 19 participants were needed for the paired t-test. In addition, a medium effect size of d_z 0.58 was detectable for the paired t-test with the current number of participants.

2.2. Apparatus

To investigate the modulation of EEG responses according to mechanical parameters of keyboard switches, we prepared five experimental switches with different PTs. **Figure 1A** depicts a schematic explanation of the PT of the mechanical switches. The PTs of the experimental switches increased linearly from Switch 1 with the shortest PT to Switch 5 with the longest PT. The experimental switches were mounted on a metallic board to follow the A, S, D, W, and X key positions of a commercially available keyboard (**Figure 1B**). These keys were chosen because they are frequently used to move characters in FPS games. To prevent a switch position from affecting ERP amplitudes, we prepared two switch devices which set the experimental switches in different positions. In both devices, experimental switches 1, 3, and 5 were set in the horizontal positions (i.e., A, S, and

FIGURE 1 | (A) Schematic explanation of pre-travel (PT) value of a mechanical keyboard switch. PT represents the distance from a free position to an operating position of a switch. **(B)** An image of the experimental device. The five experimental switches were set on the metallic board. The arrangement of the experimental switches follows the positions of A, S, D, W, and X of a commercially available keyboard. **(C)** To avoid the potential confounding effect of switch position on ERP amplitudes, two experimental devices were prepared with the experimental switches set at different positions. In both devices, the experimental switches 1, 3, and 5 were set in the horizontal positions (i.e., A, S, and D positions); and switches 2 and 4 were set in the upper or lower positions (i.e., W and X positions). **(D)** PTs of the five experimental switches in devices 1 and 2. The PTs of both devices increased linearly from Switch 1 to 5.
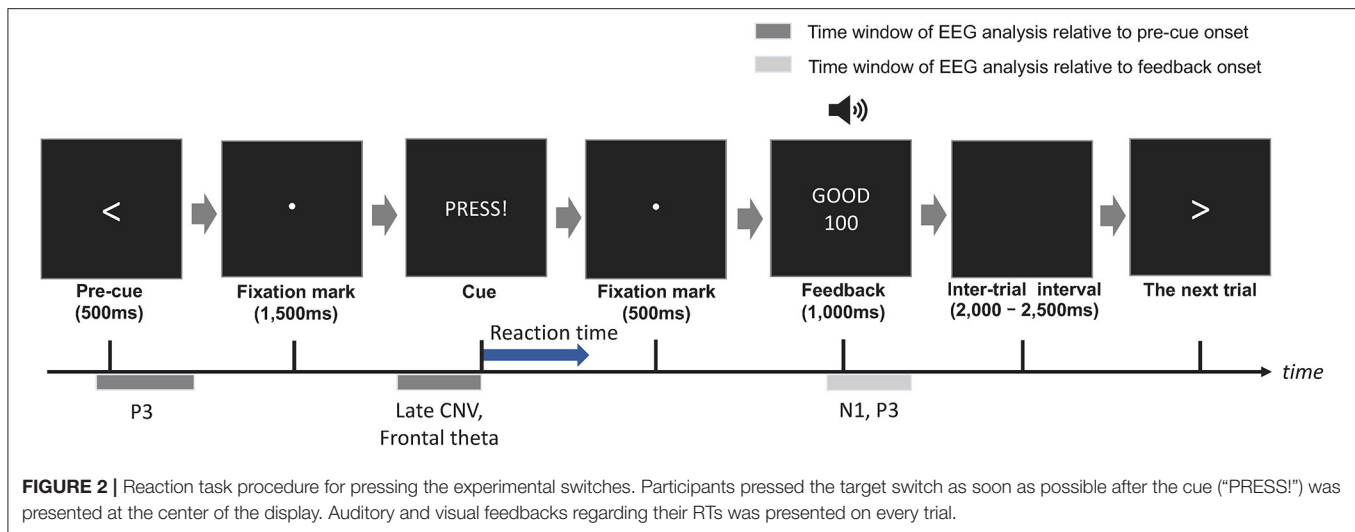
D positions); and switches 2 and 4 were set in the upper or lower positions (i.e., W and X positions; **Figure 1C**). Half of the participants used device 1, and the remaining participants used device 2. **Figure 1D** depicts the PTs of the experimental switches in both devices. The PT of Switch 3 was considered the closest to the commonly used value in mechanical switches based on an investigation of the PTs of 57 commercially available mechanical switches which revealed a mean value of 1.80 ± 0.35.

EEGs were measured using a wireless wearable system (PolymateMini AP108; Miyuki Giken Co., Ltd., Tokyo, Japan) with dry electrodes (Unique Medical Co., Ltd., Tokyo, Japan). The electrodes were positioned on Fz, Cz, and Pz sites according to the International 10-20 system. The ground and reference electrodes were set on the left and right earlobe, respectively. To monitor eye-related activity, horizontal and vertical electrooculograms (EOGs) were recorded from electrodes placed on the top and side to the lateral canthus of the participants' left eyes. All signals were sampled at 500 Hz.

2.3. Data Collection

Participants sat on a comfortable chair in a dimly lit soundproof room and performed a reaction task to press the experimental

switches. The monitor display and switch device were set in front of the participants. Because the A, S, D, W, and X keys are assigned on the left side of the keyboard, in order to reproduce the key presses during actual game play, we asked the participants to press switches with their non-dominant left hand. A trial procedure is summarized in **Figure 2**. Participants placed their left middle finger on the center switch (i.e., S switch) before the start of each trial. First, a pre-cue to indicate the switch to be pressed by the participants (i.e., target switch) appeared at the center of the display for 500 ms. The pre-cues >, ≡, <, ∧, and ∨ represented the right, center, left, upper, and lower switches on the experimental device (i.e., D, S, A, W, and X switches; cf. **Figure 1B**), respectively. Participants set their left middle finger on the target switch to prepare for a response. To indicate a fixation point, the pre-cue was replaced with a white dot for 1,500 ms. Participants pressed the target switch as quickly as possible after the presentation of the cue string "PRESS!." The cue was changed back to the fixation mark immediately after the participants' response or automatically 500 ms after the onset of the cue. Visual and auditory feedback on the participant's responses was presented 500 ms after the participant's response and the visual feedback lasted for 1,000 ms. The inter-stimulus

FIGURE 2 | Reaction task procedure for pressing the experimental switches. Participants pressed the target switch as soon as possible after the cue ("PRESS!") was presented at the center of the display. Auditory and visual feedbacks regarding their RTs was presented on every trial.

interval from the feedback end to the beginning of the next trial was randomly set from 2,000 to 2,500 ms. The interval was chosen to avoid contamination of feedback-related components with the baseline time region for ERP calculation. Participants were instructed to avoid blinking as much as possible during the period from the onset of a pre-cue to the onset of feedback to avoid contaminating EEG data with eye-related activity.

During the feedback section in a trial, an earned point and response category were presented on the display. These parameters were calculated based on participants' RT. The response categories were *False start* (RT < 150 ms), *Good* (150 ms ≤ RT < 330 ms), *Late* (330 ms ≤ RT < 700 ms), and *Over* (700 ms ≤ RT). The time range of *False start* was chosen given that simple RT generally falls within 200–300 ms (e.g., Eckner et al., 2010). The trials in which participants pressed a non-target switch were categorized as *Error*. The earned point was calculated using Equation (1):

$$
points = \begin{cases}
-500 \\
\quad (response = \textit{False start or response} = \textit{Over}) \\
\frac{300 \times (RT_{upper} - RT)}{RT_{upper} - RT_{lower}} \\
\quad (response = \textit{Good}) \\
-100 \\
\quad (response = \textit{Late or response} = \textit{Error})
\end{cases}
$$

(1)

where RT_{lower} and RT_{upper} are the lower and upper limits of RTs in *Good* trials (i.e., 150 and 330 ms), respectively. In the case of *Good* trials, the equation transformed participants' RT in a range from 0 to 300 points. In the feedback section, feedback on whether the RT was categorized as *Good* or not was provided aurally, and the category of participants' responses and earned points were presented on the display. After completion of a trial session, the total points across switches and per switch were displayed on the display. Participants performed 12 trials per switch in a session (i.e., 60 trials/session). The order of trials was randomized per session and participant. Participants

performed one practice session and five main sessions. The EEG data collection procedure, including recording preparation, lasted ~60 min.

