Measuring Human Trust in a Virtual Assistant using Physiological Sensing in Virtual Reality

Kunal Gupta*
Empathic Computing Lab
The University of Auckland
Auckland, New Zealand

Ryo Hajika†
Empathic Computing Lab
The University of Auckland
Auckland, New Zealand

Yun Suen Pai‡
Empathic Computing Lab
The University of Auckland
Auckland, New Zealand

Andreas Duenser§
CSIRO
Hobart, Australia

Martin Lochner¶
CSIRO
Hobart, Australia

Mark Billinghurst‖
Empathic Computing Lab
The University of Auckland
Auckland, New Zealand

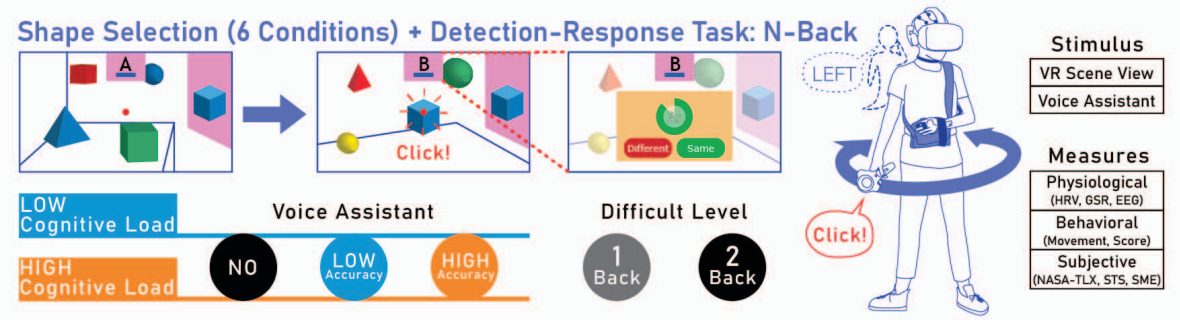

Figure 1: The system design of our VR task for measuring trust towards a virtual assistant under different cognitive load levels. We propose a shape selector task with integrated N-Back where N = 1 or 2 depending on the selected cognitive load level. The voice assistant can either be not present, present with 50% accuracy, or 100% accuracy to assist the user. From there, we measure the physiological, behavioral and subjective feedback.

Abstract

With the advancement of Artificial Intelligence technology to make smart devices, understanding how humans develop trust in virtual agents is emerging as a critical research field. Through our research, we report on a novel methodology to investigate user's trust in auditory assistance in a Virtual Reality (VR) based search task, under both high and low cognitive load and under varying levels of agent accuracy. We collected physiological sensor data such as electroencephalography (EEG), galvanic skin response (GSR), and heart-rate variability (HRV), subjective data through questionnaire such as System Trust Scale (STS), Subjective Mental Effort Questionnaire (SMEQ) and NASA-TLX. We also collected a behavioral measure of trust (congruency of users' head motion in response to valid/ invalid verbal advice from the agent). Our results indicate that our custom VR environment enables researchers to measure and understand human trust in virtual agents using the matrices, and both cognitive load and agent accuracy play an important role in trust formation. We discuss the implications of the research and directions for future work.

*e-mail: kgup421@aucklanduni.ac.nz
†e-mail: ryo.hajika@auckland.ac.nz
‡e-mail: yspai1412@gmail.com
§e-mail: andreas.duenser@data61.csiro.au
¶e-mail: mlochner@uwaterloo.ca
‖e-mail: mark.billinghurst@auckland.ac.nz

Index Terms: H.5.2 [User Interfaces]: Evaluation/methodology; H.1.2 [User/Machine Systems]: Human factors; H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities

1 Introduction

VR provides a platform where virtual environments can be visualized and interacted with at a very high level of immersion and realism, making it a popular choice for professional use cases like simulations and therapy, and entertainment like gaming and immersive storytelling. These environments can include a virtual agent which acts as a guide, advisor or collaborator and can play a critical role in determining the effectiveness of the VR experience. Virtual agents may exist in many different forms from a photo-realistic graphical avatar to a disembodied voice, yet one of the key factors that influences user experience when interacting with these agents is users' trust in them. Trust in the agent is determined by several factors, mainly the accuracy and reliability of the agent's performance. But other factors such as their intonation, accent, motion, and appearance, also contribute towards the interaction with and experience of the system [8, 30, 38]. Outside of VR, virtual agents also already exist in most modern smartphones (e.g. Google Assistant, Amazon Alexa, etc.) and even in living rooms with devices like The Apple homepod, Google Home, and so on. So the research reported in this paper may have broader applicability.

There have been various definitions for trust [17] depending on context. Generally, it describes the "willingness to be vulnerable" [29] or "willingness towards behavioural dependence" [25]. We relate closer to the latter definition, as our research aims at understanding or parameterizing how willing a user is to depend on

2642-5254/20/$31.00 ©2020 IEEE
DOI 10.1109/VR46266.2020.00099

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on July 27,2020 at 04:59:06 UTC from IEEE Xplore. Restrictions apply.

the instructions or advice from a virtual agent. This is more evident when the presented task becomes more challenging, where we may find ourselves trusting these agents more as well, akin to trusting collaborators when the goal of a group work becomes difficult to achieve. However, unlike real collaborators, a virtual agent's performance can be tweaked to maximize output. Before we establish these factors however, an effective way to measure or parameterise trust under different VR scenarios is needed.

In this work, we present a VR task specifically designed for inducing different cognitive load levels using a combination of psychologically established methods as well as methodologies designed for VR. The task incorporates a VR shape-selection task that has been proven to induce cognitive load [9] with the popular N-Back task from psychology to induce memorization, multi-tasking and temporal load [23]. The differences between high and low cognitive load conditions are the time limit, number of virtual objects in the environment, and the difficulty of the N-Back task. An audio-based virtual agent provides directional advice to the user in searching for the correct shape. Then, we measure the user's physiological signals (EEG, GSR and HRV) while they perform the task varying in difficulty and agent accuracy (50% and 100% accuracy).

Our research makes the following significant contributions:

1. We present a custom VR environment meant to maximize or minimize cognitive load (CL) in the presence of a reliable or unreliable virtual agent (VA).

2. We explore the relationship between agent trust and induced cognitive load based on gathered physiological signals, behavioral aspects and subjective feedback.

3. We found the EEG alpha channel, frequency domain-mean power, peak frequency for GSR and total power spectral density of GSR at .05Hz as well as .12Hz, the rounds completed per second, and the subjective questionnaires to be indicative of the user's trust and cognitive state.

2 Related Work

In this section, we look into the related work on measuring both cognitive load and trust for both VR and non-VR interfaces. We summarize our findings in Table 1.

2.1 Physiological Sensing

Physiological signals, or biosignals are signals from the human body that correlate to certain activities that we perform or that we react to. For example, EEG, GSR, HRV, heart rate (HR), electromyography (EMG), skin temperature (SKT), respiration (RESP), etc. are some of the example of physiological signals that can be measured, allowing researchers to use these signals for two main use cases; as an input and interaction modality [27, 35], or as implicit feedback. The latter is the main focus of this work, seeing as previous work has shown that physiological signals can reflect the humans' physiological [1], cognitive [4, 40] or emotional state [7, 46].

