Title
Feasibility of Virtual and Augmented Reality Devices as Psychology Research Tools: A Pilot Study

Author
Garduno Luna, Cristopher Daniel

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Santa Barbara

Feasibility of Virtual and Augmented Reality Devices as Psychology Research Tools: A Pilot Study

A Thesis submitted in partial satisfaction of the
requirements for the degree Master of Arts
in Dynamical Neuroscience

by

Cristopher Daniel Garduno Luna

Committee in charge:

Professor Barry Giesbrecht, Chair

Professor Tobias Höllerer

Professor Scott Grafton

December 2020

The thesis of Cristopher Daniel Garduno Luna is approved.

Scott Grafton

Tobias Höllerer

Barry Giesbrecht, Committee Chair

December 2020

Abstract

Feasibility of Virtual and Augmented Reality Devices as Psychology Research Tools: A Pilot Study

by

Cristopher Daniel Garduno Luna

The recent proliferation of VR and AR devices has led to an increase in the use of these devices as research tools. As these technical developments continue, researchers can leverage these hardware improvements to create realistic and controlled environments for experimentation in life-like scenarios. In cognitive research, these devices will often be coupled with neurophysiological recordings, which poses the challenge of dealing with movement artifacts. In this study, three experiments were conducted using oddball tasks and semantic processing tasks to assess EEG data quality using VR/AR to display stimuli. The first experiment showed that the VR oddball task elicited comparable neural activity as a traditional desktop oddball task. The subsequent experiments systematically introduced movement artifacts in VR and AR, and showed that these neural data were usable with minor movement artifacts, while neural signals recorded under walking/free motion conditions were heavily contaminated with movement artifacts. Although there have been a variety of approaches for removing movement artifacts from neural data, many of them are specific to the experimental design, or have other constraints limiting their generalizability.

Introduction

The proliferation of virtual and augmented reality (VR/AR) technologies has made VR and AR tools with applications in varied fields, including behavioral and cognitive research applications. Traditional experimentation in these fields of research have often used simplified and highly controlled environments that have struggled to reflect the environments in which humans experience the behavioral or cognitive phenomena in question. With the goal of understanding these phenomena as they occur in realistic environments, various labs have turned to VR and AR environments. The Virtual Environment Navigation Laboratory (VENLab) at Brown University and the CAVE™from the University of Illinois at Chicago are two examples of VR and AR environments aimed to be as life-like as possible (Tarr and Warren, 2002). Although these two examples employed different systems to achieve their goal, the VENLab used a tethered head-mounted display (HMD) in a 40' by 40' room while the CAVE™used a projection-based system with 3D glasses in a 9' by 9', both of these systems were still relatively inaccessible and suffered from graphical and spatial limitations.

As these technologies continued to develop and increase their technological capacities we've seen more research groups investigate these tools as a viable option for life-like research environments (Wilson and Soranzo, 2015). Applications have ranged from using VR as an educational aid to using AR-based brain-computer interfacing (BCI) systems to control a mobile robot (Makransky et al., 2019; Si-Mohammed et al., 2020). Foerster et al. (2016) addressed a technical concern on the usability of a consumer-grade VR device as a research tool, and showed that the Oculus Rift DK2 (Oculus VR, Irvine, CA) is as reliable as a traditional CRT display for experiments assessing visual processing speed, threshold of conscious perception and capacity for visual working memory. Others have shown that virtual environments are capable of producing realistic physiological responses. Meehan et al. (2002) used a 'virtual pit' where users walked to a ledge and looked down into a large pit, and this reliably induced a stress response from users. Similarly, people completing a beam-walking task using the Oculus Rift DK2 were also shown to experience a stress response Peterson et al. (2018).

Another way to use these tools is to combine them with other research tools. Various groups have combined VR and AR technologies with electroencephalography (EEG) and other data acquisition tools to create BCI systems, and to examine behavioral and perceptual effects of these environments (Meehan et al., 2002; Kober and Neuper, 2012). With the addition of EEG to any VR/AR system, the need for users to walk quickly becomes a concern. EEG is very sensitive to noise and artifacts from the environment as well as from the users muscle movements, so this poses a challenge in building a system that allows for life-like environments and behaviors. Islam et al. (2016); Rahman et al. (2019) review the various types of artifacts that typically emerge in EEG studies, and examine different methods and their efficacy in removing the various types of artifacts. While some approaches involve novel hardware designs to minimize movement artifacts, others have implemented multi-modal data acquisition systems to characterize and remove movement artifacts (Mihajlovic et al., 2014; Arad et al., 2018).

EEG has been an important tool for understanding cognitive states during VR/AR scenarios, however, given the various sources of noise (due to movements of various magnitudes and electrical noise), the present study aimed to determine the feasibility of acquiring reasonable EEG data in VR/AR. Three experiments (with small sample sizes) were conducted using variants of a classic EEG task (odd ball) in VR using an HTC Vive HMD, and an AR analogue of a EEG semantic processing task using a Magic Leap 1 HMD. The task variants aimed to introduce noise systematically to move towards application scenarios.

Experiment I

Aim

In this experiment, participants completed a 3-stimulus oddball task in which subjects were instructed to identify target stimuli in the presence of standard non-target stimuli and distractor stimuli. The aim in this experiment was to elicit and compare the P300 signal generated by rare visual stimuli using a desktop monitor and a VR HMD.

Methods

Participants

Three adult participants at University of California, Santa Barbara participated in this study: two males ages 26 and 21, one female age 23. All participants reported normal or corrected to normal vision.

Visual Stimuli

Visual stimuli consisted of 8-bit grey scale images of faces and cars originally obtained from the Max Planck Institute for Biological Cybernetics face database (Troje and Bülthoff, 1996). The image set consisted of twelve car and twelve face images, where half of the faces were left-facing (45°) and the other half were right-facing (45°). Additional filters were applied as described in Bullock et al. (2015).

Procedure and Design

Each participant completed two testing conditions (*Desktop* and *VR*) of an oddball task during a single session. In the *Desktop* condition, participants were seated 70 cm from the display (Dell UltraSharp 2408WFP 24-inch widescreen LCD monitor, 1920x1200 resolution, 60 Hz refresh rate). In the *VR* condition stimuli were presented on a grey frame background using an HTC Vive HMD that uses an OLED panel for each eye (1080x1200 resolution each), 110° field of view, and 90 Hz refresh rate. Participants remained seated while using the HMD. Responses were recorded using the HTC Vive controllers for both conditions - left trigger: initiate block; right trigger: respond to target. In both display conditions were presented on a grey frame background and subtended ∼8.2°x8.2° of visual angle.

Prior to data collection, participants were given verbal instructions and a brief practice session (see Figure 1). The task consisted of 5 blocks of 200 trials for each condition (1000 trials total), and the order of stimulus presentation was randomized for each subject. The proportions of stimuli were 80% standard non-targets (cars), 10% distractor non-targets (left-facing faces), and 10% targets (right-facing faces). Upon initiating a block, stimuli were presented at fixation for 200 ms, followed by an inter-stimulus interval (ISI) of 800 ms ±200 ms. Participants were encouraged to take brief breaks between blocks. Figure 2 illustrates an example of the sequence of trials presented.

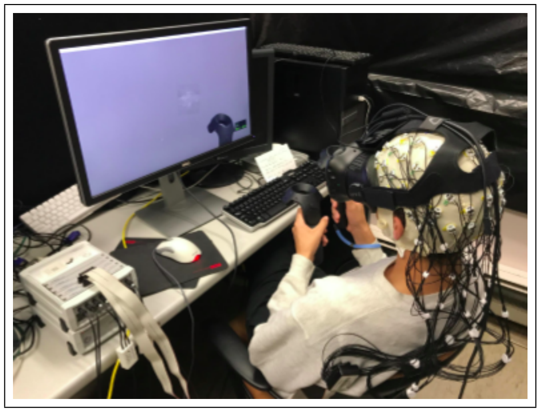

Figure 1: An example of a participant familiarizing themselves with the testing equipment and protocol during the practice task. Note: room lights were turned off during data collection.