After the reaction task, a questionnaire survey was administered to investigate the subjective feelings of pressing the switches. Participants ranked the switches according to each question item on a 5-point scale while referring to the mean earned points of each switch across sessions (1: the best; 5: the worst). Participants were allowed to press the experimental switches anytime when answering the questions to re-confirm the feelings of the switches. A tied rank was assigned for cases in which participants judged that there was no difference in subjective feelings between switches. The question items are summarized in **Table 1**.

2.4. EEG Analysis

We used the EEGLAB toolbox (Delorme and Makeig, 2004) for MATLAB (The MathWorks, Inc., U.S.A) to analyze EEG data. The data were preprocessed separately per participant and session. Raw continuous EEG and EOG data were off-line filtered using a 16,500th finite impulse response (FIR) high-pass filter at 0.1 Hz and a 330th low-pass FIR filter at 20 Hz. The artifacts of the raw data were suppressed using an Artifact Subspace Reconstruction method (Mullen et al., 2015) using an open-source plug-in function *clean_rawdata* in the toolbox. To correct for eye-related activity, data were decomposed using independent component analysis. The components related to eye movements and blinks were determined by visual inspection, and data were reconstructed using the components excluding the eye-related components.

To analyze data where participants appropriately responded to switches, we used the "*Good*" trials for further analysis. To observe the effects of switch types on both EEG responses relative to pre-cue onset and feedback onset, we analyzed data separately per onset type. The data were split in epochs in a range [−500, 2,100 ms] and [−100, 600 ms] relative to the pre-cue and feedback onsets, respectively. Since differences in

TABLE 1 | The question items in the questionnaire survey.

Question item	Explanation
1. Pleasantness of the pressing sound	The sound when pressing the switch feels pleasant.
2. Pitch of the pressing sound	The sound when pressing the switch has a high pitch.
3. Loudness of the pressing sound	The sound when pressing the switch feels loud.
4. Response speed	The response of the switch feels quick.
5. Niceness of responses	The response of the switch feels nice.
6. Click feeling	The switch clicks when pressing.
7. Lightness	The switch feels light when pressing.
8. Operability	The operability of the switch feels good.
9. Suitability for gaming	The switch feels suitable for gaming purposes.
10. Comfortableness	The switch feels comfortable when pressing.
11. Reliability	The switch feels reliable to use.
12. Satisfaction	The switch is satisfying to use.
13. Attractiveness	The switch feels attractive.
14. Liking	The switch is to my liking.

the mean amplitude of the baseline period across conditions influence the post-stimulus ERPs, we employed a two-way repeated measures analysis of variance (ANOVA; channel × switch) for each type of dataset (i.e., pre-cue onset and feedback onset) to determine if the mean amplitude of the pre-stimulus time window in ERPs differed across conditions. Since Mauchly's sphericity test revealed a violation of the sphericity assumption in all main effects and interactions in both datasets ($p < 0.05$), the degrees of freedom were corrected by the Greenhouse-Geisser epsilon. No significant main effects and interactions were observed in the pre-cue onset [switch: $F_{(2.03, 48.79)} = 0.56$, $p = 0.58$, channel: $F_{(1.59, 38.05)} = 2.48$, $p = 0.11$, switch × channel: $F_{(1.80, 43.29)} = 1.03$, $p = 0.36$] and feedback onset dataset [switch: $F_{(2.11, 50.65)} = 0.32$, $p = 0.74$, channel: $F_{(1.54, 36.93)} = 0.24$, $p = 0.73$, switch × channel: $F_{(1.85, 44.50)} = 0.88$, $p = 0.42$]. The epoch trials were baseline-corrected using the pre-stimulus region [−100, 0 ms]. In each dataset, epochs including signals exceeding a range of ±80 μv were rejected from further analysis. We excluded one participant's data from the pre-cue onset dataset as the number of remaining trials was <20 in all switch conditions due to a low signal-to-noise (S/N) ratio in ERP data. To confirm whether the number of rejected trials differed significantly across switches, the number of rejected trials was submitted to a one-way repeated measures ANOVA. Since Mauchly's sphericity test revealed a violation of the sphericity assumption (pre-cue onset: $W = 0.0001$, $p < 0.01$, feedback onset: $W = 0.04$, $p < 0.01$), the degrees of freedom were corrected by the Greenhouse-Geisser epsilon. No significant main effect was observed [pre-cue onset: $F_{(1.34, 30.73)} = 0.06$, $p = 0.87$, feedback onset: $F_{(1.48, 35.62)} = 2.53$, $p = 0.11$]. To consider the time-series variation of EEG modulations by the implicit learning of differences in the mechanical parameters, the EEG trial data in sessions 1, 2, and 3 were categorized as session 1 and the remaining sessions (session 4 and 5) as session 2. Single participant ERP waveforms were obtained by averaging across trials per participant, session, and switch types in each dataset.

We identified the mean amplitudes of three types of ERP components: N1 (auditory feedback onset), P3 (auditory feedback onset and pre-cue onset), and late CNV (pre-cue onset). For ERP analysis, a recently reported statistical analysis can determine the detailed time windows when significant effects are observed using a permutation test (Maris and Oostenveld, 2007). However, in the current study, and from the perspective of ergonomics research, we adopted the calculation of mean amplitudes in a predefined time window for simplicity. The time window of these ERPs was determined based on grand-averaged data where all switches and sessions were included (Luck and Gaspelin, 2017). To calculate the mean amplitude of the feedback-locked N1, we chose a time window of ±20 ms centered on the peak latency of the grand-averaged ERPs (N1: [98, 138 ms]). Since the auditory N1 response shows fronto-central distribution, we selected the Fz channel for N1 analysis. For the pre-cue-locked and feedback-locked P3, time windows of ±100 ms and ±50 ms, respectively, were selected centered on the peak latency of the grand-averaged ERPs (pre-cue-locked: [298, 498 ms], feedback-locked: [292, 392 ms]). The length of these time windows was determined by visual inspection of the grand-averaged ERPs. It seems that a range of these time windows is not incompatible with previous studies (Schevernels et al., 2016) that analyzed the pre-cue-locked and positive feedback-locked P3 in a motor response task involving the motor preparatory process between the pre-cue and the cue signal (pre-cue-locked: [400, 600 ms], feedback-locked: [300, 450 ms]). The Pz channel was used for P3 analysis because the component distributed parietally (Krebs et al., 2013; Ma et al., 2014; Schevernels et al., 2016; Carsten et al., 2021). We used a time window between [1,500, 2,000 ms] for the late CNV relative to the pre-cue onset (Schevernels et al., 2016). Considering that the late CNV reflects motor preparation (Leuthold et al., 2004) and is considered to partially overlap with the early CNV, we normalized the late CNV amplitudes to early CNV amplitudes [1,000, 1,500 ms] using Equation (2):

$$CNV_{normalized} = CNV_{late} - CNV_{early} \qquad (2)$$

where CNV_{late} and CNV_{early} are mean amplitudes of the late and early CNV, respectively. The Cz channel where the CNV is dominantly observed (Verleger et al., 2000) was used for the calculation.

To contrast the our proposed CNV index with one used in neuroergonomics research (Aricó et al., 2016), we also examined the modulation of frontal theta power as an index of mental workload before pressing the switch, using the same time window as the late CNV. To this end, the mean power across the theta frequency band (4–8 Hz) at Fz in the time window during motor response preparation [1,500, 2,000 ms] was obtained using the Fourier transform with Hanning window tapering. The frontal theta power was converted to a decibel scale relative to the theta power in the prestimulus region [−500, 0 ms] using Equation (3):

$$frontal\_theta = 10 log_{10}(\frac{power_{post}}{power_{pre}}) \qquad (3)$$

where $power_{post}$ and $power_{pre}$ are the frontal theta power in [1,500, 2,000 ms] and [−500, 0 ms] relative to the pre-cue onset, respectively.