Measuring cognitive load using physiological signals has been explored since a long time now. In 2010, Pavlo et. al. [3] described cased studies employing EEG to collect and analyse cognitive load data while learning from hypertext and multimedia. Whereas GSR, HR, EEG, temperature and pupil dilation was used to assess mental workload during web browsing on desktop PC [21]. Previous research to analyse cognitive load in virtual environment has provided many physiological sensing parameter such as EEG, Heart Rate, and GSR to understand the cognitive load and stress effects of heights exposure in VR environment when the participant is beam-walking [36]

2.2 Trust

The definition of trust varies across the literature and is often divided into several sub-categories of trust. For example, trust has previously been categorized into persistence, technical competence and fiduciary responsibility [5], and others claiming that can be divided into dispositional, situational and learned [18]. In general though, it can be defined as a firm belief about another's intention and one's willingness to act by following their words, expressions, decisions, or actions [44]. For trust towards technology, it is a multidimensional concept based on the interaction between system attributes and users' attributes [41]. Researchers have been trying to measure trust using various means. On the physiological measure side, Hu et al. [19], Akash et al. [2] and Dong et al. [10] measured trust using EEG with GSR, and only EEG respectively. They found that both EEG and GSR can be used to model human trust with a relatively high accuracy of 71%, and that human-like cues are important in influencing EEG signals in a trust game.

On the subjective measure side, Hale et al. [14] measured trust using virtual avatar mimicry on a desktop. They found that mimicry, or imitating one's movements, does not always affect trust. This idea originated from the concept of social glue [26], where business, teaching and even therapy use some form of mimicry to induce rapport and trust. Salanitri et al. [41] instead developed a VR task alongside the Technology Trust Measure questionnaire, and found that presence in VR is a strong influence towards trust in VR technologies in general. This also shows a strong correlation between presence and trust, at least based on subjective feedback. However, this study only relied on subjective feedback from the participants via questionnaires, and the measured trust is towards the VR technology as a whole, instead of a virtual agent.

2.3 Relationship between Cognitive Load and Trust

The measure of cognitive load leans towards measuring the amount of working memory in a particular situation or task. For example, Dey et al. [9] used EEG to measure the cognitive load level of a participant, then developed an adaptive learning tool in VR that changed in difficulty depending on the measured cognitive load. This is similar to the work done by Gerry et al. [12] who also used EEG signals with a similar VR task to detect cognitive load.

McDuff et al. [31] on the other hand used HR and HRV signals that are able to detect cognitive stress from 3 meters away. The stress was induced using a desktop ball control task and a card sorting task. Another popular task would be driving, used by Zhang et al. [49] who measured cognitive load based on a VR driving game. The measured signals were EEG, eye gaze, ECG, EMG, SKT, RESP and GSR. However, it was the eye gaze with EEG data that provided the highest accuracy in detecting cognitive load through hybrid sensor fusion.

In cases where both cognitive load and trust is being measured at the same time, such as work by Samson et al. [42] and Khawaji et al. [22], it is highly likely that the physiological signals are effected by both of these factors. The work by the former author was more focused on a trust game with subjective measures, but the latter author used GSR signals in a Desktop text chat environment. It was found that GSR signals correlate with trust more than cognitive load when the task is easy, whereas it correlates more towards cognitive load when the task is difficult. Although this work is the most related to ours, we propose the measure of trust towards a virtual agent in a VR environment instead.

In summary, previous works have touched on the link between cognitive load and trust, which can be identified using GSR. Looking at trust alone, it has been proven to be measurable with both EEG and GSR, whereas cognitive load alone has proven to be measurable using a myriad of physiological signals. We summarized these findings in Table 1. Our approach differs from this work in three key categories: 1) we use a combination of EEG, GSR and HRV to

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on July 27,2020 at 04:59:06 UTC from IEEE Xplore. Restrictions apply.

Table 1: List of Previous Work Regarding Sensing of Cognitive Load and Trust

Author	CL and/ or Trust	Physiological Measure	Experiment Task	Subjective Measure
Samson et al.	CL, Trust	-	Trust game	Raven P Matrix Test (CL), EIS Trust Scale (Trust)
Hu et al.	Trust	EEG,GSR	Desktop driving game	-
Dey et al.	CL	EEG	VR Shape selector, N-Back (separate task)	-
McDuff et al.	CL	HR, HRV	Desktop ball control, Berg card sorting	Dundee Stress
Akash et al.	Trust	EEG, GSR	Desktop driving game	-
Dong et al.	Trust	EEG	Desktop matrix game	-
Zhang et al.	CL	EEG, Eye gaze, ECG, EMG, SKT, RESP	VR driving	-
Khawaji et al.	CL, Trust	GSR	Desktop text chat	-
Hale et al.	Trust	-	Virtual avatar mimicry	-
Salanitri et al.	Trust	-	VR task	Technology Trust Measure
Gerry et al.	CL	EEG	VR shape selector, N-Back (separate task)	-
Gupta et al.	CL, Trust	EEG, GSR, HRV	VR shape selector with voice assistant, N-Back (separate task)	Nasa TLX, System Trust Scale
Our Method	CL, Trust	EEG, GSR, HRV	VR-shape selector with voice assistant and N-Back combined	Nasa TLX, subjective mental effort, System Trust Scale

evaluate trust under various cognitive load levels, 2) we design and implemented a custom VR environment specifically to measure these signals for a virtual agent in VR, and 3) we relate our measurements with behavioural and subjective measures, by measuring both the user's movement and subjective feedback to identify the gap between these methods, as well as establish correlations.

3 System Design

Here we describe both the hardware and software components of the system we have developed. The system consists of six components, where the hardware setup is illustrated in Figure 2:

Hardware

- OpenBCI EEG cap with Cyton daisy module for 16 channel support (wet electrodes) at 125 Hz sampling rate. It is also possible to use the dry electrodes and open source 3D printable headset, though the results may slightly differ.
- Shimmer GSR+ Sensing device for sensing GSR and HRV signals at 128Hz.
- HTC Vive VR HMD to display the VR environment and to enable interactions. It is also compatible with Windows Mixed Reality HMDs.

Software

- OpenBCI GUI for EEG data streaming (notch filter at 50Hz, bandpass filter from 1 to 50 Hz)
- Unity 3D game engine for data acquisition and rendering the VR environment. The full sample project will be available open source.
- Java application for Shimmer data streaming and acquisition. The full code will be available open source.

3.1 Virtual Reality System

We chose the HTC Vive Head Mounted Display (HMD)[1] as VR hardware, although the Vive Pro and/or Vive Pro Eye should also be compatible. The computer powering the entire system is running on an Intel Core i7 8700 central processing unit (CPU) and an Nvidia RTX 2070 graphical processing unit (GPU). The system also runs on laptops powered by an Nvidia GTX 1060 GPU, albeit at lower frame rates.