EEG Data Acquisition

EEG data were recorded using a Brain Products ActiCHamp system (Brain Vision LLC, Morrisville, NC) with 64 electrodes in a actiCAP cap and arranged according to the 10-20 system. The TP9 and TP10 electrodes were placed on the right and left mastoids (average mastoid signal used as reference during data collection). Prior to starting each session, impedance was kept below 15 kΩ for all electrodes. The data were sampled at 1000 Hz.

Data Processing & Analysis

MATLAB (version 2018b, Massachusetts, The MathWorks, Inc., Natick, MA) and the EEGLAB toolbox (Delorme and Makeig, 2004) were used for offline EEG data processing. The continuous data were re-referenced to the average reference. Then the data were band pass filtered from 0.1 Hz to 30 Hz. The data were then epoched from -200 ms pre-stimulus to 1000ms post-stimulus. Electrodes Fp1 and Fp2 had poor quality data and were identified via visual inspection and were interpolated using data from surrounding electrodes. Eye blink correction was done

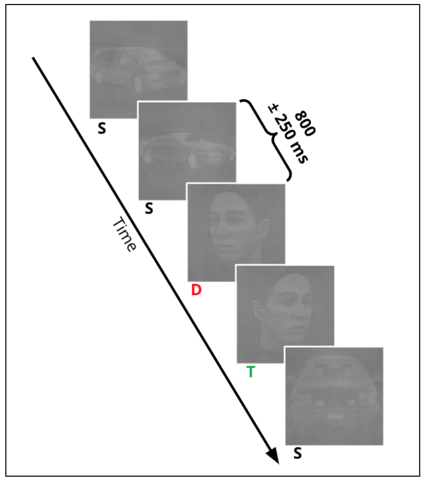

Figure 2: Experiment I oddball task trial sequence example. S: standard non-target (vehicles); D: distractor non-target (left-oriented faces); T: target (right-oriented faces). This figure shows an example of the stimuli presented in sequential trials; stimulus duration: 200 ms, ISI: 800 ± 250 ms.

using a conventional recursive least squares regression method from EEGLAB (crls_regression.m). Trials exceeding $\pm 150\mu V$ were excluded.

Event-related potentials (ERPs) were computed by averaging over the P3, Pz, and P4 electrodes for each subject then averaging across subjects. Figure 3 shows the ERPs along with the corresponding scalp topography averaged over 400-700 ms (set as the P300 time window). ERP differences were computed by subtracting the data standard (Std) stimuli ERPs from the distractor (Dis) or target (Tar) stimuli ERPs. Figure 4 shows the ERP differences along with the corresponding scalp topography differences.

Statistical Analysis. Given the limited sample size in this experiment ($n = 3$), t-tests were conducted across blocks for all subjects such that $n = 15$ (3 subjects, 5 blocks each). This approach allowed us to conduct significance testing, with an added assumption of equal variance across subjects. ERP difference waveforms were averaged over $400 - 700ms$ to compute the t statistics.

Results

ERPs and scalp topographies are shown in Figure 3 and Figure 4. The t-test results are presented in Figure 5. The P300 signal elicited from target and distractor stimuli was distinguishable from the signal elicited by standard stimuli in both display conditions. The signal elicited from target stimuli was statistically different from distractor stimuli in the desktop display condition, but not in the VR display condition. See Figure 5 for full results.

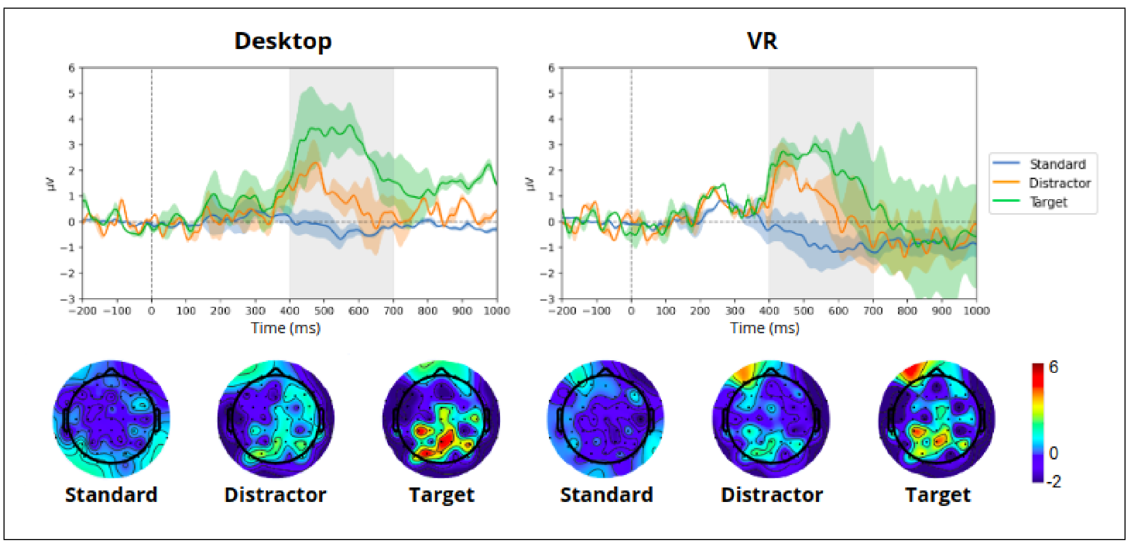

Figure 3: 3-Stimulus Oddball task ERPs with standard error (P3, Pz, P4) and scalp topography (400-700 ms). The grey shaded regions in the ERP plots indicate the region over which the data were averaged to compute the scalp topographies and for significance testing.

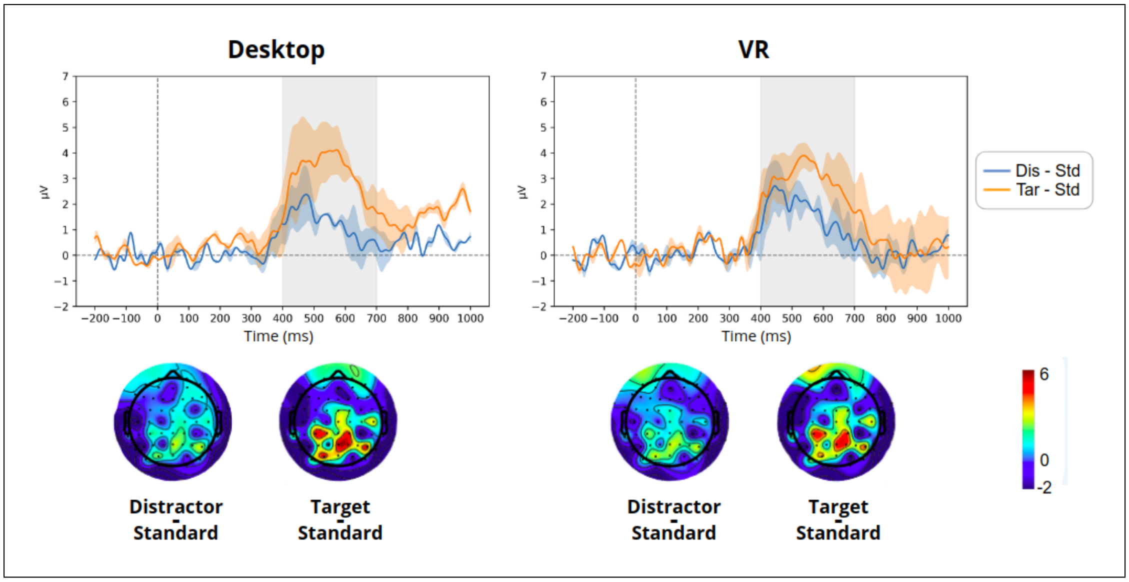

Figure 4: 3-Stimulus Oddball task ERP difference waveforms with standard error (P3, Pz, P4) and scalp topography (400-700ms).