2.5. Statistical Analysis

Because of the possibility that switch positions in the upper/lower (W and X) or horizontal position (A, S, or D) (cf. **Figure 1B**) affected participants' task responses, we performed a paired t-test between switch positions (upper/lower or horizontal position). As PT values linearly increased from Switch 1 (Shortest) to Switch 5 (Longest), the mean PT values of the upper/lower switches (Switch 2, 4) and horizontal position (Switch 1, 3, 5) were almost identical. Since all parameters except PT are identical across switches, PT is the only factor that can change RTs when switches are pressed with the same RT and speed. Therefore, we did not expect a significant difference in the mean RTs between switches with approximately the same PT, unless participants' behaviors are modulated. Therefore, if a significant difference in RTs is detected depending on switch position, it is likely this modulated the participant's behavior. In that case, switch position may also have affected EEG results, therefore, this analysis was performed to investigate the possible confounder of position effects.

To evaluate the effect of switch types on participant behavior, the RTs of *Good* trials and the obtained points in the task were submitted to linear mixed effect models (LME). The model included switch types, sessions, and interactions as fixed effects. Participants were processed as a random effect. The model for RTs in the Wilkinson notation is $RT \sim switch + session + switch:session + (1 \mid participants)$ (Model 1). The same model was fitted using points in the sessions as a target variable.

To identify the effects of switch type on EEG responses, we fitted an LME per EEG response. An effect of RTs was also included for investigating effects of participants' performance on EEG responses. Participants were processed as a random effect. The model in the Wilkinson notation is $EEG \sim switch + session + RT + RT:session + switch:session + switch:RT + switch:session:RT + (1 \mid participants)$ (Model 2).

A questionnaire survey was administered to investigate whether it could detect differences in the psychological feeling of switch pressing. Switch rankings were submitted to LME analysis as a target variable per question item. Since participants' performances may have affected the detection of differences in psychological feelings across switch types, the model included switch types, RTs, and their interactions as fixed effects. Since the questionnaire survey was performed after all sessions, RTs were averaged across sessions for this analysis. Participants were processed as a random effect. The Wilkinson notation of the model is $ranking \sim switch + RT + switch:RT + (1 \mid participants)$ (Model 3). To control for the false discovery rate across question items, p-values were corrected using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). In both analyses of EEG responses and questionnaires, RT data were centered using the mean value.

For LME modeling, the R software (R Core Team, 2020) and an *lme4* package (Bates et al., 2015) were used. The type II Wald χ^2 tests were used to determine the significance of the fixed effects based on *car* packages (Fox and Weisberg, 2019). For the *post-hoc* tests, the multiple comparisons of switch types or sessions were performed based on the fitted LMEs with the Kenward-Roger method and the Tukey method to correct p-values for multiple comparisons, if necessary. For the significant interaction of RTs by sessions or switch types in LME analysis of EEG responses or questionnaire surveys, the significance of a coefficient of a RT trend per level of the factor and the differences in coefficients across levels were tested. All *post-hoc* tests were performed using the *emmeans* packages (Lenth, 2021).

3. RESULTS

3.1. Behavioral Data

To investigate whether switch positions significantly affected participant performance, a paired t-test was performed based on the mean RTs between switch positions (upper/lower or horizontal position). The result showed that the RTs of the upper/ lower switches were almost significantly longer than horizontally-positioned switches ($t = 1.97$, $p = 0.06$, Cohen's $d = 0.39$), suggesting that switch position affected the participants' behavioral performance. As this effect is not intrinsically related to the mechanical parameters of the switches, we decided to use only switches arranged in the same direction. To use as many switches with equally spaced PTs as possible, and considering that switches 2 and 4 have relatively small PT deviations from Switch 3 with a normal PT, horizontally arranged switches (i.e., Switch 1, 3, and 5) were submitted to further analysis.

The mean RTs of *Good* trials and mean points earned in each session across participants in the remaining switches are summarized in **Figures 3A,B** for both sessions, respectively. The effect of switch type on participants' behavioral performance was investigated using Model 1, which analyzed the mean RTs of *Good* trials and the mean points earned in sessions as a target variable, respectively, and switch types (switch: Switch 1, 3, and 5) and sessions (session: 1 and 2), and their interactions as fixed effects. For RTs, the type II Wald χ^2 test revealed a

FIGURE 3 | (A) The mean RTs of *Good* trials and **(B)** mean points per session and switch type.
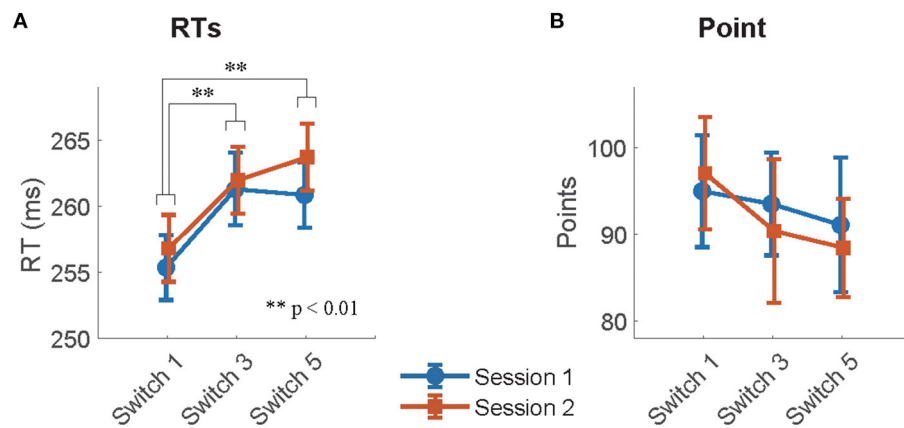
TABLE 2 | Estimated coefficients of fixed effects in LMEs for RTs and points.

Fixed effects	RTs		Points	
	Estimates	CI	Estimates	CI
Switch 1 (Intercept)	255.37	[250.50, 260.25]	95.01	[81.88, 108.14]
Switch 3	5.92	[2.33, 9.51]	−1.47	[−12.56, 9.62]
Switch 5	5.48	[1.89, 9.07]	−3.90	[−14.99, 7.19]
Session 2	1.45	[−2.14, 5.04]	2.07	[−9.01, 13.16]
Switch 3:Session 2	−0.77	[−5.85, 4.31]	−5.20	[−20.88, 10.48]
Switch 5:Session 2	1.40	[−3.68, 6.48]	−4.70	[−20.37, 10.98]

CI denotes a 95% confidence interval.

TABLE 3 | Summaries of the Wald χ^2 tests in the LMEs for RTs and points.

Fixed effects	RTs			Points		
	χ^2	df	p	χ^2	df	p
Switch	27.52	2	<0.01**	2.52	2	0.28
Session	2.47	1	0.12	0.14	1	0.71
Switch:Session	0.72	2	0.70	0.51	2	0.77

*df, degrees of freedom. **p < 0.01.*

significant effect of switch type [switch: $\chi^2_{(2)} = 27.52$, $p < 0.01$]. Multiple comparisons with *p*-value corrections based on the Tukey method revealed that the mean RTs of Switch 1 were significantly quicker than those of Switch 3 and 5 (Switch 1–Switch 3: $t = -4.19$, $p < 0.01$, Switch 1–Switch 5: $t = -4.68$, $p < 0.01$). On the other hand, no significant differences were observed between Switch 3 and 5 (Switch 3–Switch 5: $t = -0.49$, $p = 0.88$). The effect of sessions [session: $\chi^2_{(1)} = 2.47$, $p = 0.12$] and the interaction of switches by sessions [switch × session: $\chi^2_{(2)} = 0.72$, $p = 0.70$] did not reach significance. For task points, no fixed effect reached significance [switch: $\chi^2_{(2)} = 2.52$, $p = 0.28$, session: $\chi^2_{(1)} = 0.14$, $p = 0.71$, switch × session: $\chi^2_{(2)} = 0.51$,

$p = 0.77$]. LME analysis and Wald χ^2 tests for RTs and points are summarized in **Tables 2**, **3**.