[1] https://www.vive.com/us/

3.1.1 VR Environment

The main VR task was built with the Unity game engine, where each of the independent variables are easily modifiable for each condition. In Gupta et al.'s research [13], a major drawback with their system was that the difficulty level was not high enough. Furthermore, the N-Back task was carried out before shape selector, acting as two separate tasks where the N-Back was first for establishing the baseline, and the shape selector serves as the main task to assess cognitive load and trust. In order to make the tasks more challenging, we tried few iterations by using various stress-inducing techniques such as Stroop test with shape seletor, Stroop test with only text, and stroop test with target shape from the text and color from the shape. After each prototype, we conducted a quick pilot study with some participants and concluded the final interface that has a combination of N-Back test with shape selector task (Figure 3). With this, we are able to increase the induced cognitive load because participants are required to complete both the N-Back and shape selector task simultaneously. This combined task also allows us to induce all kinds of cognitive load, mainly memorization, multi-tasking and temporal load not seen in other VR-based task [23].

3.1.2 Study Task

Our implementation is divided into a primary and secondary task. The primary task is based on the Shape Selector Task shown by Dey et al. [9, 13] modified for the purpose of this study. For this task, there will be a target object, either in the shape of a pyramid, cube or sphere at a specific color (red, yellow, blue, green), visible in front of a pink background that the participant can always see. The participant is required to search for the exact same game object which can be located in any direction. Once the object is found, participants need to place the reticle that follows their head gaze onto the target object and pull the trigger on the Vive controller. Participants were told to complete this task as fast and as accurately as possible. To assist the participant, a voice assistant is present (depending on the

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on July 27,2020 at 04:59:06 UTC from IEEE Xplore. Restrictions apply.

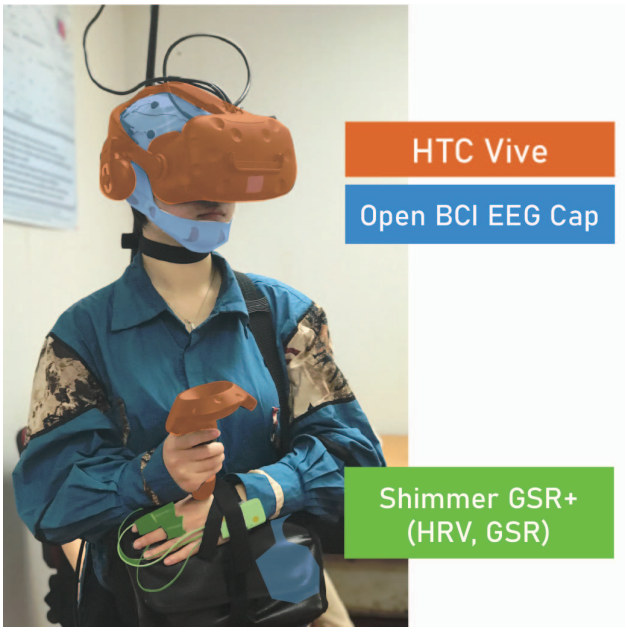

Figure 2: Harware and sensor setup. To minimize motion artifact, the HMD cables are routed upwards, the OpenBCI cables are attached to the participant's neck via velcro, and the hand with the Shimmer sensor is attached on a shoulder bag where the OpenBCI board is kept. The dominant hand holds a VR controller, where he/she simply pulls the trigger when they locate the target object and/or answer the N-Back question. Aiming with the controller is not necessary.

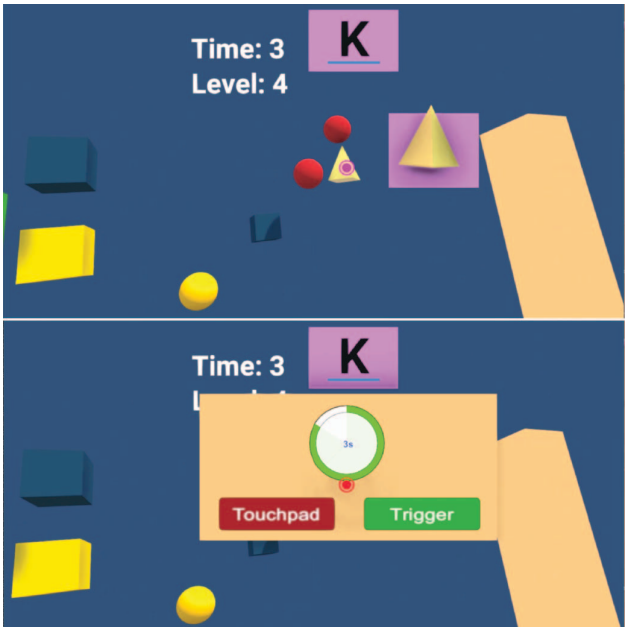

Figure 3: Screenshot of the VR task showing the (top) shape selector task being performed while the letters of the N-Back task are shown simultaneously. After the shape is selected, (bottom) the participant needs to select true or false for the displayed letter, depending on N = 1 or 2.

experimental condition) to guide participants towards the game object's direction. The participant hears either "left" or "right" as an indicator of where the game object can potentially be. The period for searching for the game object is either 5 (difficult) or 10 (easy) seconds, depending on the condition.

While the main task is being performed, we introduce a secondary task based on N-Back [23]. The N-Back task is a standard working memory task meant to further induce the cognitive load. While the participant is performing the main task, letters are constantly appearing every 1 second in the visual field (after each 1 second interval either the same or a different letter appears; letter are only visible for a period of 0.3 seconds). After participants have found the target object, they are asked whether the currently shown letter corresponds to a previous letter. In the n = 1 case, the participant has to evaluate whether the letter that was shown previous to the currently displayed one was the same letter or not. If the current letter matches the previous letter, participants pull the trigger again. If it does not match, they press the trackpad. Participants are given four seconds to answer this, visualized using a circular progress bar. In the n=2 case, the participant has to memorize the two characters that appeared before the current one and determine if it matches. The level of the N-Back is determined by the experimental condition.

If the participant has successfully completed both the primary and secondary task, she/he will advance to the next level and the task is repeated with a new target object for the main task and a new sequence of letters for the secondary task. A total of 20 trials are presented, with a time limit of 7 minutes per session. The screenshot of the task can be seen in Figure 3.

3.2 Physiological Signal Collection

To collect the physiological signals, we used two devices; the openBCI Cyton board[2] for EEG, and the Shimmer GSR+[3] for GSR and HRV. For the Cyton board, we also used a Daisy Chain module to increase the spatial resolution to 16 electrodes with a sampling rate of 125 Hz. Gel-based, or wet electrodes were used together with an EEG cap for better signal-to-noise ratio. We focused on the data from electrodes placed near the pre-Frontal lobe responsible for decision making, and ability to concentrate i.e. FP1, and FP2, and the electrodes at the parietal and occipital lobe [9] for measuring cognitive load, i.e. P3, Pz, P4, O1 and O2.