A)

DIS - STD	Screen	VR
M (μV)	1.26	1.91
SD (μV)	1.49	1.74
t	3.35	4.11
p	< 0.01	< 0.001
df	14	14

B)

TAR - STD	Screen	VR
M (μV)	2.60	2.33
SD (μV)	2.15	2.20
t	4.81	3.97
p	< 0.001	< 0.001
df	14	14

C)

TAR - DIS	Screen	VR
M (μV)	1.34	0.42
SD (μV)	2.03	2.38
t	2.46	0.66
p	< 0.05	0.26
df	14	14

Figure 5: Significance testing across blocks for all subjects. $\boldsymbol{M}$ was computed by averaging the ERP difference waveforms over $400 - 700$ ms. Distractor-standard (**A**) stimuli, target-standard (**B**) stimuli, and target-distractor (**C**) stimuli.

Summary

The aim of experiment I was to elicit and compare the P300 signal using a desktop and VR display. ERP results showed a robust P300 signal and were generally consistent across display devices, with the exception being that target and distractor stimuli did not elicit statistically different P300 signal using the VR display. Scalp topographies were consistent across display devices and generally matched the expected topography of a P300 signal.

Experiment II

Aim

In this experiment participants completed a 3-stimulus and 2-stimulus oddball task. The 2-stimulus oddball task was aimed at reducing task difficulty and making the P300 signal more pronounced. All of the stimuli were presented using a VR headset. The aim of this experiment was to systematically introduce small predictable movements of the eyes and head, and observe how these movements affect the data quality.

Methods

Participants

Three adult participants at the University of California, Santa Barbara participated in this study: two males ages 26 and 21, one female age 23. Participants reported normal or corrected to normal vision. All three participants completed the *Fixation* testing condition, but only one participant (male, age: 26) completed the *Eye Motion* and *Head Motion* conditions.

Visual Stimuli

Visual stimuli in the 3-stimulus task were the same as described in experiment 1, and were all presented inside a blank virtual environment with a grey frame background. In the 2-stimulus task the distractor non-targets were replaced with targets such that 80% of the stimuli were standard non-targets, and 20% of the stimuli were targets (oddballs). The Gaussian white noise added by Bullock et al. (2015) was removed in the 2-stimulus task. The aim in removing the noise and distractor non-targets was to accentuate the P300 signal.

Procedure and Design

For both the 3- and 2-stimulus tasks, there were 3 conditions: *Fixation*, *Eye Motion/Movement*, *Head Motion/Movement*. In the *Fixation* condition all stimuli were presented at fixation. In the *Eye Movement* condition the stimuli were presented along an invisible circle 12° viewing angle away from fixation such that the images were completely within the field of view, but viewing the images required that the participant move their eyes toward the stimuli. In the *Head Movement* condition the stimuli were presented along the border of the field of view such that the participant required a head movement to view the entire image.

The 3- and 2-stimulus tasks were completed on separate days. The three testing conditions for each task were completed on the same day. Each condition consisted of 5 blocks of 200 trials each, for a total of 1000 trials per condition per task (6000 trials total over two days). The duration of each stimulus (and ISI) and the controller settings were the same as in experiment 1. Figure 6 shows an example of the stimulus presentation for the 3-stimulus oddball task.

EEG Data Acquisition

EEG data were recorded for each participant using a g.Nautilus Pro wireless system (gTec, Austria) consisting of 32 active gel-based electrodes arranged according to the 10-20 system. The ground electrode was located at AFz, and the reference electrode was clipped to the right earlobe. Prior to data collection, impedance for

Figure 6: Example of 3-Stimulus Oddball Task from experiment 2. Stimulus duration = 200 ms. ISI = 800 ± 250 ms. S: standard non-target; D: distractor non-target; T: target. Fixed: stimuli at fixation; Eye: stimuli just inside border of field of view; Head: stimuli just outside border of field of view.

all electrodes was < 30 kΩ, with the exception of FT10 and TP10, which were non-functional. The data were sampled at 500 Hz, and downsampled offline to 250 Hz.

Data Processing & Analysis

MATLAB and the EEGLAB toolbox (Delorme and Makeig, 2004) were used for offline data processing. The EEG data processing steps were as described in experiment I.

Statistical Analysis

As with experiment I, t-tests were conducted across blocks for all subjects. In this experiment, there were three subjects in the *Fixation* condition and one subject in the *Eye Movement* and *Head Movement* conditions ($n_{fix} = 15$, $n_{eye} = n_{head} = 5$).

Results

The ERP and scalp topographies are shown in Figure 7 and Figure 8 for the 3-stimulus oddball task, and Figure 9 and Figure 10 for the 2-stimulus oddball task. The t-test results are shown in Figure 11 for the 3-stimulus oddball task. These results showed that the P300 signals elicited by target and distractor stimuli were statistically different from the signal elicited by standard stimuli, but the P300 signals elicited by the target and distractor stimuli were not statistically different from each other. See Figure 11 for full results.

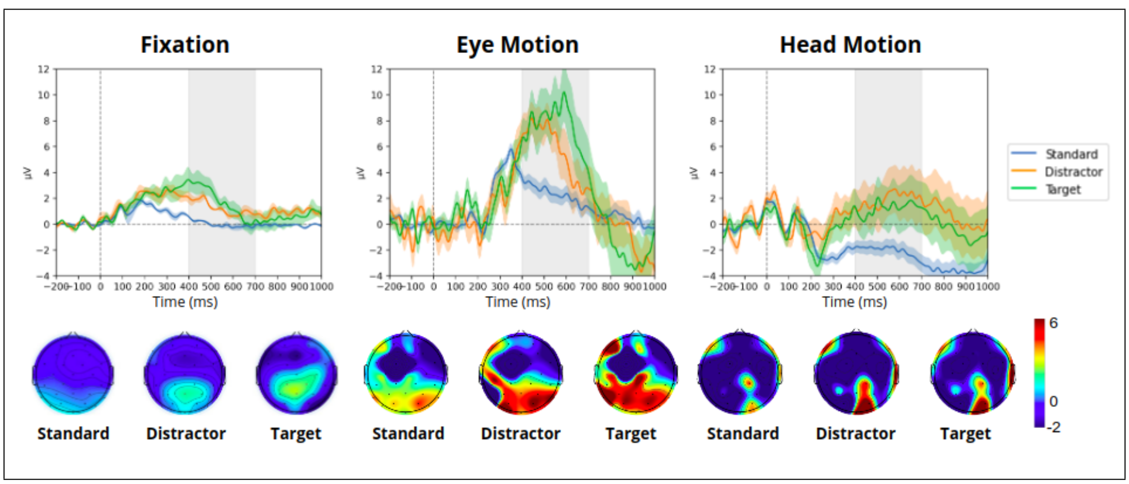

Figure 7: 3-Stimulus Oddball task ERPs with standard error (P3, Pz, P4) and scalp topography (400-700 ms). The grey shaded regions in the ERP plots indicate the region over which the data were averaged to compute the scalp topographies and for significance testing.