3.2. Pre-cue Onset
3.2.1. P3
The summaries of LME fittings and Wald χ^2 tests for all EEG responses relative to pre-cue onsets are summarized in **Tables 4**, **5**, respectively. **Figure 4A** shows the grand-average ERP at Pz relative to the pre-cue onset per session and switch type. The effects of switch type, participants' performance (RTs), and sessions on mean P3 amplitudes were investigated using Model 2. The Wald χ^2 test showed that an almost significant interaction

TABLE 4 | Estimated coefficients of fixed effects in LMEs for EEG responses relative to pre-cue.

Fixed effects	P3		CNV		Frontal theta	
	Estimates	CI	Estimates	CI	Estimates	CI
Switch 1 (Intercept)	2.07	[1.14, 3.01]	−0.86	[−1.51, −0.21]	−0.60	[−1.27, 0.08]
Switch 3	−0.28	[−1.06, 0.51]	−0.06	[−0.88, 0.75]	0.40	[−0.50, 1.30]
Switch 5	−0.49	[−1.27, 0.29]	0.11	[−0.71, 0.92]	0.38	[−0.52, 1.27]
Session 2	−0.16	[−0.94, 0.62]	−0.17	[−0.99, 0.66]	0.98	[0.07, 1.89]
RT	0.00	[−0.05, 0.05]	0.04	[−0.01, 0.09]	0.01	[−0.04, 0.07]
Session 2:RT	−0.03	[−0.09, 0.03]	−0.03	[−0.09, 0.03]	−0.01	[−0.08, 0.06]
Switch 3:Session 2	0.50	[−0.59, 1.58]	0.49	[−0.65, 1.64]	−1.24	[−2.50, 0.02]
Switch 5:Session 2	0.32	[−0.77, 1.42]	−0.71	[−1.86, 0.44]	−1.05	[−2.32, 0.23]
Switch 3:RT	0.03	[−0.03, 0.09]	−0.04	[−0.10, 0.02]	0.00	[−0.07, 0.07]
Switch 5:RT	0.01	[−0.05, 0.07]	0.00	[−0.06, 0.07]	−0.02	[−0.09, 0.05]
Switch 3:Session 2:RT	0.03	[−0.05, 0.11]	0.09	[0.00, 0.18]	−0.05	[−0.15, 0.04]
Switch 5:Session 2:RT	0.05	[−0.04, 0.13]	0.02	[−0.07, 0.11]	−0.04	[−0.14, 0.06]

CI denotes a 95% confidence interval.

TABLE 5 | Summaries of the Wald χ^2 tests in the LMEs for EEG responses relative to pre-cue.

Fixed effects	P3			CNV			Frontal theta		
	χ^2	df	p	χ^2	df	p	χ^2	df	p
Switch	2.37	2	0.31	2.15	2	0.34	0.24	2	0.89
Session	0.57	1	0.45	0.55	1	0.46	0.42	1	0.52
RT	0.61	1	0.44	7.24	1	<0.01**	0.94	1	0.33
Session:RT	0.03	1	0.86	0.26	1	0.61	3.83	1	0.05[†]
Switch:Session	0.56	2	0.76	5.55	2	0.06[†]	3.65	2	0.16
Switch:RT	5.12	2	0.08[†]	0.64	2	0.73	2.30	2	0.32
Switch:Session:RT	1.19	2	0.55	4.19	2	0.12	1.23	2	0.54

*df, degrees of freedom. [†]p < 0.10, **p < 0.01.*

of switch types by RTs [switch × RT: $\chi^2_{(2)} = 5.12$, $p = 0.08$]. The *post-hoc* tests revealed that the coefficient of an RT trend did not reach significance for any of the switches (Switch 1: coeff. = −0.02 ± 0.02, $t = −0.72$, $p = 0.47$; Switch 3: coeff. = 0.03 ± 0.02, $t = 1.54$, $p = 0.13$, Switch 5: coeff. = 0.01 ± 0.02, $t = 0.67$, $p = 0.50$). In addition, there were no significant differences in the coefficients of RT trends across the switches (Switch 1–Switch 3: $t = −2.14$, $p = 0.09$, Switch 1–Switch 5: $t = −1.34$, $p = 0.38$, Switch 3–Switch 5: $t = 0.77$, $p = 0.72$). Other fixed effects did not reach significance in LME analysis (see **Table 5** for details). The mean P3 amplitudes by switch and session are depicted in **Figure 4B**.

3.2.2. CNV
The grand-average CNV at Cz per session and switch type is shown in **Figure 4C**. The results of the Wald χ^2 test showed that a significant effect of RTs [RT: $\chi^2_{(1)} = 7.24$, $p < 0.01$]. All coefficients of RT trends showed a positive relationship between CNV amplitudes and RTs (c.f. **Table 4**). Since the enhancement in mean CNV amplitudes shows more negative deflection, this positive correlation indicates that quicker RTs increase mean CNV amplitudes. An almost significant interaction of switch

types by sessions [Switch × session: $\chi^2_{(2)} = 5.55$, $p = 0.06$] was found. Multiple comparisons of mean CNV amplitudes across switch types per session revelated that larger mean amplitudes of Switch 5 (Longest PT) than those of Switch 3 (Normal PT) in session 2 (Switch 3–Switch 5: $t = 2.39$, $p < 0.05$). There were no other significant differences in session 1 (Switch 1–Switch 3: $t = 0.15$, $p = 0.99$, Switch 1–Switch 5: $t = −0.25$, $p = 0.97$, Switch 3–Switch 5: $t = −0.41$, $p = 0.91$) and session 2 (Switch 1–Switch 3: $t = −1.00$, $p = 0.58$, Switch 1–Switch 5: $t = 1.36$, $p = 0.36$). The p-values were corrected using the Tukey method. Other fixed effects did not reach significance in LME analysis (see **Table 5** for the details). The mean CNV amplitudes per switch and session are depicted in **Figure 4D**.

3.2.3. Frontal Theta
The mean frontal theta power at Fz per session and switch type is depicted in **Figure 4E**. The Wald χ^2 tests showed that the interaction of sessions by RTs was almost significant [session × RT: $\chi^2_{(1)} = 3.83$, $p = 0.05$]. The *post-hoc* tests showed that a coefficient of a RT trend in session 2 was similarly almost significant, but did not reach significance (session 1: coeff. = 0.01 ± 0.02, $t = 0.50$, $p = 0.62$, session 2: coeff. = −0.03 ± 0.02,
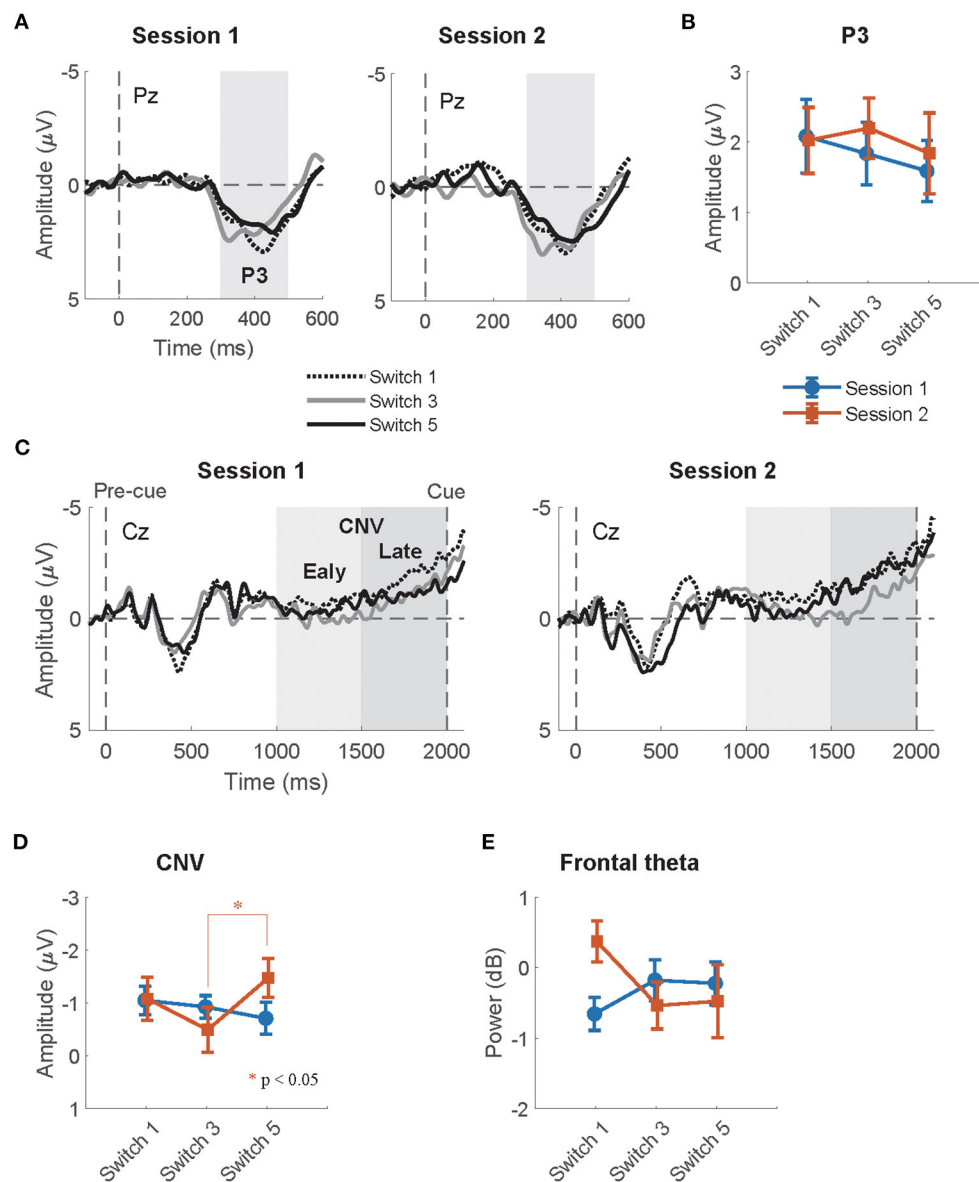
FIGURE 4 | (A) The grand-average ERP relative to pre-cue onsets at Pz per session and switch type. The highlighted time window was used for the calculation of the mean amplitudes of P3. **(B)** The mean amplitude of the P3 relative to pre-cue onsets at Pz per session and switch type. **(C)** The grand-average CNV waveforms relative to pre-cue onset at Cz per session and switch type. The time windows of the early and late CNV are highlighted in light and dark gray, respectively. **(D)** The mean amplitudes of the normalized CNV at Cz per session and switch type. **(E)** The mean frontal theta power (dB) at Fz per session and switch type.