For the Shimmer Sensor, it is placed on the participant's non-dominant hand (strapped to the wrist) with the sensor being in contact with the index and middle finger. Both these sensors are worn by the participant for the duration of the experiment. A developed JAVA stand-alone application was developed to obtain the signals from the Shimmer sensor and broadcasts them to Unity via Lab Streaming Layer (LSL).

4 Experimental Evaluation

In this study, we evaluate the effectiveness of our virtual environment in terms of inducing cognitive load and trust, as well as measuring these aspects using physiological, behavioural and subjective means.

4.1 Participants and Design

We conducted a 3x2 within-subjects study based on two independent variables; induced cognitive workload, and varied the accuracy of the voice assistant as shown in Table 2. A total of 24 participants aged between 23 to 35 (12 Male, mean: 28, SD: 3.1) were recruited to complete the six conditions. Each of the condition names are abbreviated to avoid confusion (LCL for low cognitive load, HCL

[2] https://openbci.com/
[3] https://www.shimmersensing.com/

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on July 27,2020 at 04:59:06 UTC from IEEE Xplore. Restrictions apply.

Table 2: Experimental Conditions

	No VA	Low Accuracy VA	High Accuracy VA
Low CL	LCL-NOVA	LCL-LAVA	LCL-HAVA
High CL	HCL-NOVA	HCL-LAVA	HCL-HAVA

for high cognitive load, NOVA for no voice assistant, LAVA for low accuracy voice assistant, and HAVA for high accuracy voice assistant).

The order of conditions for each participant was arranged in a Latin Square to reduce potential ordering effects. The low cognitive load condition involves the N-Back task with n=1, ten seconds to perform the shape selector task, and fewer distractor objects. For the high cognitive load condition, participants need to complete the N-Back task with n=2, are only given five seconds for the main task, and are presented with more distractor objects to increase the task difficulty. This particular task was designed based on the three main components of cognitive load; time load, mental effort load and psychological stress load [11]. For the conditions regarding the voice assistant, we include three variations; no assistant, an inaccurate assistant, and an accurate assistant. The conditions with no assistant serves as a baseline. The inaccurate voice assistant was set to 50% accuracy to minimize predictability, and the accurate assistant is 100% accurate. All of the participants were above 18 years of age, native English speakers or fluent in English, familiar with computers and smartphones, and had some experience with virtual environments. They also all had some experience with using virtual assistants like Google Assistant, Apple Siri, Bixby, or Amazon Alexa for tasks such setting an alarm, searching for a nearby cafe, and setting up a destination for car navigation.

4.1.1 Cognitive Load Measures

We measured cognitive load with subjective and physiological measures. The subjective measure is based on the weighted NASA Task Load Index (TLX) [15]. We also included the subjective mental effort (SME) single question questionnaire which focused on mental load. For the physiological measures, we gathered and analyzed the EEG, GSR and HRV signals.

4.1.2 Trust Measures

The trust measures only apply for the LCL-LAVA, HCL-LAVA, LCL-HAVA and HCL-HAVA conditions as it is meant to evaluate the participant's trust towards the voice assistant. We measured trust in three ways: subjective, physiological and behavioural. We used the System Trust Scale (STS) questionnaire [20] as a subjective measure. The psychological measures, like cognitive load, include EEG, GSR and HRV signals. As a behavioural measure, we record (using head tracking) the direction of the head movement of the participant relative to the target (left or right) game object during the shape selector task, along with the direction informed by the agent (left or right as well) within the same timestamp throughout the experiment. An equal direction between the participant and the voice assistant is labeled as trust.

4.2 Procedure

The experiment was conducted in a room with minimal radio frequency interference as there was a risk of extra noise in the physiological signals due to such interference (however, the main computer used to run the system was present). After welcoming the participants, they were first given a copy of the Consent Form and Participant Information Sheet to fill at the start of the session with an opportunity to ask any questions about the study. Once they signed the CF, they were asked to complete the pre-task questionnaire including questions regarding demography, previous VR and virtual assistant experience.

The participants then had a training phase to familiarise themselves with the task. We started by explaining and letting them try the N-Back task implemented in VR, where the right trigger controller is pressed for matching letters and the track pad for not matching. Participants had to complete the task for n=1,2 and 3 which took about five minutes. After that, we explained the actual task (both the main and secondary tasks) and asked participants to run a trial round with low cognitive load settings and no voice assistant for another five minutes. After this we asked them to wash and dry their hands and then put the GSR and HRV sensors on their non-dominant hand. We then setup the OpenBCI EEG cap with gel in the electrodes followed by the Vive VR HMD. While filling the gel, we made sure that the impedance level for each electrode was no more than 40kohm. Periodically, we also filled the gel between trials to keep the impedance level low and prevent the gels from drying up. To minimize motion artifacts, we put the Cyton board into a shoulder bag and requested the participant to wear it. Participants could rest the non-dominant hand on the bag and we secured the hand onto the bag using velcro tape. We also used velcro tape as a choker to secure the cables from the electrodes around the participant's neck, while ensuring that the participant were comfortable with the tightness.

The entire setup had to be carefully completed as the EEG cap electrodes could be displaced from their position because of the HMD straps as well, resulting in faulty EEG data. After resting for 1 minute allowing the electrode gel to settle in, we started the main task. At the end of each session, participants were asked to fill out the weighted NASA TLX questionnaire, the SME questionnaire, and the STS questionnaire (only for conditions LCL-LAVA, HCL-LAVA, LCL-HAVA and HCL-HAVA). At the end of the experiment, we conducted non-structured, open-ended interviews with the participants to understand their perspective and experiences while performing the tasks. Participants were compensated with a $20 gift card voucher. Each session took approximately one hour and 30 minutes.

4.3 Data Acquisition and Preprocessing

The experiment's data pre-processing steps are reported in this section. First, we extracted time window data, including the last 2 minutes for each trial. This was done to balance the sample length across all trials. It also enabled us to use the first segment of each trial as a calibration stage for trust and cognitive load in that condition. By the final two minutes in each trial, participants were familiar with the task, allowing for an accurate representation of their physiological state in relation to that condition. [43].

4.3.1 EEG Processing

Since we collected raw EEG signals during the experiment, we first needed to preprocess the data to get the individual band frequencies. From the 16 acquired channels, we focused only on FP1, FP2, O1, O2, P3, P4 and Pz channels which are located on the pre-frontal cortex (FP1, FP2), occipital lobe (O1, O2) and parietal lobe (P3, P4 and Pz) according to the 10-20 layout [34]. We choose these channels because the pre-frontal cortex is in charge of decision making, cognitive state and problem solving [32], the occipital lobe is tasked with vision processing [28] and parietal lobe informs about the attentional demands [24].