The t-test results for the 2-stimulus oddball task are shown in Figure 12. These

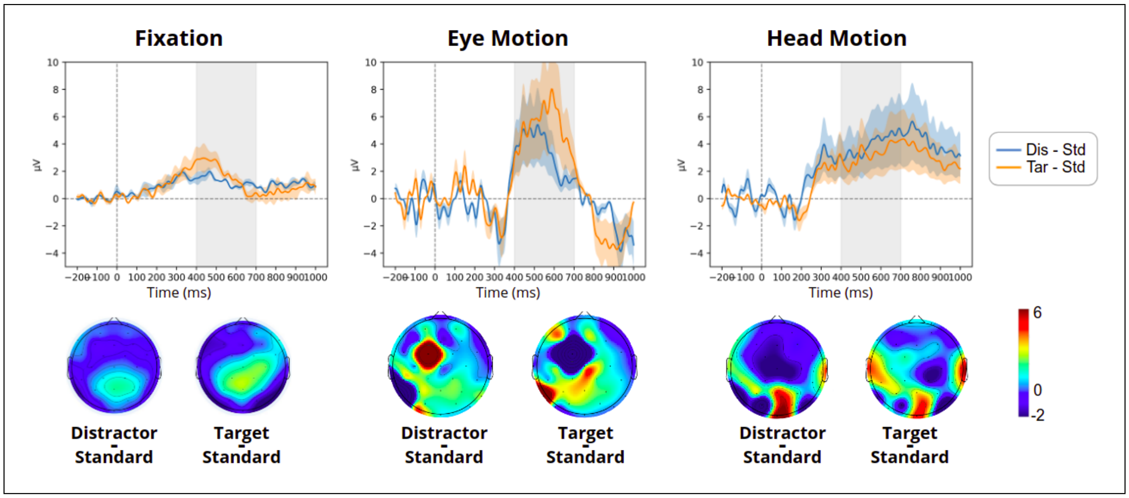

Figure 8: 3-Stimulus oddball task ERP difference waveforms with standard error (P3, Pz, P4) and scalp topography (400-700 ms).

results showed that the P300 signal elicited by target stimuli were statistically different from the signal elicited by standard stimuli only in the *Eye Movement* condition.

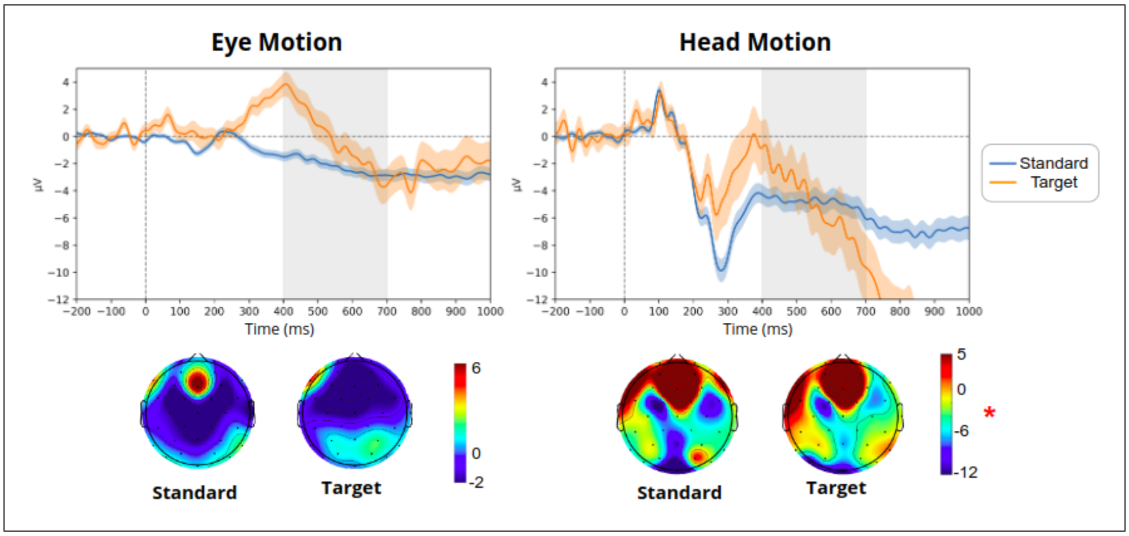

Figure 9: 2-Stimulus oddball task ERPs with standard error (P3, Pz, P4) and scalp topography (400-700 ms). Note: the scalp topography scale in the *Head Motion* condition was adjusted due to high amplitude deflections.

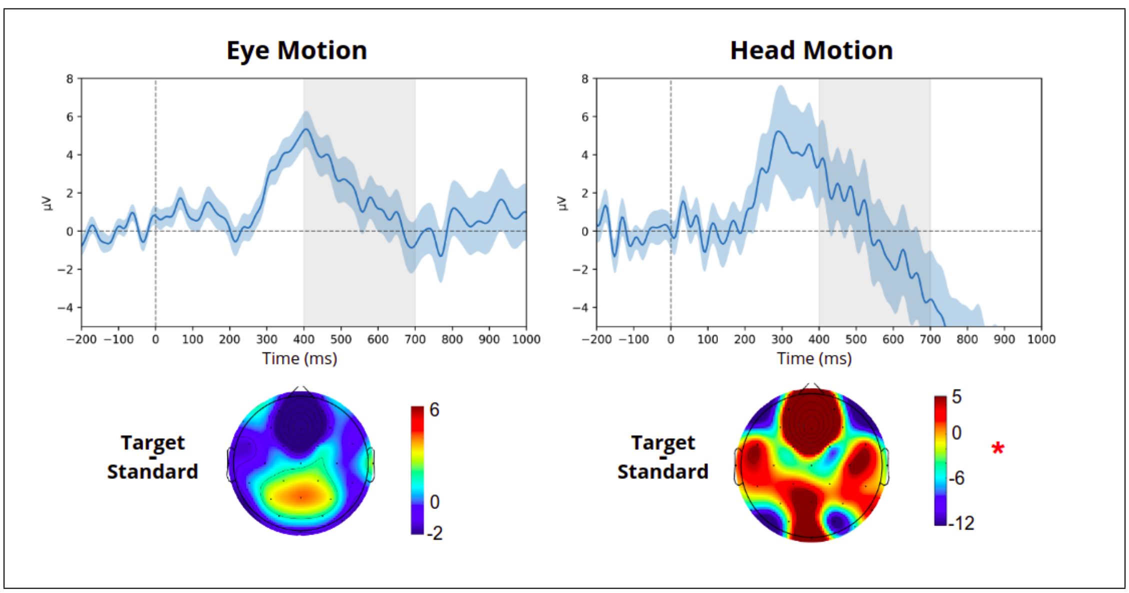

Figure 10: 2-Stimulus oddball task ERP difference waveforms with standard error (P3, Pz, P4) and scalp topography (400-700 ms). Note: The scalp topography scale in the *Head Motion* condition was adjusted due to high amplitude deflections.

A) Fixation	*M (μV)*	*SD (μV)*	*t*	*p*	*df*
TAR - STD	1.72	1.72	3.76	< 0.01	14
DIS - STD	1.27	0.58	8.22	< 0.001	14
TAR - DIS	0.46	1.37	1.25	0.12	14

B) Eye Movement	*M (μV)*	*SD (μV)*	*t*	*p*	*df*
TAR - STD	5.26	2.38	4.42	< 0.01	4
DIS - STD	3.56	2.68	2.66	< 0.05	4
TAR - DIS	1.70	3.07	1.11	0.165	4

C) Head Movement	*M (μV)*	*SD (μV)*	*t*	*p*	*df*
TAR - STD	3.27	1.17	5.59	< 0.01	4
DIS - STD	3.94	1.28	6.16	< 0.01	4
TAR - DIS	-0.67	1.00	-1.35	0.125	4

Figure 11: 3-Stimulus oddball task significance testing across blocks for all subjects. $\boldsymbol{M}$ was computed by averaging the ERP difference waveforms over $400 - 700$ms. Note: there were three subjects in the *Fixation* condition, and two in each of the other conditions.

TAR - STD	Eye Movement	Head Movement
M (μV)	2.14	-0.15
SD (μV)	1.71	2.00
t	2.49	-0.15
p	< 0.05	0.45
df	4	4

Figure 12: 2-stimulus oddball task significance testing across blocks for all subjects. $\boldsymbol{M}$ was computed by averaging the ERP difference waveforms over $400 - 700$ms.

Summary

The aim of this experiment was to systematically introduce small predictable movements while users completed tasks in VR, and to observe the effect of these motion artifacts on data quality. The P300 ERP component was robust for target and distractor stimuli (3-stimulus oddball task) in all movement conditions, although these movements distorted the waveforms, particularly with head movements. The scalp topographies were also distorted with these movements, and head movements added the most distortion. The P300 ERP component and scalp topography for the *Eye Movement* condition was as expected and comparable to the signal and topography at *Fixation*, but this was not the case for the *Head Movement* condition, which was likely due to the high amplitude motion artifacts.