$t = -1.90$, $p = 0.06$). The difference in the coefficient of RT trends between session 1 and 2 was almost significant (session 1–session 2: $t = 1.82$, $p = 0.07$). Any other fixed effects also showed no significant effects in LME analysis (see **Table 5** for details).

3.3. Feedback Onset
3.3.1. N1
The summaries of LME fittings and Wald χ^2 tests for all EEG responses relative to feedback onsets are summarized in **Tables 6**, **7**, respectively. The grand-average ERP at Fz relative to feedback onsets per session and switch type is shown in **Figure 5A**. The

Wald χ^2 tests showed a significant effect of sessions [session: $\chi^2_{(1)} = 5.84$, $p = 0.02$]. The *post-hoc* test showed significantly larger mean amplitudes in session 1 than in session 2 (session 1–session 2: $t = -2.17$, $p = 0.03$). Any other fixed effects showed no significant effects in LME analysis (see **Table 7** for details). The mean amplitudes per switch and session are depicted in **Figure 5B**.

3.3.2. P3
Figure 5C shows the grand-average ERP at Pz relative to the feedback onset per session and switch type. The Wald χ^2 test

TABLE 6 | Estimated coefficients of fixed effects in LMEs for EEG responses relative to feedback onset.

Fixed effects	N1		P3	
	Estimates	CI	Estimates	CI
Switch 1 (Intercept)	−3.54	[−4.61,−2.46]	2.91	[1.65, 4.17]
Switch 3	0.58	[−0.31, 1.46]	0.37	[−0.78, 1.51]
Switch 5	−0.18	[−1.06, 0.70]	−0.05	[−1.20, 1.09]
Session 2	0.40	[−0.47, 1.28]	0.18	[−0.96, 1.32]
RT	0.00	[−0.06, 0.06]	−0.01	[−0.09, 0.07]
Session 2:RT	−0.01	[−0.08, 0.06]	0.02	[−0.08, 0.11]
Switch 3:Session 2	−0.17	[−1.38, 1.04]	−0.35	[−1.93, 1.22]
Switch 5:Session 2	0.68	[−0.54, 1.90]	0.28	[−1.31, 1.87]
Switch 3:RT	−0.02	[−0.09, 0.05]	0.01	[−0.08, 0.09]
Switch 5:RT	0.00	[−0.07, 0.07]	0.00	[−0.09, 0.09]
Switch 3:Session 2:RT	0.03	[−0.06, 0.13]	−0.04	[−0.17, 0.08]
Switch 5:Session 2:RT	0.01	[−0.08, 0.11]	−0.04	[−0.17, 0.09]

CI denotes a 95% confidence interval.

TABLE 7 | Summaries of the Wald χ^2 tests in the LMEs for EEG responses relative to feedback onset.

Fixed effects	N1			P3		
	χ^2	df	p	χ^2	df	p
Switch	2.46	2	0.29	0.29	2	0.86
Session	5.84	1	0.02*	0.10	1	0.75
RT	0.02	1	0.89	0.37	1	0.54
Session:RT	0.22	1	0.64	0.23	1	0.63
Switch:Session	1.99	2	0.37	0.65	2	0.72
Switch:RT	0.30	2	0.86	0.53	2	0.77
Switch:Session:RT	0.41	2	0.81	0.57	2	0.75

*df, degrees of freedom. *$p < 0.05$.*

showed no significant results in any fixed effects (see **Table 7** for the details). The mean P3 amplitudes per switch and session are depicted in **Figure 5D**.

3.4. Questionnaire

The mean rankings per question item are summarized in **Figure 6**. To assess whether participants consciously perceived differences in the psychological perception of switches, the ranking data in the questionnaire survey were used to construct Model 3 per question item. The models included questionnaire rankings as a target variable, switch types, RTs and its interaction as fixed effects. The *p*-values of fixed effects were adjusted using the false discovery rate across question items according to the BH-method (Benjamini and Hochberg, 1995). Any fixed effects in every question item did not reach significance ($p > 0.05$). The results of the Wald χ^2 tests are summarized in **Table 8**.

4. DISCUSSION

The current research aimed to elucidate whether differences in the PTs of switches modulate the late CNV amplitudes

associated with motor preparation to press mechanical switches. This research was underpinned by the motivation to incorporate the allocated attentional resources to switch operation into the evaluation of mechanical switch parameters for gaming keyboards. We predicted that switches with PT values deviating from the normal range would increase the attentional resources allocated to switch operation, which would be reflected in a mean CNV amplitude increase (i.e., mean CNV amplitude of Switch 1, Switch 5 > Switch 3).