We inspected the collected signals for each participant and condition manually and remove trials with large interruptions in the signal such as missing signals, large continuous spikes, or streams of a fixed constant value for a long duration of the trial. This accounted for 5.56% of the overall collected EEG data. Then, we applied a bandpass filter to only acquire signals that fall in the frequency range of 1Hz to 40Hz. Next, we use the Wiener Filter to remove motion artifacts as it was found to be one of the most efficient algorithm for motion artifact removal [45]. We found this procedure to be a necessity because both the constant turning of the participants together with the slight shift in the HMD's position during turning

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on July 27,2020 at 04:59:06 UTC from IEEE Xplore. Restrictions apply.

introduced many artifacts that needed to be removed. Following that, we performed independent component analysis (ICA) to check and remove electrooculography (EOG) signals generated from eye movements. Once the fitering was complete, we performed a Fast Fourier Transform (FFT) to extract each of the bands from the signals: Delta (1-4Hz), Theta (4-8Hz), Alpha (8-12Hz), Beta (12-30Hz) and Gamma (30-45Hz) waves. For further analysis, we followed Gupta et al. [13] and used only the alpha band.

4.3.2 GSR and HRV Processing

We visually inspected the GSR signals and excluded trials with large interruptions in the signal (e.g., due to poor electrode connectivity). This accounted for 16.8% of the data overall. These interruptions arise from poor electrode connectivity, displacement of the sensor placement from moving too rapidly, or the logging sometimes freezes after a period of time. Nevertheless, there are still well over 30,000 samples (before down-sampling) present for analysis. Individual pairwise proportions tests indicated that the removal rate did not differ significantly between conditions ($p>0.05$ in all cases). We then extracted a time window, including the last 2 minutes for each trial. This was done to balance the sample length across all trials as few participants took 2 minutes to complete tasks especially the low cognitive load tasks. It also enabled us to use the first segment of each trial as a calibration stage for trust and cognitive load in that condition. By the final two minutes in each trial, participants were familiar with the task, allowing for an accurate representation of their physiological state in relation to that condition. [43]. Finally, we down-sampled the signals from 128 Hz to 20 Hz, to improve the computational efficiency of the following analyses.

Both time-domain (raw and standardized GSR) and frequency-domain (Fast Fourier Transform, Welch Power Spectral Density [WPSD]) analyses were performed on the GSR data, using Linear Mixed-Effects Models [6], with participant as a random factor in all analyses. The FFT was performed using the base R stats package [47], while the WPSD [48] was performed using the R 'bspec' library [39]. The means for raw and standardized GSR did not differ statistically across either the CL or Trust factors, $p>.05$. However, Nourbaksh et al. [33] has investigated frequency domain of GSR in relation to Cognitive Load, with some success through arithmetic and reading tasks. To this end, we first computed the gradient (slope) of the GSR signal for each block, and submitted the resulting gradient to a Fast Fourier Transform using the R base package. The FFT and WPSD plots of GSR in the CL and Trust conditions are presented in Figure 4.

Next, we looked into the collected HRV signals which were gathered alongside the GSR signals using the Shimmer sensor. We also manually inspected and removed signals we deemed unfit for analysis, such as those with continuous logging of 0 values mainly due to poor electrode connectivity. Moreover, we also manually cleaned duplicated data which emerged from how Unity's framerate upsampled the signals superficially. In total, we removed 7.64% of the data. For the remaining, we applied a lowpass Butterworth filter of 10Hz first, followed by the Hampel correction to identify remaining outliers and replaced them with more representative values using the Heartpy Python package [4]. Finally, we extracted the beats per minute (BPM), root mean square of the successive differences(RMSSD), interbeat intervals (IBI), and low frequency-to-high frequency ratio (LF/HF), , which can be indicative of a person's cognitive state [22]. We were mainly interested in the LF/HF value since it has proven to directly reflect a person's cognitive state [31].

4.3.3 Head Movement Data Processing

For the head movement data, we observed that during the first half of a condition, the participant's head motions may reflect a reaction to a sudden audio stimulus rather than them trusting the advice or not. Therefore, we focused on the latter half set of data for all conditions. For example, if a participant was able to complete the condition with 54 rounds (including the incorrect rounds) in the given time, we only looked at the last 27 trials data. Next, we only extract the one second time window of data immediately after they heard the assistant's voice for the analysis. We deemed that the participant trusts the voice assistant if she follows the direction suggested by it.

[4] https://python-heart-rate-analysis-toolkit.readthedocs.io/en/latest/

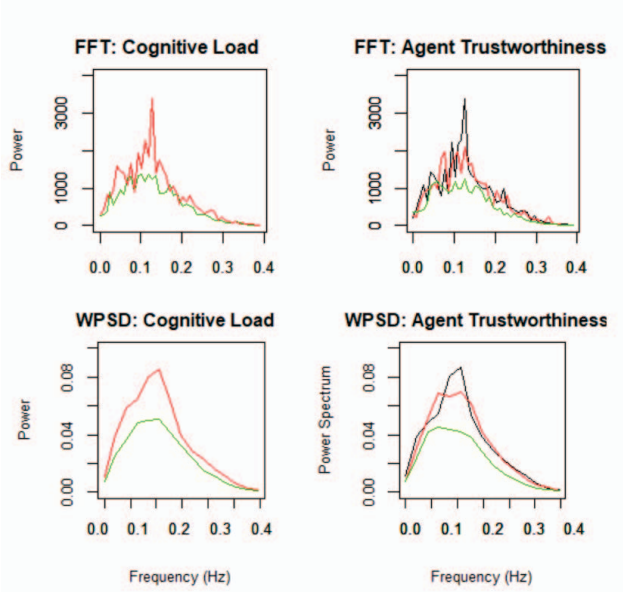

Figure 4: Fast Fourier Transform (top) and Welch Power Spectral Density (bottom) of the gradient of the GSR signal, plotted by Cognitive Load (left) and Accuracy (right). For the CL plot, Low load is green, and High load is red. For the Accuracy plot, No Advice is black, High Accuracy is green and Low Accuracy is red.

4.3.4 System Trust Scale Processing (STS)

We calculated the final trust score from the STS responses by first reversing the rating for the negative valence questions i.e. Q1-5. For example, if someone rating 4 for Q2, we subtracted it from 6 (Scale Length + 1) in order to reverse the response, so the updated rating was 2. Then we calculated a Trust score by averaging all the ratings from the System Trust Scale questionnaire.

5 Results

In this section, we analyse the obtained results from the user study and categorize them into analysis for each physiological signal, behavioural, and subjective measures.

5.1 Physiological Measures

The statistical analysis and results from the experiment are reported in this section. We have divided it into three major subsections: Subjective Measures (NASA TLX, STS, and SME), Behavioral Measures (rounds per second, error rate, and head movement), and physiological measures (EEG, GSR, and HRV).

5.1.1 EEG Signal

We performed a channel-wise band power analysis on the EEG pre-processed data. The normality test on the mean alpha band power described it as non-parametric. For factorial analysis of the mean alpha band power, we used the Aligned Rank Transform for nonparametric factorial analysis using ANOVA procedures (α =

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on July 27,2020 at 04:59:06 UTC from IEEE Xplore. Restrictions apply.