Experiment III

Aim

In this experiment we compared a semantic processing ERP component (N400) using an LCD display and an Magic Leap 1 augmented reality (AR) headset. Similar to experiment 2, this experiment systematically introduced muscle movement artifacts with the aim of recovering the N400 signal with the presence of heavily artifact-contaminated data.

Methods

Participants

Twenty-four adult participants from the University of California, Santa Barbara (UCSB) community took part in this study primarily via SONA, but only six of those participants (three male, three female) were used in the analysis due to technical errors resulting in corrupted data. All participants provided informed consent and were compensated financially ($15/h). The mean age for participants used in the analysis was 20.7 years with a standard deviation of 4.2 years. All participants reported having normal or corrected to normal vision.

Visual Stimuli

The stimuli used in each condition were from the same stimulus list of 120 related word-object pairs. The words were obtained from a list of word-word pairs used in an prior study of semantic processing (Swaab et al., 2002). Each pair was presented two times: once with a congruent pairing and another time with an incongruent pairing, where the order in which these congruent/incongruent pairs appear is randomized. Words were presented in white upper-case sans serif font on a grey frame background.

Procedure and Design

Each participant completed all four testing conditions (*Desktop, AR-Fixed, AR-Head Movement, AR-Walking*) during a single session. Each participant was required to complete 2 blocks of 120 trials for each test condition (960 trials total) in which a word is shown, followed by the corresponding congruent/incongruent object pairing. Subjects were given up to five minutes of rest in between blocks and testing conditions. The order in which testing conditions were completed was randomized.

In the *Desktop* condition all stimuli were displayed using a 24 inch LCD monitor placed 60 cm away from the participant. In the three *AR-* conditions the

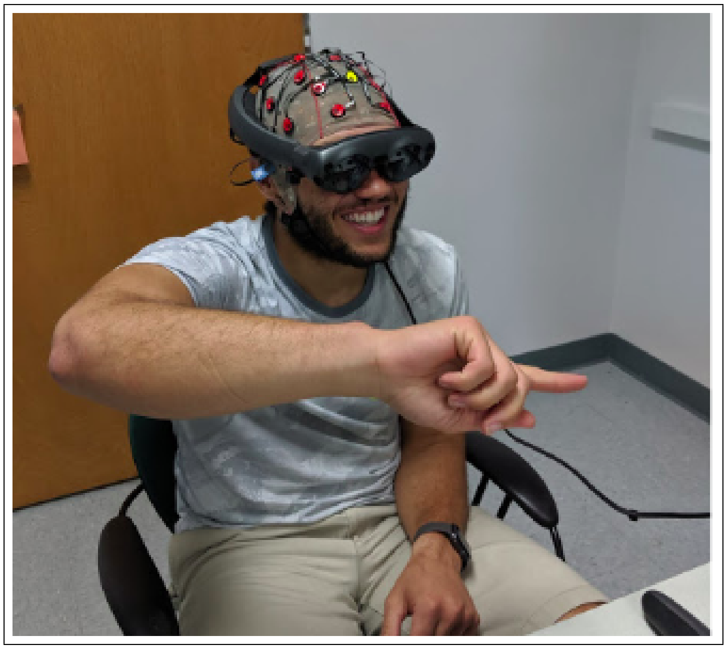

Figure 13: Example of a typical participant familiarizing themselves with the Magic Leap HMD.

stimuli were displayed using a Magic Leap 1 AR HMD.

Prior to starting data collection participants were taken to the testing room where the AR display was calibrated using Magic Leap's built-in eye calibration application, then they were given a brief practice version of the testing conditions with ten word-object pairs. Figure 13 shows an example of a participant familiarizing themselves with the AR HMD. In the *Desktop* condition participants were instructed to record their responses with keyboard response keys where "J" indicates a congruent pair and "F" indicates a incongruent pair. In the *AR-* conditions participants were instructed to record their responses using the Magic Leap 1 handheld controller touchpad where the right half of the touchpad indicates a congruent trial and the left half of the touchpad indicates an incongruent trial.

In the *Desktop* condition the participants were $\sim 70cm$ from the display. In the *AR-Fixed* condition the plane was set such that the stimuli were presented on a stand $\sim 140cm$ from the participant. In the *AR-Head Movement* condition the plane was set such that the *word* stimuli were presented on a stand $\sim 140cm$ from the participant, and *object* stimuli were presented randomly on a stand either $110cm$ to the left or $110cm$ to the right of the center stand. In the *AR-Walking* condition the plane was set on the center stand and participants were instructed to walk in a circular motion at their own pace around the stand while they completed the task. Participants were encouraged to switch the direction they walked in during this task to avoid neck discomfort from constantly facing in a single direction.

Task. The task was self-paced, and in order to initiate a block or trial participants simply press the *space* key in the *Desktop* condition or the trigger for the *AR-* conditions. One additional step in the *AR-* conditions was to select the surface plane for words and objects to appear - this was done by the researcher to ensure a standardized placement of words/objects. For the *Desktop*, *AR-Fixed*, and *AR-Walking* conditions the order of events are similar. The participant initiates a trial and a fixation cross appears either at the center of the screen (*Desktop*) or on the fixation surface at the center of the field of view (*AR-Fixed/Walking*) until the participant fixates on the cross for 1000 ±200 ms. The fixation cross is then removed and immediately followed by a word that is displayed at fixation for 1 second. The word is then removed, and after a 1 second delay (±200 ms, 0.8-1.2 second inter-stimulus interval) an object is presented at fixation and the participant indicates whether they think the word and object make a congruent or incongruent pair. Once the participant responds then the object is removed from the display and the participant starts the next trial at their own pace. The *AR-Head Movement condition* is designed to induce a head movement by simultaneously displaying an arrow when the object appears, but in this case the object is outside the field of view and the arrow indicates the direction that the participant's head must move in order to view the object. A key difference in the *AR-Walking* condition is that the participant is instructed to walk at their own pace around a stand where the stimuli are displayed. A built-in break in between blocks began after 120 trials and lasted up to five minutes. Participants were also encouraged to take the allotted 5 minute break in between conditions.

Hardware Setup. The hardware implementation consisted of three main components (shown in Figure 14): the main computer used to display in the *Desktop* condition and record the EEG data, the Magic Leap 1 HMD used for stimulus presentation and logging user data in the *AR-* conditions, and the EEG headset for recording neural data. Network latency between components was accounted for empirically by measuring the average round trip (10 trips) time of a very small package containing the timestamps of when the package was sent and received. The EEG system utilized Bluetooth communication to send the EEG data packets to the g.Nautilus Base Station, wired directly to the main computer. The stimulus display data was logged directly onto the Magic Leap 1 device and was merged & aligned with the EEG data offline using the main computer's Unix timestamp. See Figure 14 for a simplified visualization of the components.

EEG Data Acquisition

EEG data were recorded for each participant using a g.Nautilus RESEARCH wireless system (gTec, Austria) consisting of 64 active gel-based electrodes with

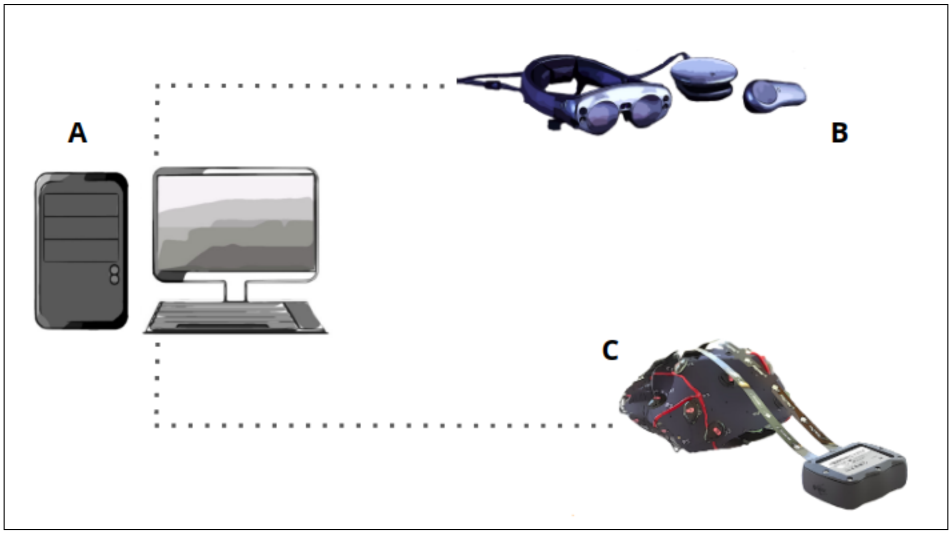

Figure 14: Hardware setup for experiment 3. The dotted lines indicate wireless communication. **A**) Main computer used to record the EEG data and display stimuli in the *Desktop* condition. **B**) Magic Leap 1 device used to display stimuli in the *AR*- conditions and log user responses. **C**) g.Nautilus Research EEG headset used for recording EEG data.

ground and reference electrodes (wireless transmission) arranged according to the 10-20 system. The ground electrode was located at AFz, and the reference electrode was located on the right earlobe using a built-in ear clip. Signa gel (Parker Laboratories, New Jersey) was used for all electrodes to keep the impedance below 30 $k\Omega$ using the internal impedance check. Although the analog to digital converter operates at 1024 Hz, the data were down sampled offline to 250 Hz.

Electrodes. Data were recorded from all 64 channels, but two channels were non-functional (TP10, FT10) and the data was discarded offline. Additional channels were removed due to poor quality data and are listed in the table below. In order to prevent unbalanced electrode removal from biasing the data towards either hemisphere the corresponding electrode on the opposite hemisphere was also removed. After removing the channels the missing data was interpolated using surrounding channels.

Channels Removed Per Subject	
Subject	Channels Removed
4	FT10, TP10
5	FT10, TP10, O1, O2
6	FT10, TP10, AF8, AF7, F7, F8, F3, F4, F6, F5, FC6, FC5, FT8, FT7
10	FT10, TP10, Cz, AF3, AF4, FPz, AF7, AF8
15	FT10, TP10, F1, F2, AF7, AF8
16	FT10, TP10, F1, F2, P9, P10

Table 1: These electrodes were removed from analysis in all conditions due to issues with poor data quality.

Data Processing

EEG Data Processing. MATLAB (version 2018b, Massachusetts, The MathWorks, Inc., Natick, MA) was used for offline EEG data processing, as well as the EEGLAB toolbox (Delorme & Makeig, 2004). The continuous data were bandpass filtered between 0.1 to 30 Hz to remove noise and irrelevant physiological signals. The data were epoched between -0.2 to 1.0 sec for the *Desktop*, *AR-Fixed*, and *AR-Walking* conditions, and between -0.2 and 1.5 sec for the *AR-Head Movement* condition. The data were then re-referenced using an average reference. Trials exceeding $\pm 100\mu V$ were removed from data in the *Desktop*, *AR-Fixed*, and *AR-Walking* conditions, and trials exceeding $\pm 150\mu V$ were removed from data in the *AR-Walking* condition. Contamination from eyeblinks was removed using a conventional recursive least squares regression method from EEGLAB (crls_regression.m).

The average ERP waveforms in all conditions were computed time-locked to the onset of the object stimulus. The *Desktop*, *AR-Fixed*, and *AR-Walking* conditions included a 200 ms pre-stimulus baseline and 1000 ms post-stimulus interval. The *AR-Walking* condition included a a 200 ms pre-stimulus baseline and 1500 ms post-stimulus interval. To present the N400 component more clearly the congruent ERP waveform was subtracted from the incongruent ERP waveform.

Statistical Analysis

The magnitude of the N400 was computed as the average amplitude of the difference waves in the N400 region. The N400 region in the *Desktop*, *AR-Fixed*, and *AR-Walking* conditions were defined as the region between 300-500 ms, and 800-1000 ms in the *AR-Head Movement* condition. The reason for the N400 region shift in the *AR-Head Movement* condition was that from prior pilot work it was observed that head movements in this scene had a duration of about 500 ms. The electrodes used to compute the N400 were FC1, FC2, FCz, C1, C2, Cz, CP1, CP2, and CPz. One-tailed t-tests were used for significance testing of the N400 component.

LDA Classifier. A leave-one-out linear discriminant analysis (LDA) classifier was used to classify the data. 1000 random permutations were generated for permutation testing, then the average value as well as 95th percentile values were computed for each time window to construct a mean accuracy waveform and 95th percentile waveform for the permuted data. Classifier performance values on real data exceeding the 95th percentile waveform are analogous to $p < 0.05$.

Results

The ERP grand averages and the ERP difference waveforms are shown in Figure 15 and Figure 16. In the *AR-Fixed* condition, the N400 signal elicited by incongruent trials had a larger amplitude ($M = -0.97$, $SD = 0.96$) than the signal elicited by congruent trials, $t(5) = -2.23$, $p = 0.038$. In each of the other conditions the observed differences were not statistically significant: *Desktop*: ($M = -0.61$, $SD = 0.72$) $t(5) = -1.90$, $p = 0.058$; *AR-Head Movement*: ($M = 1.34$, $SD = 1.62$) $t(5) = 1.84$, $p = 0.06$; *AR-Walking*: ($M = -1.48$, $SD = 4.50$), $t(5) = -0.74$, $p = 0.248$.

Figure 17 shows the scalp topography averages as well as their difference (incongruent - congruent). The *Desktop* and *AR-Fixed* conditions show comparable topographies that reflect the N400 ERP component. The *AR-Head Movement* and *AR-Walking* conditions had heavily distorted scalp topographies due to the movement artifacts, and had limited interpretability without further data cleaning.

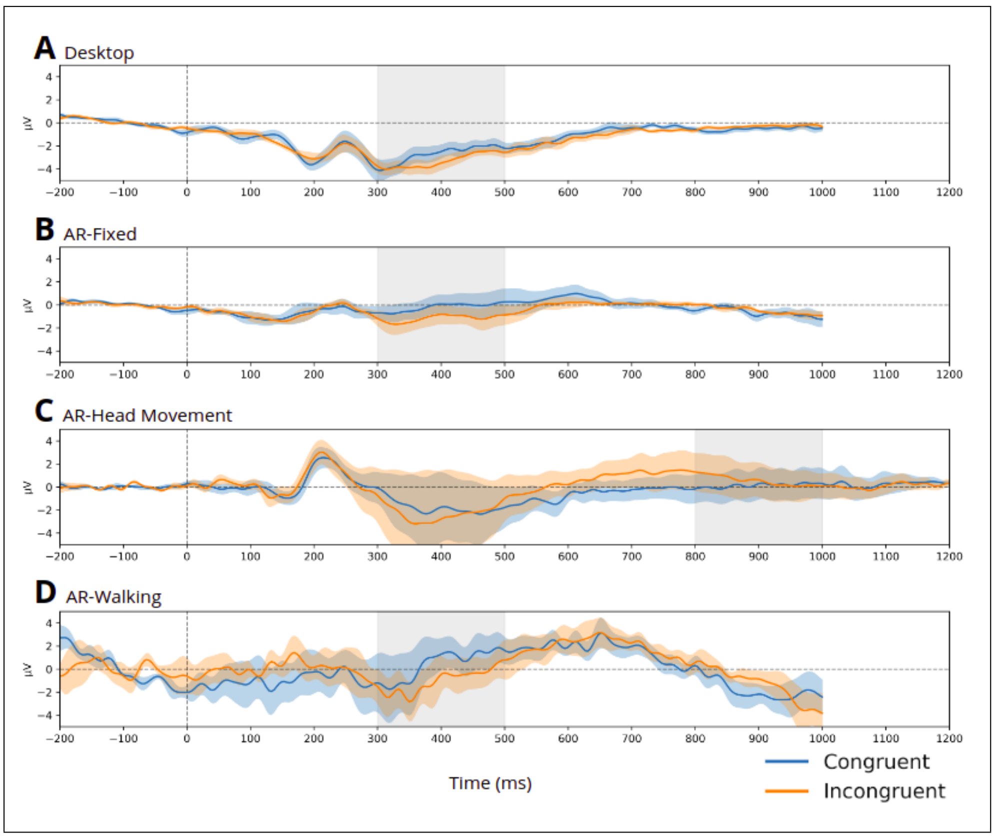

Figure 15: ERP grand averages with shaded standard error. See *Statistical Analysis* for electrodes. $n = 6$. Grey shaded region indicates the N400 region used for statistical analysis.

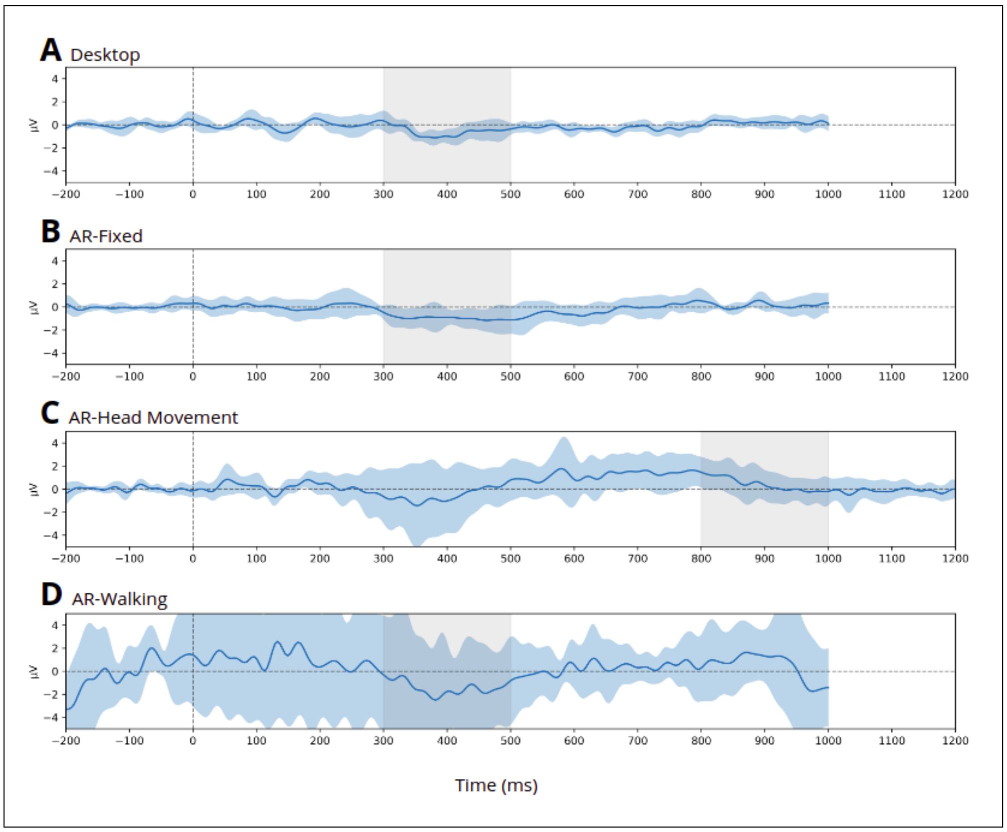

Figure 16: ERP grand average differences waveforms (incongruent - congruent) with shaded standard error. See *Statistical Analysis* for electrodes. $n = 6$.

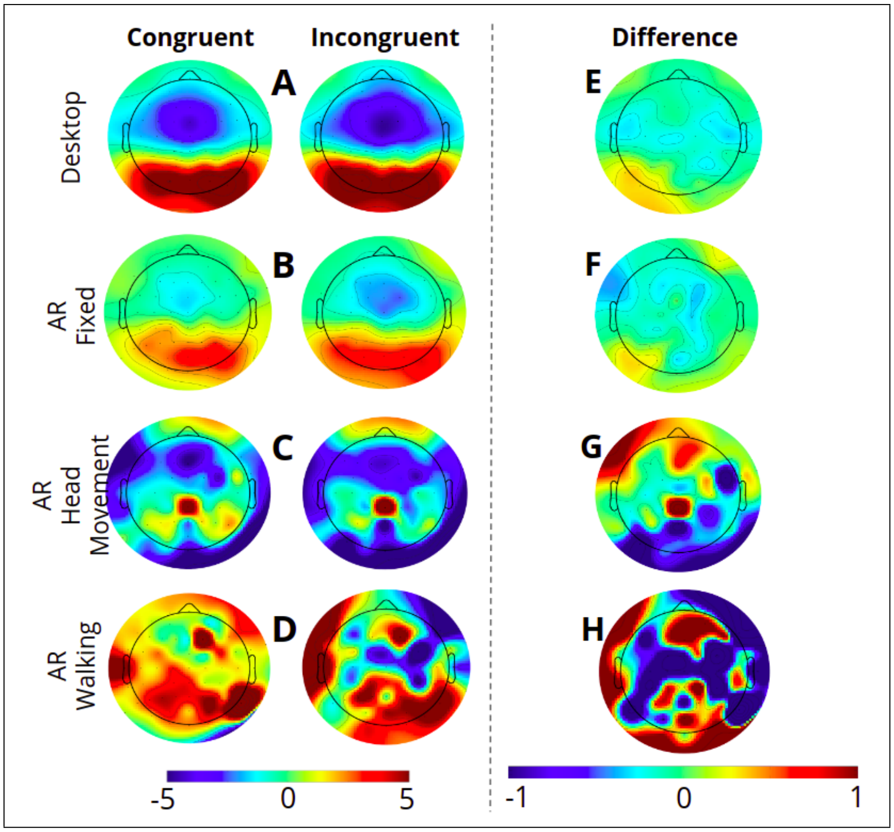

Figure 17: Scalp topography for congruent and incongruent word-object pairs (A-D), and incongruent - congruent differences (E-H). Electrodes used in analysis: FC1, FC2, FCz, C1, C2, Cz, CP1, CP2, CPz. $n = 6$. **A,E**) Desktop, 300-500ms **B,F**) AR-Fixed, 300-500ms **C,G**) AR-Head Movement, 800-1000ms **D,H**) AR-Walking, 300-500ms.

The LDA classifier performance for each condition is shown in Figure 18. The performance of the classifier regularly exceeded chance performance, but only the *Desktop* and *AR-Fixed* regularly exceeded the 95th percentile of the classifier performance on permuted data. This is analogous to statistical significance with $\alpha = 0.05$. The performance on data with movement artifacts was not stable and did not regularly exceed the 95th percentile of classifier performance on permuted data.

Figure 19 shows the LDA classifier peak performance for individual subjects as well as the average proportion correct for each condition. These results show that the classifier has variable performance across subjects. Table 2 shows the times at which these peak performances occurred during the N400 window.

Figure 18: LDA classifier performance, LOOCV. $n = 6$. *Real* indicates the classifier performance on ERP data. *Permuted: 95th Percentile* indicates the 95th percentile classifier performance on the ERP data with shuffled labels. *Permuted: Average* indicates the average classifier performance on ERP data with shuffled labels.

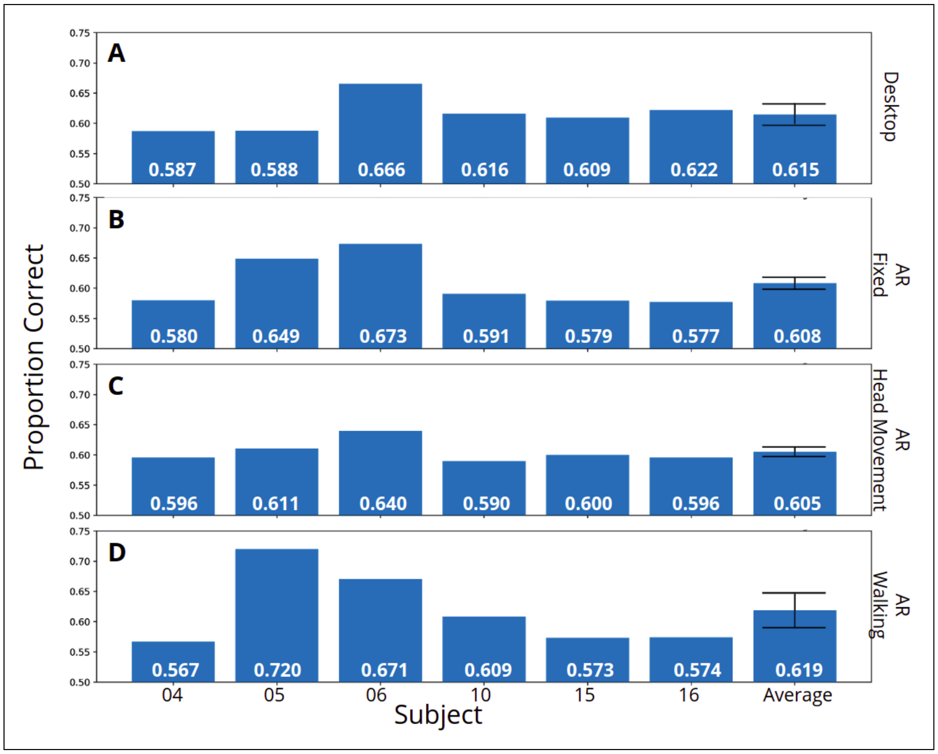

Figure 19: Peak LDA classifier performance for individual subjects during N400 region (see *Statistical Analysis* for details), LOOCV. $N = 6$. Error bars indicate SEM.

LDA Peak Accuracy Time (ms)				
Subject	Desktop	AR-Fixed	AR-Head Movement	AR-Walking
4	304	448	856	376
5	472	352	808	304
6	424	376	952	424
10	376	448	1000	304
15	304	400	1000	304
16	424	328	808	328
Average	384	392	908	340

Table 2: LDA classifier peak accuracy time during N400 region (see *Statistical Analysis* for details).

Summary

The goal of this experiment was to systematically introduce like-like movements into the task and observe the effect on the N400 ERP component. The *Desk-*

top and *AR-Fixed* had comparable ERPs and scalp topographies, but only the *AR-Fixed* had a statistically significant difference between the N400 component elicited by congruent and incongruent trials. The *AR-Head Movement* and *AR-Walking* conditions had much larger variability and distorted scalp topographies due to movement artifacts. Consistent with the ERP results, the LDA classifier exceeded the 95th percentile of performance on permuted data in the *Desktop* and *AR-Fixed* conditions, but not in the others.

Discussion

The proliferation of VR/AR devices for research applications has been observed in cognitive and behavioral studies (Foerster et al., 2016; Meehan et al., 2002; Peterson et al., 2018; Kober and Neuper, 2012). The use of these devices offers a way to create life-like settings to conduct studies, but often introduce movement artifacts as subjects interact with these environments. To assess the feasibility of using VR/AR tools in cognitive and behavioral research we conducted a series of experiments and systematically introduced movement noise. The key findings were: (1) using a traditional desktop display produced similar results as a VR display, (2) minor head and eye movements contaminated the ERP waveform in both VR and AR environments, and (3) walking introduces heavy contamination to the ERP waveform.

Although these results are promising and are consistent with the growing literature of VR/AR applications in research, there are notable limitations with these results. Because of the limited sample sizes in experiments 1 and 2, t-tests were conducted across blocks for all subjects, which limited the generalizability of these results. Additionally, movement artifacts from minor head and eye movements contaminate the scalp topographies in experiments 1 and 2, making they difficult to interpret. Movement artifacts from walking added heavy contamination and distortion to the ERP waveform and scalp topographies in experiment 3. One of the challenges with requiring head movement to move the object into the field of view is that a delay occurs while the subject is moving their head. From our pilot tests of experiment 3, we estimated that participants required 500 ms to move their head into position. Because of this, our statistical analyses for experiment 3 - head movement condition used an adjust N400 time window of 800 to 1000 ms. This adjustment was based on an estimated average time for previous participants, so this approach could potentially smear the signal when averaged across participants. Another potential source of noise was the network latency. Although we accounted for this delay using an empirical approach, our solution assumed a relatively constant network speed throughout a block of trials, which could not be guaranteed in our design.

Our results suggest that using VR/AR tools for cognitive and behavioral research is feasible and could help researchers create novel and life-like testing environments. One of the benefits of these devices is that they often come with additional useful data that can be leveraged by the researcher to improve the data quality. Head position and accelerometer data are often available and could be used in our design to estimate head and body movements for artifact filtering (Arad et al., 2018). Various types of movement artifacts and potential solutions are discussed in Islam et al. (2016); Stone et al. (2018) and Rahman et al. (2019). Eye tracking methods have also been explored (Wilson and Soranzo, 2015) and could be incorporated in our design.

References

Arad, E., Bartsch, R., Kantelhardt, J., and Plotnik, M. (2018). Performance-based approach for movement artifact removal from electroencephalographic data recorded during locomotion. *PLoS ONE*, 13(e0197153).

Bullock, T., Cecotti, H., and Giesbrecht, B. (2015). Multiple stages of information processing are modulated during acute bouts of exercise. *Neuroscience*, 307(29):138–150.

Delorme, A. and Makeig, S. (2004). Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21.

Foerster, R., Poth, C., Behler, C., Botsch, M., and Schneider, W. (2016). Using the virtual reality device oculus rift for neuropsychological assessment of visual processing. *Scientific Reports*, 6(37016):1–9.

Islam, M., Rastegarnia, A., and Yang, Z. (2016). Methods for artifact detection and removal from scalp eeg: A review. *Clinical Neurophysiology*, 46:287–305.

Kober, S. and Neuper, C. (2012). Using auditory event-related eeg potential to assess presence in virtual reality. *International Journal of Human-Computer Studies*, 70:577–587.

Makransky, G., Terkildsen, T., and Mayer, R. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, 60:225–236.

Meehan, M., Insko, B., Whitton, M., and Brooks, F. (2002). Physiological measures of presence in stressful virtual environments. *Association for Computing Machinery Transactions on Graphic*, 21(3):645–652.

Mihajlovic, V., Patki, S., and Grundlehner, B. (2014). The impact of head movements on eeg and contact impedance: An adaptive filtering solution for motion artifact reduction. *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC*, 2014:5064–5067.

Peterson, S., Furuichi, E., and Ferris, D. (2018). Effects of virtual reality high heights exposure during beam-walking on physiological stress and cognitive loading. *PLoS ONE*, 13:1–17.

Rahman, M., Karwowski, W., Fafrowicz, M., and Hancock, P. (2019). Neuroergonomic applications of electroencephalography in physical activities: A systematic review. *Frontiers in Human Neuroscience*, 13:1–21.

Si-Mohammed, H., Petit, J., Jeunet, C., Agrelaguet, F., Spindler, F., Evain, A., Roussel, N., Casiez, G., and Lecuyer, A. (2020). Towards bci-based interfaces for augmented reality: Feasibility, design and evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 26(3):1608–1621.

Stone, D., Tamburro, G., Fiedler, P., Haueisen, J., and Comani, S. (2018). Automatic removal of physiological artifacts in eeg: The optimized fingerprint method for sports science applications. *Fontiers in Human Neuroscience*, 12.

Swaab, T., Baynes, K., and Knight, R. (2002). Separable effects of priming imageability on word processing: An erp study. *Cognitive Brain Research*, 15(1):99–103.

Tarr, M. and Warren, W. (2002). Virtual reality in behavioral neuroscience and beyond. *Nature Neuroscience*, 5:1089–1092.

Troje, N. and Bülthoff, H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research*, 36(12):1761–1771.

Wilson, C. and Soranzo, A. (2015). The use of virtual reality in psychology: A case study in visual perception. *Computational and mathematical models in medicine*, 2015(151702).