In line with our prediction, we observed that the mean amplitude of the normalized CNV in Switch 5 (longest PT) was significantly larger than that in Switch 3 (medium PT among the experimental switches studied herein) in the second half session. Given that attention increases correlate with the enhancement of late CNV amplitudes (McCallum and Walter, 1968; Tecce, 1972; Liebrand et al., 2017), the longest PT of Switch 5, which considerably deviates from that of commonly used switches, may have increased attention to the switch during motor preparation. The switch type effects on mean CNV amplitudes were observed only in the second half sessions. Participants might have gradually noticed the differences and
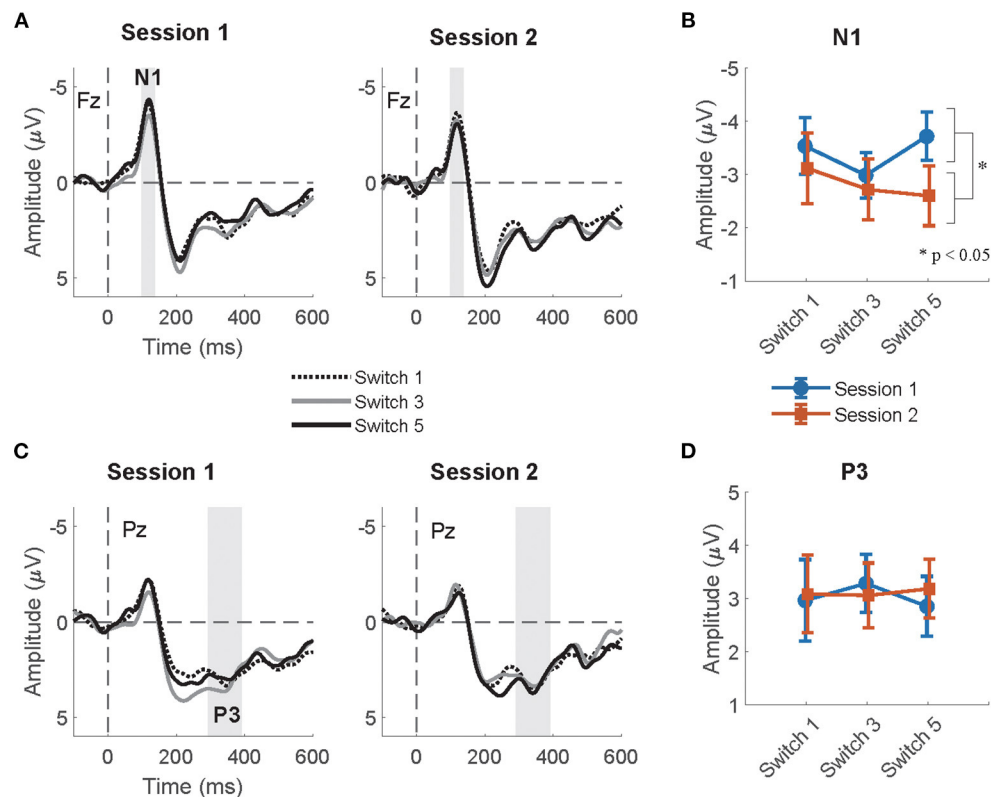
FIGURE 5 | (A) The grand-average ERP relative to feedback onset at Fz per session and switch type. The highlighted time window was used for the calculation of the mean amplitudes of N1. **(B)** The mean amplitude of the N1 relative to feedback onsets at Fz per session and switch type. **(C)** The grand-average ERP relative to feedback onset at Pz per session and switch type. The highlighted time window was used for the calculation of the mean amplitudes of P3. **(D)** The mean amplitude of the P3 relative to feedback onsets at Pz per session and switch type.

modulated their attention to switch operation as the sessions progressed. The results that the mean N1amplitude relative to auditory feedback, which is enhanced by top-down attention to auditory input (Hillyard et al., 1973; Näätänen and Picton, 1987), was significantly larger in session 1 than in session 2 might also suggest that the gap increased their attentions to the feedback at the beginning of the task. In addition, while other EEG responses (N1, P3, and frontal theta) to pre-cue and feedback onsets were also measured in the current research, these responses did not significantly differ between switches in *post-hoc* tests. Thus, among all EEG responses measured in the current research, CNV seems to be the most reliable feature for detecting increases in attention allocation to switch operations.

We predicted a significant difference between Switch 1 and 3, but the difference did not reach statistical significance. However, the mean CNV amplitude of Switch 1 tended to be larger than that of Switch 3. Since the RT of Switch 1 was significantly quicker than that of the other switches due to the short PT and RTs are correlated with sustained attention in reaction tasks (Buck, 1966), the attention increase due to PT deviation may have been smaller than that of Switch 5 because a faster RT with Switch 1 can be more easily obtained than with Switch 5. In addition, the small number of participants in the current study, which

resulted in a medium detectable effect size, may have contributed to the lack of a significant difference. Since we cannot confirm a difference between Switch 1 (shortest PT) and Switch 3 (normal PT) in the current study, an experiment with more participants is warranted.

The results of the questionnaire survey showed no significant effect of the type of switch on the subjective switch ranking. This suggests that PT differences and the corresponding increase in attentional resources cannot be extracted by questionnaire surveys, which supports the usefulness of brain activity characterization for marketing purposes. In particular, among the measures of brain activity used in neuromarketing, the use of EEG has an advantage in terms of evaluation in real-world environments. Indeed, EEG is preferred in various marketing conditions because its measurement systems have developed in recent years and permit wearable measuring devices that are compact and do not require application of conductive gels on the user's scalp (Higashi et al., 2017). To date, EEG measurements have been widely demonstrated to be a powerful tool with high temporal resolution for neuromarketing studies (Dmochowski et al., 2014; Golnar-Nik et al., 2019; Aldayel et al., 2020). As an extension of these EEG-based neuromarketing studies, the current research fulfilled the purpose of demonstrating that
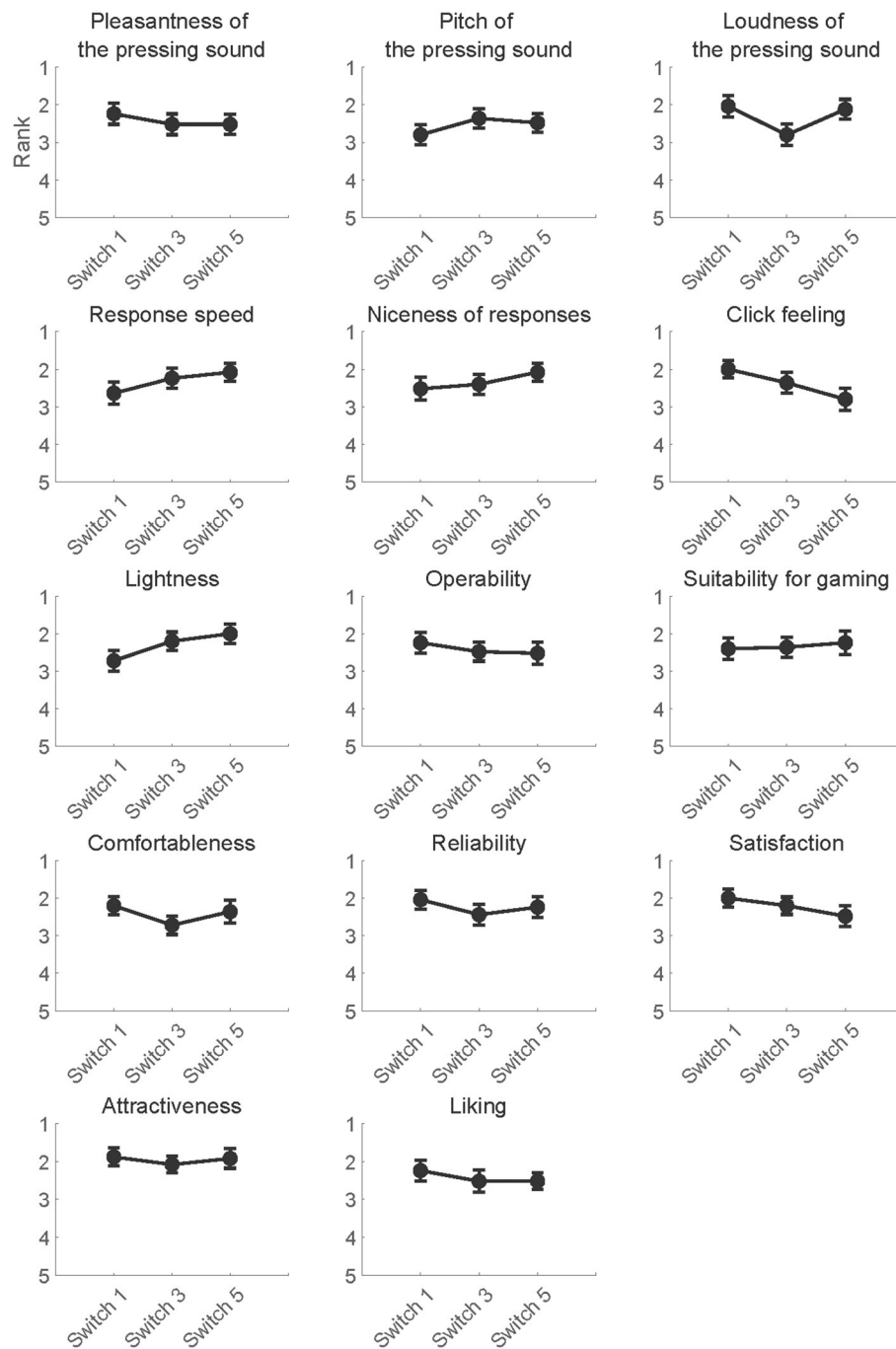
FIGURE 6 | The mean subjective rankings of switches per question item.