0.05) with the Bonferroni test for post hoc comparisons. There was a significant main effect of cognitive load ($F(1,144) = 4.425$, $p=0.037$) on the alpha band power. No significant main effect of accuracy ($F(1,144) = 0.262$, $p=0.770$) and no interaction effect of cognitive load and accuracy ($F(1,144) = 0.030$, $p=0.971$) was found on the alpha band power.

Descriptive statistics showed that on average the alpha band power decreased to about 71% in HCL-LAVA ($M=23.250$, $SD = 36.170$) as compared to LCL-LAVA ($M=81.21$, $SD = 30.642$), and about 27% increase in alpha band power in HCL-HAVA ($M=99.96.35$, $SD = 29.338$) compared to LCL-HAVA ($M=72.75$. $SD=30.395$).

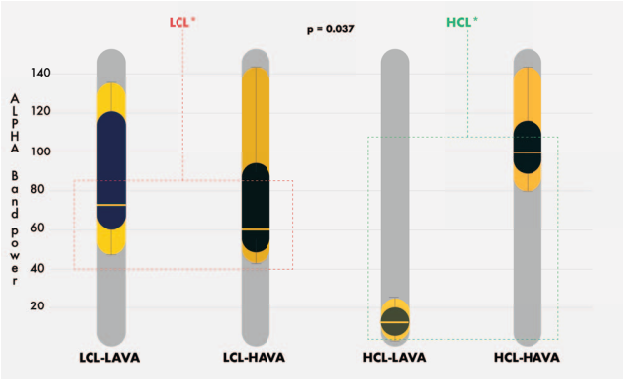

Figure 5: Box Plot of the alpha band power from FP1, FP2, O1,O2, P3, P4, and Pz Channels across voice assistant conditions.

5.1.2 GSR and HRV Signals

We tested for significant differences in the FFT signal using three main metrics: Mean Frequency, Peak Frequency, and Total Power, as described in [37]. Mean Frequency assesses the 'centre of mass' of a waveform along the X-axis; Peak Frequency assess the frequency at which the maximum amplitude was obtained; and Total Power assess the amplitude of the signal at a given frequency. For the FFT analysis, both Mean ($F(2,92)=4.13$, $p=0.019$) and Peak Frequency ($F(2,94)=4.59$, $p=0.013$) metrics differed significantly across levels of Trust. The effect of Cognitive Load was non-significant for both metrics ($p>0.05$), and there were no significant interactions. For Total Power of the FFT, the difference between Low and High cognitive load was significant at two points in the spectrum, at approximately .05 Hz ($F(1,85)=3.99$, $p=0.049$), and again at approximately .12 Hz ($F(1,88)=5.0$, $p=0.028$).

We tested the normality of the LF/HF ratio feature data across all the conditions using the Shapiro-Wilk test that reported it as non-parametric. For non-parametric factorial analysis, we used ART with the Bonferroni post hoc test. The test reported that there was no significant main effect of CL ($F(1,24) = 0.562$, $p=0.455$), and Accuracy ($F(1,24) = 0.597$, $p=0.552$). No significant interaction effect of CL and Accuracy ($F(1,24)=0.632$, $p=0.533$) was reported as well.

Descriptive statistics showed that on an average the LF/ HF ratio increased from LCL-NOVA ($M=65.42$, $SD=43.636$) to HCL-NOVA ($M=81.458$, $SD=35.202$) by about 24%. It also increased LCL-LAVA ($M=68.42$, $SD=39.289$) to HCL-LAVA ($M=82.00$, $SD=39.232$) by about 20% whereas it decreased by around 2% in HCL-HAVA ($M=68.00$, $SD=47.426$) as compared to LCL-HAVA ($M=69.71$, $SD=45.454$). This is illustrated in Figure 7.

5.2 Behavioural Measures

In this section, we discuss the analysis of head movement as a behavioural measure and overall performance score.

5.2.1 Head Movement

We performed a two-way repeated measure ANOVA on this trust factor for all the conditions and determined that there was no significant main effect of CL ($F(1,24)=0.328$, $p = 0.572$) or Accuracy ($F(1,24)=0.100$, $p = 0.754$) and no significant interaction effect between CL and Accuracy ($F(1,24)=0.644$, $p = 0.431$).

From the descriptive statistics, we determined that on an average there was a decline in about 4% of trust factor in the HCL-LAVA ($M=0.5206$, $SD = 0.06761$) as compared to LCL-HAVA ($M=0.5433$, $SD = 0.088$). Whereas, on an average there was a around 0.7% increment of trust factor when participant was performing HCL-HAVA ($M=0.5380$, $SD = 0.062$) than LCL-HAVA ($M=0.5341$, $SD = 0.10$).

5.2.2 Performance

We computed rounds completed per second (RPS) by dividing total number of rounds each participant played in each condition by the total time taken (figure 6 C). A two-way repeated measure ANOVA reported that there was a significant main effect of Accuracy ($F(1,24)=11.367$, $p = 0.003$) but no significant main effect of CL ($F(1,24)=0.257$, $p = 0.617$) or interaction effect between CL and Accuracy ($F(1,24)=0.175$, $p = 0.679$) was found.

5.3 Subjective Measures

For this section, we analyse the STS, SME and TLX questionnaires. The results from these questionnaires can be seen on Figure 6 A,B.

5.3.1 System Trust Scale

A two-way repeated measure ANOVA was performed on the Trust score for all the conditions to learn about the participant's trust perception on the Voice Assistant. The test showed a significant main effect of both CL ($F(1,24)=12.569$, $p=0.002$) and Accuracy ($F(1,24)=108.585$, $p<0.001$). There was also a significant interaction effect between CL and Accuracy ($F(1,24) = 13.571$, $p=0.001$).

5.3.2 Subjective Mental Effort Questionnaire

A Shapiro-Wilk test reported the SMEQ responses as non-parametric. We found a significant main effect of CL ($F(1,144)=34.050$, $p<0.001$) on the overall task's perceived difficulty level using ART factorial analysis. There was no significant effect of Accuracy ($F(1,144)=0.662$, $p=0.518$) and no interaction effect of CL and Accuracy ($F(1,144)=0.666$, $p=0.515$).

5.3.3 NASA TLX

A two-way repeated measure ANOVA on the weighted average NASA TLX Score [16] showed a significant effect of both CL ($F(1,24)=57.594$, $p<0.001$) and Accuracy ($F(1,24)=4.843$, $p=0.012$). No significant interaction was found between CL and Accuracy ($F(1,24)=0.497$), $p=0.611$).

6 Discussion

As previously reported [9], alpha bandpower is inversely proportional to Cognitive Load (i.e. higher alpha bandpower, lower cognitive load and vice versa), our significant main effect of LCL and HCL conditions from EEG analysis clearly shows high cognitive load in HCL-LAVA condition and a low cognitive load in HCL-HAVA. This indicates that an inaccurate agent reduced the alpha band power hence increased the physiological cognitive load as compared to an accurate agent who increased the alpha band power i.e. decreased physiological cognitive load in a high cognitive load condition. This provides additional insight to the results reported by Gupta et al. [13] where they couldn't report any relationship of cognitive load and trust for high cognitive load conditions caused because of tasks not being difficult enough to differentiate between low and high cognitive load tasks.

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on July 27,2020 at 04:59:06 UTC from IEEE Xplore. Restrictions apply.

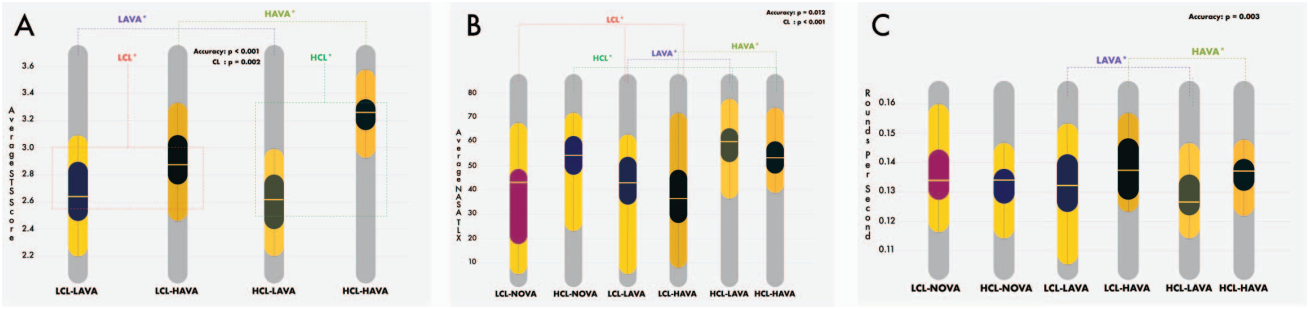

Figure 6: Results for STS (A), TLX (B) and Rounds Per Second (C) analysis across all conditions. The STS plot does not include the NOVA conditions since they evaluate trust, and NOVA has no voice assistant present

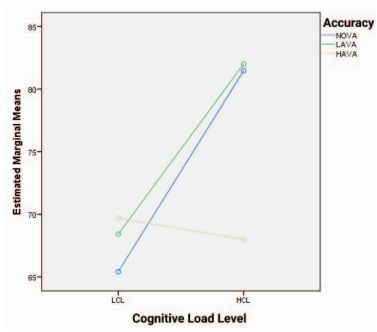

Figure 7: Results for the estimated marginal means of LF/HF for both cognitive load levels and No Assistant, Low Accuracy agent and High Accuracy agent

The significant effect of cognitive load on GSR FFT power aligns with previous research suggesting that the designed tasks induced significantly different cognitive load but didn't report any significant effect of change in accuracy level. However, other FFT features like mean and peak frequency showed an effect of agent accuracy. Overall, In contrast with other studies, our GSR analysis didn't show any significant relationship between trust and cognitive load. It was in the total power analysis in the FFT spectrum that we found two points of statistical significance, namely at 0.05Hz and 0.12Hz between low and high cognitive load. These results suggests that at these frequencies, the amplitude of the GSR signal can reflect a person's cognitive load level, thus could potentially be used as a key feature for machine learning algorithms in cognitive load detection [33]. Even though the WPSD analysis did not elicit significance, the trend of the plot shown in Figure 4 reflects the FFT plot.

Through the analysis of LF/HF HRV feature, we didn't find any significant main or interaction effect. However, the comparison of means of LF/HF in Figure 7 indicates that on an average, the mean for the condition when the agent was accurate lowered down indicating reduced cognitive load as compared to the condition where the agent was inaccurate that increased the mean even more than the condition when there was no assistant. This aligns with the views of six participants, with one of them reporting "Helper made my job easy. but I would prefer No assistance if the helper is inconsistent".

We couldn't find any significance in head movement, however, Seven participants mentioned that despite them knowing that agent may be inaccurate, they prefer to follow the directions suggested by the voice assistant. According to one of the participant "I take the direction as a Heads-up! I still have a 50% chance that I will find the object". This also explains the task performance analysis using rounds completed per second (RPS) indicating participants being able able to complete more rounds in high accuracy conditions regardless of cognitive load. However,after few trials in high cognitive load conditions, three participants started to follow opposite direction to what the agent suggested as they got frustrated by the inaccurate agent. One of the participant (P19) shouted at the agent saying "I don't trust you anymore!!".

7 Conclusion and Future Work

In this work, we introduce a novel methodology on assessing the trust level towards a virtual agent in VR under different cognitive load levels. The developed method incorporates a shape selector task with the N-Back task that allows the control on required memorization, multi-tasking and temporal load. During the process, we explored few techniques from non-VR scenario to induce cognitive load and developed those in VR to find the most optimum technique to control the cognitive load. We also found that human trust towards virtual agents can be measured using the collected physiological, behavioral, and subjective measures.

In the future, we wish to further expand the data capture compatibility to include additional sensing modalities like electromyography, eye tracking, and so on. Furthermore, we will look into real-time signal preprocessing and classification so that the system can be further expanded as a real-time trust detector in VR. In this research, we chose alpha band for EEG analysis but in future it could be interesting to investigate if there is any relationship in other bands (i.e. Delta, Theta, Beta, and Gamma).

8 Acknowledgments

Part of the work was funded by the Data61, UTAS UniSA Automation, trust and workload CRP.

References

[1] M. Abouelenien, M. Burzo, R. Mihalcea, K. Rusinek, and D. Van Alstine. Detecting human thermal discomfort via physiological signals. In *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '17, pp. 146–149. ACM, New York, NY, USA, 2017. doi: 10.1145/3056540.3064957

[2] K. Akash, W.-L. Hu, T. Reid, and N. Jain. Dynamic modeling of trust in human-machine interactions. In *2017 American Control Conference (ACC)*, pp. 1542–1548. IEEE, 2017.

[3] P. Antonenko, F. Paas, R. Grabner, and T. Van Gog. Using electroencephalography to measure cognitive load. *Educational Psychology Review*, 22(4):425–438, 2010.

[4] O. Augereau, B. Tag, and K. Kise. Mental state analysis on eyewear. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18, pp. 968–973. ACM, New York, NY, USA, 2018. doi: 10.1145/3267305.3274119

[5] B. Barber. *The logic and limits of trust*, vol. 96. Rutgers University Press New Brunswick, NJ, 1983.

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on July 27,2020 at 04:59:06 UTC from IEEE Xplore. Restrictions apply.

[6] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01

[7] H. Chen, A. Dey, M. Billinghurst, and R. W. Lindeman. Exploring the design space for multi-sensory heart rate feedback in immersive virtual reality. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, pp. 108–116. ACM, 2017.

[8] A. Davis, J. D. Murphy, D. Owens, D. Khazanchi, and I. Zigurs. Avatars, people, and virtual worlds: Foundations for research in metaverses. *Journal of the Association for Information Systems*, 10(2):90, 2009.

[9] A. Dey, A. Chatourn, and M. Billinghurst. Exploration of an eeg-based cognitively adaptive training system in virtual reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 220–226. IEEE, 2019.

[10] S.-Y. Dong, B.-K. Kim, K. Lee, and S.-Y. Lee. A preliminary study on human trust measurements by eeg for human-machine interactions. In *Proceedings of the 3rd International Conference on Human-Agent Interaction*, HAI '15, pp. 265–268. ACM, New York, NY, USA, 2015. doi: 10.1145/2814940.2814993

[11] V. J. Gawron. *Human performance measures handbook.* Lawrence Erlbaum Associates Publishers, 2000.

[12] L. Gerry, B. Ens, A. Drogemuller, B. Thomas, and M. Billinghurst. Levity: A virtual reality system that responds to cognitive load. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, pp. LBW610:1–LBW610:6. ACM, New York, NY, USA, 2018. doi: 10.1145/3170427.3188479

[13] K. Gupta, R. Hajika, Y. S. Pai, A. Duenser, M. Lochner, and M. Billinghurst. In ai we trust: Investigating the relationship between biosignals, trust and cognitive load in vr. In *25th ACM Symposium on Virtual Reality Software and Technology*, VRST '19, pp. 33:1–33:10. ACM, New York, NY, USA, 2019. doi: 10.1145/3359996.3364276

[14] J. Hale and F. D. C. Antonia. Testing the relationship between mimicry, trust and rapport in virtual reality conversations. *Scientific reports*, 6:35295, 2016.

[15] S. G. Hart. Nasa task load index (tlx). volume 1.0; paper and pencil package. 1986.

[16] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988.

[17] B. Hernandez-Ortega. The role of post-use trust in the acceptance of a technology: Drivers and consequences. *Technovation*, 31(10-11):523–538, 2011.

[18] K. A. Hoff and M. Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434, 2015.

[19] W.-L. Hu, K. Akash, N. Jain, and T. Reid. Real-time sensing of trust in human-machine interactions. *IFAC-PapersOnLine*, 49(32):48–53, 2016.

[20] J.-Y. Jian, A. M. Bisantz, and C. G. Drury. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71, 2000.

[21] A. Jimenez-Molina, C. Retamal, and H. Lira. Using psychophysiological sensors to assess mental workload during web browsing. *Sensors*, 18(2):458, 2018.

[22] A. Khawaji, J. Zhou, F. Chen, and N. Marcus. Using galvanic skin response (gsr) to measure trust and cognitive load in the text-chat environment. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 1989–1994. ACM, 2015.

[23] W. K. Kirchner. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, 55(4):352, 1958.

[24] W. Klimesch. Eeg-alpha rhythms and memory processes. *International Journal of psychophysiology*, 26(1-3):319–340, 1997.

[25] D. Knights, F. Noble, T. Vurdubakis, and H. Willmott. Chasing shadows: control, virtuality and the production of trust. *Organization studies*, 22(2):311–336, 2001.

[26] J. L. Lakin, V. E. Jefferis, C. M. Cheng, and T. L. Chartrand. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior*, 27(3):145–162, 2003.

[27] J. C. Lee and D. S. Tan. Using a low-cost electroencephalograph for task classification in hci research. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, pp. 81–90. ACM, 2006.

[28] R. Malach, J. Reppas, R. Benson, K. Kwong, H. Jiang, W. Kennedy, P. Ledden, T. Brady, B. Rosen, and R. Tootell. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, 92(18):8135–8139, 1995.

[29] R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[30] R. McDonnell and M. Breidt. Face reality: investigating the uncanny valley for virtual faces. In *ACM SIGGRAPH ASIA 2010 Sketches*, p. 41. ACM, 2010.

[31] D. J. McDuff, J. Hernandez, S. Gontarek, and R. W. Picard. Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 4000–4004. ACM, New York, NY, USA, 2016. doi: 10.1145/2858036.2858247

[32] E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

[33] N. Nourbakhsh, Y. Wang, F. Chen, and R. A. Calvo. Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, pp. 420–423, 2012.

[34] R. Oostenveld and P. Praamstra. The five percent electrode system for high-resolution eeg and erp measurements. *Clinical neurophysiology*, 112(4):713–719, 2001.

[35] Y. S. Pai, T. Dingler, and K. Kunze. Assessing hands-free interactions for vr using eye gaze and electromyography. *Virtual Reality*, 23(2):119–131, Jun 2019. doi: 10.1007/s10055-018-0371-2

[36] S. M. Peterson, E. Furuichi, and D. P. Ferris. Effects of virtual reality high heights exposure during beam-walking on physiological stress and cognitive loading. *PloS one*, 13(7):e0200306, 2018.

[37] A. Phinyomark, S. Thongpanja, H. Hu, P. Phukpattaranont, and C. Limsakul. The usefulness of mean and median frequencies in electromyography analysis. In *Computational intelligence in electromyography analysis-A perspective on current applications and future challenges*. IntechOpen, 2012.

[38] L. Qiu and I. Benbasat. Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars. *International journal of human-computer interaction*, 19(1):75–94, 2005.

[39] C. Roever and M. C. Roever. Package 'bspec'. 2015.

[40] D. Rozado and A. Dunser. Combining eeg with pupillometry to improve cognitive workload detection. *Computer*, 48(10):18–25, 2015.

[41] D. Salanitri, G. Lawson, and B. Waterfield. The relationship between presence and trust in virtual reality. In *Proceedings of the European Conference on Cognitive Ergonomics*, ECCE '16, pp. 16:1–16:4. ACM, New York, NY, USA, 2016. doi: 10.1145/2970930.2970947

[42] K. Samson and P. Kostyszyn. Effects of cognitive load on trusting behavior–an experiment using the trust game. *PloS one*, 10(5):e0127680, 2015.

[43] Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen. Galvanic skin response (gsr) as an index of cognitive load. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '07, pp. 2651–2656. ACM, New York, NY, USA, 2007. doi: 10.1145/1240866.1241057

[44] D. Susan and J. G. Holmes. The dynamics of interpersonal trust: Resolving uncertainty in the face of risk. *Cooperation and Prosocial Behavior; Cambridge University Press: New York, NY, USA*, p. 190, 1991.

[45] K. Sweeney. *Motion Artifact Processing Techniques for Physiological Signals*. PhD thesis, National University of Ireland Maynooth, 2013.

[46] W. Szwoch. Emotion recognition using physiological signals. In *Proceedings of the Mulitimedia, Interaction, Design and Innnovation*, MIDI '15, pp. 15:1–15:8. ACM, New York, NY, USA, 2015. doi: 10.1145/2814464.2814479

[47] R. C. Team et al. R: A language and environment for statistical com-

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on July 27,2020 at 04:59:06 UTC from IEEE Xplore. Restrictions apply.

puting. 2013.
[48] P. Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.
[49] L. Zhang, J. Wade, D. Bian, J. Fan, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar. Cognitive load measurement in a virtual reality-based driving system for autism intervention. *IEEE transactions on affective computing*, 8(2):176–189, 2017.

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on July 27,2020 at 04:59:06 UTC from IEEE Xplore. Restrictions apply.