EEG measurements have the potential to provide a human-centric evaluation of the PTs of mechanical switches for gaming keyboards in a way that does not impose extra allocation of attentional resources to switch operation.

Since Switch 1 showed the shortest RT and it does not significantly increase the attentional resources compared to a switch with normal PT value (Switch 3), in terms of behavioral data and attentional resources, the switch with the shortest PT seems to be the best one in the current study. However, in game situations that require more complex switch operations than the current reaction task, switches with shorter PTs are more prone to operation errors because when a finger is placed on a switch, subtle finger movements unrelated to switch pressing are often misinterpreted as input. In addition, such characteristics

TABLE 8 | Summaries of the Wald χ^2 tests in the LMEs for question item in the subjective evaluation.

Question item	Switch			RT			Switch × RT		
	χ^2	df	p	χ^2	df	p	χ^2	df	p
Pleasantness of the pressing sound	1.30	2	0.79	1.21	1	0.64	4.40	1	0.37
Pitch of the pressing sound	1.07	2	0.79	1.13	1	0.64	2.83	1	0.48
Loudness of the pressing sound	6.42	2	0.32	1.29	1	0.64	0.99	1	0.85
Response speed	2.90	2	0.66	0.00	1	0.99	0.47	1	0.96
Niceness of responses	1.17	2	0.79	0.61	1	0.64	3.92	1	0.37
Click feeling	6.19	2	0.32	0.86	1	0.64	0.23	1	0.96
Lightness	3.98	2	0.64	0.96	1	0.64	1.91	1	0.67
Operability	0.45	2	0.86	0.33	1	0.72	6.39	1	0.29
Suitability for gaming	0.08	2	0.96	1.22	1	0.64	1.49	1	0.74
Comfortableness	3.02	2	0.66	0.77	1	0.64	6.43	1	0.29
Reliability	0.95	2	0.79	0.56	1	0.64	3.67	1	0.37
Satisfaction	1.85	2	0.79	0.03	1	0.93	0.04	1	0.98
Attractiveness	0.74	2	0.81	0.11	1	0.86	0.29	1	0.96
Liking	1.71	2	0.79	0.70	1	0.64	4.14	1	0.37

df, degrees of freedom. p-values were adjusted by the BH-method.

may also cause fatigue when gaming for long periods of time since users continue to pay attention to the switch so as not to produce errors. In the current study, we demonstrated that PT length induces an increase in the attentional resources allocated to switch pressing, and that there is no significant increase for a switch with short PT. On the other hand, the pursuit of RT alone does not necessarily lead to the development of optimal switches. Therefore, a multifaceted evaluation of mechanical switches, including the attentional resources allocated to switch operation in addition to RT, will lead to more human-centered product development. In the future, it may be possible to evaluate a newly developed switch in terms of the attentional resources allocated to its operation. Alternatively, when selecting which mechanical switch to use in a computer game, selecting that with the lowest CNV among several with different parameters may prevent the increase in the attention assigned to switch operations.

The current study has several limitations. First, during recruitment, the absence of motor disorders was not considered as selection criterion. Thus, if participants with a history of such disorders were involved in the current study, motor preparation might have been affected. Second, due to the low number of participants in the current study, only a medium effect size could be detected in case of a paired *t*-test. Thus, the lack of significant difference of mean CNV amplitudes between Switch 1 and 3 was probably due to the small number of subjects. Thus, in the future it will be necessary to investigate the switch effects on mean CNV amplitudes using a larger number of subjects.

5. CONCLUSIONS

In conclusion, the current research provides preliminary evidence that EEG is an effective evaluation criterion for mechanical parameters of switches in terms of the attentional resources allocated to switch operations. The results suggest our

method to evaluate mechanical parameters can help prevent increasing attention allocation to switch operation. The wireless wearable EEG measurement system with few electrodes which we employed may be useful for parameter design in real-world environments. Further investigations using a larger sample size and data collection from more switch types are necessary to confirm the applicability of the current method to the actual design of mechanical switches for gaming.

DATA AVAILABILITY STATEMENT

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethical Committee for Human and Animal Research of the National Institute of Information and Communications Technology. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

KF and KNakat developed the experimental switches. HW, KNakaj, ST, RM, and MS performed the data collection. HW and KNakaj analyzed the data. HW, KNakaj, ST, RM, MS, YY, HK, HN, and YN discussed the results. HW wrote the manuscript. All authors discussed the experimental design and data collection procedures, reviewed and approved the manuscript.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnrgo.2021.644449/full#supplementary-material

Supplementary Figure 1 | Representative parameters of the mechanical switches related to force required to press the switches and distances to move the switches from free positions (Stroke).

REFERENCES

Aldayel, M., Ykhlef, M., and Al-Nafjan, A. (2020). Deep learning for EEG-based preference cassification in neuromarketing. *Appl. Sci.* 10:1525. doi: 10.3390/app10041525

Aricó, P., Borghini, G., Di Flumeri, G., Colosimo, A., Bonelli, S., Golfetti, A., et al. (2016). Adaptive automation triggered by EEG-based mental workload index: a passive brain-computer interface application in realistic air traffic control environment. *Front. Hum. Neurosci.* 10:539. doi: 10.3389/fnhum.2016.00539

Ariely, D., and Berns, G. S. (2010). Neuromarketing: the hope and hype of neuroimaging in business. *Nat. Rev. Neurosci.* 11, 284–292. doi: 10.1038/nrn2795

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bazzani, A., Ravaioli, S., Faraguna, U., and Turchetti, G. (2020). Is EEG suitable for marketing research? A systematic review. *Front. Neurosci.* 14:594566. doi: 10.3389/fnins.2020.594566

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Buck, L. (1966). Reaction time as a measure of perceptual vigilance. *Psychol. Bull.* 65, 291–304. doi: 10.1037/h0023207

Carsten, T., Kostandyan, M., Boehler, C. N., and Krebs, R. M. (2021). Comparing the motivational value of rewards and losses in an EEG-pupillometry study. *Eur. J. Neurosci.* 53, 1822–1838. doi: 10.1111/ejn.15064

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Dmochowski, J. P., Bezdek, M. A., Abelson, B. P., Johnson, J. S., Schumacher, E. H., and Parra, L. C. (2014). Audience preferences are predicted by temporal reliability of neural processing. *Nat. Commun.* 5:4567. doi: 10.1038/ncomms5567

Eckner, J. T., Kutcher, J. S., and Richardson, J. K. (2010). Pilot evaluation of a novel clinical test of reaction time in national collegiate athletic association division I football players. *J. Athl. Train.* 45, 327–332. doi: 10.4085/1062-6050-45.4.327

Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Fox, J., and Weisberg, S. (2019). *An R Companion to Applied Regression, 3rd Edn.* Thousand Oaks, CA: Sage.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Pers. Individ. Differ.* 7, 385–400. doi: 10.1016/0191-8869(86)90014-0

Gaillard, A. W. K. (1977). The late CNV wave: preparation versus expectancy. *Psychophysiology* 14, 563–568. doi: 10.1111/j.1469-8986.1977.tb01200.x

Golnar-Nik, P., Farashi, S., and Safari, M. S. (2019). The application of EEG power for the prediction and interpretation of consumer decision-making: a neuromarketing study. *Physiol. Behav.* 207, 90–98. doi: 10.1016/j.physbeh.2019.04.025

Guo, F., Qu, Q. X., Nagamachi, M., and Duffy, V. G. (2020). A proposal of the event-related potential method to effectively identify Kansei words for assessing product design features in Kansei engineering research. *Int. J. Ind. Ergon.* 76:102940. doi: 10.1016/j.ergon.2020.102940

Hajcak, G., Holroyd, C. B., Moser, J. S., and Simons, R. F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology* 42, 161–170. doi: 10.1111/j.1469-8986.2005.00278.x

Hajcak, G., Moser, J. S., Holroyd, C. B., and Simons, R. F. (2007). It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology* 44, 905–912. doi: 10.1111/j.1469-8986.2007.00567.x

Higashi, Y., Yokota, Y., and Naruse, Y. (2017). "Signal correlation between wet and original dry electrodes in electroencephalogram according to the contact impedance of dry electrodes," in *Conference of the IEEE Engineering in Medicine and Biology Society* (Jeju Island), 1062–1065. doi: 10.1109/EMBC.2017.8037010

Hillyard, S. A., Hink, R. F., Schwent, V. L., and Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science* 182, 177–180. doi: 10.1126/science.182.4108.177

King, M. F., and Bruner, G. C. (2000). Social desirability bias: a neglected aspect of validity testing. *Psychol. Mark.* 17, 79–103. doi: 10.1002/(SICI)1520-6793(200002)17:2<79::AID-MAR2>3.0.CO;2-0

Kiss, M., and Eimer, M. (2008). ERPs reveal subliminal processing of fearful faces. *Psychophysiology* 45, 318–326. doi: 10.1111/j.1469-8986.2007.00634.x

Krebs, R. M., Boehler, C. N., Appelbaum, L. G., and Woldorff, M. G. (2013). Reward associations reduce behavioral interference by changing the temporal dynamics of conflict processing. *PLoS ONE* 8:e53894. doi: 10.1371/journal.pone.0053894

Lenth, R. V. (2021). *Emmeans: Estimated Marginal Means, Aka Least-Squares Means.* R Package Version 1.5.5-1. Available online at: https://CRAN.R-project.org/package=emmeans

Leuthold, H., Sommer, W., and Ulrich, R. (2004). Preparing for action: inferences from CNV and LRP. *J. Psychophysiol.* 18, 77–88. doi: 10.1027/0269-8803.18.23.77

Liebrand, M., Pein, I., Tzvi, E., and Krämer, U. M. (2017). Temporal dynamics of proactive and reactive motor inhibition. *Front. Hum. Neurosci.* 11:204. doi: 10.3389/fnhum.2017.00204

Liu, J., Wu, G., Luo, Y., Qiu, S., Yang, S., Li, W., et al. (2020). EEG-based emotion classification using a deep neural network and sparse autoencoder. *Front. Syst. Neurosci.* 14:43. doi: 10.3389/fnsys.2020.00043

Luck, S. J., and Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology* 54, 146–157. doi: 10.1111/psyp.12639

Ma, Q., Meng, L., Wang, L., and Shen, Q. (2014). I endeavor to make it: effort increases valuation of subsequent monetary reward. *Behav. Brain Res.* 261, 1–7. doi: 10.1016/j.bbr.2013.11.045

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *J. Neurosci. Methods* 164, 177–190. doi: 10.1016/j.jneumeth.2007.03.024

McCallum, W. C., and Walter, W. G. (1968). The effects of attention and distraction on the contingent negative variation in normal and neurotic subjects. *Electroencephalogr. Clin. Neurophysiol.* 25, 319–329. doi: 10.1016/0013-4694(68)90172-7

Morin, C. (2011). Neuromarketing: the new science of consumer behavior. *Society* 48, 131–135. doi: 10.1007/s12115-010-9408-1

Mullen, T. R., Kothe, C. A., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., et al. (2015). Real-time neuroimaging and cognitive monitoring using wearable dry EEG. *IEEE. Trans. Biomed. Eng.* 62, 2553–2567. doi: 10.1109/TBME.2015.2481482

Näätänen, R., and Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an

analysis of the component structure. *Psychophysiology* 24, 375–425. doi: 10.1111/j.1469-8986.1987.tb00311.x

Nagamachi, M. (1995). Kansei engineering: a new ergonomic consumer-oriented technology for product development. *Int. J. Ind. Ergon.* 15, 3–11. doi: 10.1016/0169-8141(94)00052-5

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The Measurement of Meaning*. Urbana, IL: University of Illinois Press.

Pantazatos, S. P., Talati, A., Pavlidis, P., and Hirsch, J. (2012). Cortical functional connectivity decodes subconscious, task-irrelevant threat-related emotion processing. *Neuroimage* 61, 1355–1363. doi: 10.1016/j.neuroimage.2012.03.051

Plassmann, H., O'Doherty, J., and Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *J. Neurosci.* 27, 9984–9988. doi: 10.1523/JNEUROSCI.2131-07.2007

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Ramsøy, T. Z., Skov, M., Christensen, M. K., and Stahlhut, C. (2018). Frontal brain asymmetry and willingness to pay. *Front. Neurosci.* 12:138. doi: 10.3389/fnins.2018.00138

Sammer, G., Blecker, C., Gebhardt, H., Bischoff, M., Stark, R., Morgen, K., et al. (2007). Relationship between regional hemodynamic activity and simultaneously recorded EEG-theta associated with mental arithmetic-induced workload. *Hum. Brain Mapp.* 28, 793–803. doi: 10.1002/hbm.20309

Sargent, A., Watson, J., Ye, H., Suri, R., and Ayaz, H. (2020). Neuroergonomic assessment of hot beverage preparation and consumption: an EEG and EDA study. *Front. Hum. Neurosci.* 14:175. doi: 10.3389/fnhum.2020.00175

Scheffé, H. (1952). An analysis of variance for paired comparisons. *J. Am. Stat. Assoc.* 47, 381–400. doi: 10.1080/01621459.1952.10501179

Schevernels, H., Bombeke, K., Krebs, R. M., and Boehler, C. N. (2016). Preparing for (valenced) action: the role of differential effort in the orthogonalized go/no-go task. *Psychophysiology* 53, 186–197. doi: 10.1111/psyp.12558

Schevernels, H., Krebs, R. M., Santens, P., Woldorff, M. G., and Boehler, C. N. (2014). Task preparation processes related to reward prediction precede those related to task-difficulty expectation. *Neuroimage* 84, 639–647. doi: 10.1016/j.neuroimage.2013.09.039

So, W. K. Y., Wong, S. W. H., Mak, J. N., and Chan, R. H. M. (2017). An evaluation of mental workload with frontal EEG. *PLoS ONE* 12:e0174949. doi: 10.1371/journal.pone.0174949

Spence, C. (2019). Neuroscience-inspired design: from academic neuromarketing to commercially relevant research. *Organ. Res. Methods* 22, 275–298. doi: 10.1177/1094428116672003

Tecce, J. J. (1972). Contingent negative variation (CNV) and psychological processes in man. *Psychol. Bull.* 77, 73–108. doi: 10.1037/h0032177

Thurstone, L. L. (1927). The method of paired comparisons for social values. *J. Abnorm. Soc. Psychol.* 21, 384–400. doi: 10.1037/h0065439

Verleger, R., Wauschkuhn, B., van der Lubbe, R., Jaśkowski, P., and Trillenberg, P. (2000). Posterior and anterior contribution of hand-movement preparation to late CNV. *J. Psychophysiol.* 14, 69–86. doi: 10.1027/0269-8803.14.2.69

Walter, W. G., Cooper, R., Aldridge, V. J., McCallum, W. C., and Winter, A. L. (1964). Contingent negative variation: an electric sign of sensori-motor association and expectancy in the human brain. *Nature* 203, 380–384. doi: 10.1038/203380a0

Wang, X. W., Nie, D., and Lu, B. L. (2014). Emotional state classification from EEG data using machine learning approach. *Neurocomputing* 129, 94–106. doi: 10.1016/j.neucom.2013.06.046

Winkielman, P., Berridge, K. C., and Wilbarger, J. L. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Pers. Soc. Psychol. Bull.* 31, 121–135. doi: 10.1177/0146167204271309

Wu, Y., and Zhou, X. (2009). The P300 and reward valence, magnitude, and expectancy in outcome evaluation. *Brain Res.* 1286, 114–122. doi: 10.1016/j.brainres.2009.06.032

Yeung, N., and Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *J. Neurosci.* 24, 6258–6264. doi: 10.1523/JNEUROSCI.4537-03.2004

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact