Towards an Online, Objective Measure of Situation Awareness using EEG: Assessing the Relation between Attention and SA

A FIRST STEP TOWARDS AN OBJECTIVE ONLINE MEASURE OF SITUATION AWARENESS USING ELECTROENCEPHALOGRAPHY

MASTER'S THESIS
HUMAN-MACHINE COMMUNICATION
UNIVERSITY OF GRONINGEN, THE NETHERLANDS
JULY 2016

WRITTEN BY

ROBIN KRAMER
s1970755

INTERNAL SUPERVISOR

DR. MARIEKE K. VAN VUGT
University of Groningen

EXTERNAL SUPERVISOR

JULIA C. LO MSC.
ProRail Innovation and Development

university of groningen / faculty of mathematics and natural sciences / artificial intelligence and cognitive engineering

Abstract

Many years of research have shown that operator performance in safety-critical work environments is in a large degree dependent of situation awareness (SA). The currently existing methods for assessing the quality of SA, however, have some shortcomings that make them unsuitable for field studies among others. Given the importance of attention for maintaining high quality SA and the large body of research showing that attention can be captured using EEG, EEG may be a possible candidate for a new online, objective measure of SA. Therefore, in this exploratory study, we sought to capture the relation between SA and several EEG metrics of attention. In addition, we compared the data of a medical grade EEG system (128 channel BioSemi ActiveTwo) with a wireless and wearable headset (9 channel B-Alert X10), to test whether EEG can be recorded reliably in field studies.

Student participants performed a train traffic controller (TTC) task twice (once with each EEG system). During the task, SA was sampled periodically with the Situation Present Assessment Method (SPAM) and a psycho-motor vigilance (PMV) task was added as a behavioral measure of attentiveness. Several EEG metrics of attention were related to the response times (RTs) on the SPAM but no significant relation was found. The results on the PMV task did suggest that participants experienced a high level of workload in the first experiment, which is ascribed to inexperience with the task. In a pilot study with four TTCs, of which only qualitative data was analyzed and discussed, insights were gained that argue for the use of trained professionals in SA related research.

The data of BioSemi and the B-Alert was compared based on event-related potentials (ERP). The B-Alert X10 data was to a large extent in accordance with the BioSemi data, suggesting that high quality data can be obtained using the wireless wearable EEG system. Differences between the systems were found in both amplitude and latency of the P3 response, for which several possible explanations are discussed.

All in all, no relation between EEG metrics of attention and SA was found with the current experimental setup. Lessons need to be drawn from this study, in order to make future endeavors in this line of research more successful. Most importantly, in order to gain a better understanding of the dynamics between SA and the attentional demands of a task, you need to consider the level of skill of the participants and make use of trained professionals. This is not only important for gaining SA by the participants, but also for assessing SA by the researchers.

Keywords: situation awareness, attention, EEG, human factors, train traffic control, system comparison.

Contents

Introduction

In everyday life it is important to pay attention and stay aware of what is going on around oneself. For example, when checking whether it is safe to cross the road, you must look if the traffic light is green and must understand that green means you are allowed to go. This allows you to predict that it is also safe to cross the road, which you can act upon. This process of gaining an understanding of the situation is also referred to as situation awareness (SA). SA can be defined as "a generative process of knowledge creation and informed action-taking" (Smith & Hancock, 1995, p. 63). In other words, SA is a cyclical process in which perception, knowledge, anticipation and action-taking take place concurrently to create an understanding of the current situation.

SA has often been linked to the level of operator performance in safety-critical work environments. When an operator fails to perceive information relevant to the task, for instance, the operator may be less aware of the situation and, in turn, rely on bad, error-prone coping strategies during decision making (Steenhuisen, 2009). An illustration of the importance of SA can be seen in the aviation domain where 71% of airplane accidents have been attributed to human error, of which 88% were directly related to a lack of SA (Liu, Wanyan, & Zhuang, 2014). In another study it was found that of all the SA related errors in aviation, 76% were attributed to perceptual related issues (Endsley & Garland, 2000a). The importance of SA is not only limited to the aviation domain, as the the work of train traffic controllers (TTCs) is argued to be very similar to air traffic controlling. Although research paid little attention to SA of TTCs as of yet, SA is also an incredibly relevant topic in this domain (Golightly, Wilson, Lowe, & Sharples, 2010), and understanding how an operator can fail to gain or may lose SA during the job is of vital importance.

Identifying when and how this occurs allows us, among others, to make changes to the task or to the interface, such that SA can be maintained at a high level with relative ease. This design process is also referred to as situation-awareness oriented design (Endsley, 2013). The currently existing methods for assessing the quality of SA, however, all have their shortcomings. For example, query based tools, such as the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1988) or the Situation Present Assessment Method (SPAM; Durso, Dattel, Banbury, & Tremblay, 2004) - these are arguably the most valid measures of SA (Endsley & Garland, 2000b) - are not well suited for laboratory setting and are considered to be rather intrusive. Eye tracking has also been used to get an indication of SA quality (e.g. Moore & Gugerty, 2010; Yu, Wang, Li, & Braithwaite, 2014), but is difficult to use for field studies, given the setup time for fixed eye tracker systems or the manual effort necessary to extract data from wearable sensors. Several attempts have been made to find the neural correlates of SA using electroencephalography (EEG; e.g. Berka et al., 2006; French, Clarke, Pomeroy, Seymour, & Clark, 2007; Vidulich, Stratton, Crabtree, & Wilson, 1994), but these attempts have not provided fruitful results. For a more complete review on the upsides and shortcomings of different measures of SA, please see Endsley (2013) and Salmon, Stanton, Walker, & Green (2006).

There is thus a need for a new measurement tool that circumvents these issues with and intrusiveness and reliability. In this exploratory study, we took a novel approach using EEG as a new online objective measure for SA. More specifically, we sought to establish the relationship between EEG metrics of attention and the quality of SA in a high fidelity TTC simulator. Because the ultimate goal is to conduct EEG research in field studies, we repeated the experiment to allow for a comparison of a medical grade EEG system with a wearable, more usable, EEG system. Moreover, a pilot study was conducted with professional TTCs to investigate how EEG research is perceived by trained professionals and how they respond to SA research.

Related work

The job of train traffic controllers

The job of TTCs is highly automated by a system which manages train traffic and assign these trains to the correct tracks and platforms such that flow is optimal and delay is minimal. It is this system that gives directions to the train drivers. The system cannot cope with external events such as defect trains or personnel that is late for a departure. When a train is behind schedule by three minutes or more, the train will be marked as delayed. The planning lines, which contain all future actions for that particular train (see Figure 1), will subsequently turn red and automated control will be removed for that train.

The TTC must then immediately become aware of the situation and find a solution to reduce the delay or, if that is not possible, minimize the conflict with other trains. This solution must then be manually implemented and communicated to the relevant parties. If a fitting solution is found, then control can be given back to the automated system. However, the longer it takes for a TTC to notice these red lines, the more difficult it may become to find that solution and the larger the consequences may be. It is thus imperative for TTCs to stay attentive and maintain a high quality SA.

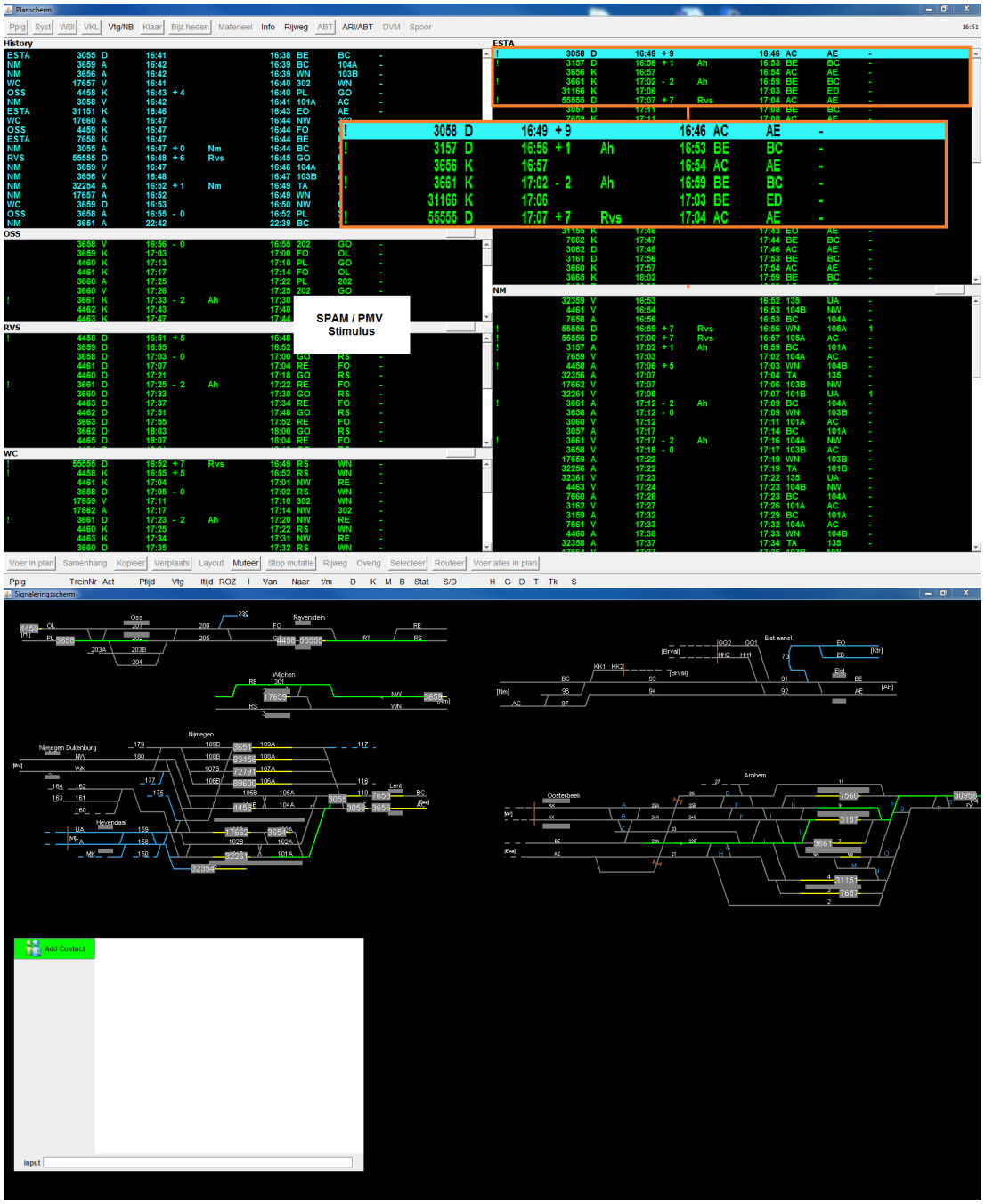


Figure 1: Graphical user interface of the TTC simulator. The top screen is the “planning screen”, with “planning lines” that contain the planned action and time for a train. The white box is the location where the SPAM button and PMV stimuli were depicted. In the bottom “signaling screen” a helicopter view is depicted of the train tracks of the operators responsible area. Moreover, the chat interface was shown on this screen. During the experiment, the two screens were presented side-by-side.

Measuring situation awareness as a product

Many different definitions of SA exist and deciding for one of the definitions can determine the focus of the research and methodology (Durso & Sethumadhavan, 2008; Salmon et al., 2006). Most SA measurement techniques are based on a different, perhaps more widely used model of SA: the three level model Endsley (1995). This model states that SA is "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1995, p. 36). SA consists thus of three levels: perception, comprehension and projection, also referred to as Level 1, Level 2 and Level 3 SA respectively. This model considers SA to be a product, which can be assessed by testing the operator's knowledge of the situation. One popular means to do so is the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1988). The SAGAT is a query based tool that asks a set of questions and tests what the operator could see of (Level 1), understands about (Level 2), or predict from (Level 3) the current situation. The questions are constructed based on a goal-directed task analysis (Endsley, 2013) or a subject-matter expert (Salmon et al., 2006) and are always situation specific. Beside the fact that constructing SAGAT queries is a time consuming process, the SAGAT is incredibly intrusive and is not well suited for field studies. For the operator to be able to answer these questions, the task must be paused and the interface hidden until the questions are answered.

In an attempt to make the measurement of SA less intrusive, the Situation Present Assessment Method (SPAM; Durso, Dattel, Banbury, & Tremblay, 2004) was developed. The SPAM is periodically test the operator's SA by asking a similar questions as the SAGAT while the task is ongoing. Besides correctness of the answer, response time (RT) is recorded. Short RTs correspond to fast retrievals of information or quick searches of information on the screen and, therefore, corresponds to high quality SA. Long RTs, on the other hand, are associated with not knowing the answer and not knowing where to find the answer and, therefore, corresponds to low quality SA. By using the SPAM, however, researchers add a secondary task, which may increase the amount of workload and, in turn, affect performance on the primary task (Pierce, 2012). Another drawback of the SPAM is its temporal resolution; it is advisable to only ask one question once every two to three minutes as to not interfere with the task too much (Durso et al., 2004; Pierce, 2012).

Measuring situation awareness as a process

In this paper, instead of the product view, we adopted the definition of Smith & Hancock (1995), which is based on the perceptual-cycle model of Neisser (1976). Smith & Hancock (1995) argue that SA is a cyclical process of attention, in which perception, sense-making, anticipation and action-taking take place concurrently, in order to generate a 'product of consciousness', that is, an understanding of the current situation. Measuring one's SA implies measuring how well the acquisition of SA is over time (Salmon et al., 2006). Even though the perceptual-cycle model argues that SA is a process, it does acknowledge the existence of a product of consciousness. This implies that it is justifiable to measure the product and relate this to the process of SA. Because the goal of this paper is to find a measure that captures SA objectively in real-time, adopting the perceptual-cycle model was considered to be more appropriate.

Tracking eye movements and measuring where and how long people fixate their gaze has been one fairly successful endeavor for assessing the quality of SA (e.g. Moore & Gugerty, 2010). However, as Salmon et al. (2006) discusses, it is difficult to use it for field studies, given the setup time for fixed eye tracker systems or the manual effort necessary to extract data from wearable sensors. Moreover, people do not always perceive the information that they focus on, as illustrated by the look-but-failed-to-see accidents (Crundall, Crundall, Clarke, & Shahar, 2012). This could, at least in part, be explained by a lack of attention (Ruby, Smallwood, Sackur, & Singer, 2013).

Another approach for measuring SA with physiological measures was attempted with electroencephalography (EEG), but these have not produced fruitful results. French et al. (2007) adopted Endsley's model of SA and marked the presentation of stimuli as either Level 1 SA (when irrelevant stimuli were presented), Level 2 SA (when stimuli were presented that were immediately relevant) and Level 3 SA (when the information was relevant for the overall mission). Following, a discriminant analysis tried to classify the stimulus markers based on the power spectral density (PSD) of the EEG data around those events, but accuracy was poor. The problem with this approach is that the researchers disregarded

how the participants acted upon the stimuli and, therefore, could not show behaviorally whether the stimuli were in fact related to different levels. Moreover, they ignored the quality of SA at each of these levels. In a different study, Berka et al. (2006) also used events that were related to Endsley's levels of SA. They compared event-related potentials (ERPs) and the PSD (1) between moments of correct and incorrect target identification, and (2) between reading questions and reading information. With these events, more focus was put on Level 2 and 3 SA, and they distinguished bad from good SA. The problem of this approach, however, is that it looks at event-related activity; it is not possible to extract these events in real-time and, therefore, impossible to implement such a system for field studies.

Vidulich et al. (1994) took a completely different approach that is more in line with the perceptual-cycle model. They manipulated the display in a target-identification task, to facilitate target identification to a greater or lesser degree. Again, PSD was calculated for individual channels. Results showed that theta power (4-7 Hz) was higher and alpha power (8-14 Hz) was lower in many channels in the most difficult conditions compared to easier conditions, which is consistent with higher attentional demands. Unfortunately, the paper was not able to determine how task difficulty had affected the quality of SA, thereby limiting the implications of their results.

Situation awareness relies on sustained attention

Much research has been conducted to sustained attention that justify the approach of Vidulich et al. (1994), showing the strong relationship of attention with learning (Niv et al., 2015) and vigilance performance (Donald & Donald, 2015), but attention has also been directly coupled to the quality of SA (Croft, Banbury, Butler, & Berry, 2004; Catherwood et al., 2014; Ratwani, McCurry, & Trafton, 2010). The general consensus is that people perform better and are more aware of what is going on when they focus their attention on task-relevant stimuli. However, when people focus their attention on self-generated thoughts, that is, "[...] mental contents that are not derived directly from immediate perceptual input" (Smallwood, 2013, p. 31.3), people are actually less capable of perceiving these external stimuli (Ruby et al., 2013). This phenomenon is also referred to as perceptual decoupling and can affect performance in a large degree. Especially in jobs that require monitoring and encoding immediate input, such as the work of TTCs, perceptual decoupling, as a result of attending to self-generated thoughts, may have large consequences (Ruby et al., 2013). Note that SA, according to Smith & Hancock (1995), does not rely on sustained attention alone. Retrieving the appropriate knowledge and making the correct decisions is also vital for having a high quality SA. In this study, however, the focus will only lie on the attentional component of SA and its neural correlates.

Neural correlates of sustained attention

Attention has been studied extensively by neuroscientists using EEG and magnetoencephalography (MEG). Research found that power in the alpha frequency band (8-14 Hz) has shown to correlate well with attending to self-generated thought. Van Dijk, Schoffelen, Oostenveld, & Jensen (2008), for example, had participants discriminate a stimulus (a small gray circle that varied in shade) that was superimposed on a mask (a larger gray circle with a fixed shade). During this task, MEG recordings were made. The results showed that an increased alpha power in posterior brain regions was accompanied with a decreased ability to discriminate the stimuli, which is consistent with the expected effect of perceptual decoupling.

Similar results were found in the study of Knyazev, Slobodskoj-Plusnin, Bocharov, & Pylkova (2011). EEG was measured in three different situations: (1) resting conditions with eyes open and eyes closed, (2) during an explicit judgment task where the hostility/friendliness of faces was to be judged and (3) during a social game task in which participants were presented with the same faces and had to say whether they would attack, avoid or make friends with the person of the picture. The authors argued that these 'social cognition tasks' would elicit self-generated thought to a greater or lesser extent. The results showed moderate to high correlations between alpha power and activity in the default-mode network, the network of brain areas associated with attending to self-generated thought (Smallwood, 2013).

In a different line of research, Pope, Bogart, & Bartolome (1995) investigated which EEG metric of task engagement would modulate adaptive automation, that is, determine the level of automation, the best. The different EEG metrics were compared based on the overall performance on the task at hand.

Results showed that participants performed best when the level of automation was determined by the following "task-engagement index" (TEI):

$$TEI = \frac{\beta}{\alpha + \theta}$$

, in which α, β and θ are alpha power (8-14 Hz), beta power (15-30 Hz), and theta power (4-7 Hz), respectively, averaged over channels Cz, Pz, P3 and P4. More specifically, when the TEI dropped, more manual control was given to the operator, whereas the level of automation increased when the TEI rose above a particular threshold. In short, according to the TEI, decreasing alpha power and theta-power, combined with an increase in beta-power, is associated with an increased task-engagement of the operator. The TEI may thus be very closely related to attention, given that beta- and theta-power also correlate well with activity in the default mode network (Scheeringa et al., 2008; Mantini, Perrucci, Del Gratta, Romani, & Corbetta, 2007). These studies show the ability of EEG to capture attentiveness and the value of attentiveness in relation to task performance, not only in controlled experiments but also in an applied setting. Recording EEG data, with a specific focus on alpha power and a combination of alpha-, beta- and theta power, may be the best candidates for measuring attentiveness and relating that to SA.

Current approach

In this study we sought to establish the relationship between the quality of SA and attention, as measured by EEG. Student participants were asked to perform relatively simple TTC tasks in a high fidelity simulator, while their brain activity is recorded. During this task, the quality of SA is assessed by periodically presenting a SPAM query, of which the RTs are recorded. Attention is measured with several EEG metrics calculated over short periods before the presentation of the questions. The EEG data is thus time-locked to, but not evoked by, the stimulus. This allows us to measure task-induced attentiveness, similar to Vidulich et al. (1994), and compare that to the quality of SA, similar to Berka et al. (2006). A psycho-motor vigilance task (PMV task; Van Dongen, Maislin, Mullington, & Dinges, 2003) was added to allow us to inspect attention more frequently throughout the task. The PMV tasks consists of a simple stimulus that is presented on the screen, which must be responded to as quickly as possible by pressing a button. It is hypothesized that RTs on both the SPAM queries and PMV stimuli would be larger when people are less attentive, that is, when they focus their attention on self-generated thought, as an effect of perceptual decoupling. Conversely, when the participants are more attentive, RTs should decrease on both tasks.

Because the ultimate goal is use EEG in the field, a subset of these participants returned for measurements with a wireless and more usable EEG headset: the B-Alert X-10 (Advanced Brain Monitoring), to see if EEG can also be applied for SA related field studies. The B-Alert X10 has shown to provide clean, high-quality data (Berka et al., 2007; Ries, Touryan, Vettel, McDowell, & Hairston, 2014) and outperforms different wearable EEG systems, such as the commercially Emotive EPOC (Ries et al., 2014). To gain a better understanding how TTCs gain SA, a pilot study with a similar experiment was conducted, after which only qualitative data about their experience was analyzed. In the remainder of the paper, the method, results, and a short discussion will be discussed for each experiment separately, followed by a general discussion of their implications and highlight some lessons for future endeavors. A formal comparison of the two EEG systems was also performed, which can be found in Appendix A.

Experiment 1

Method

Participants

A total of 24 subjects (age = 22.25 $\pm$ 2.03 years; 16 female), who were students at the University of Groningen in the academic year 2015-2016, participated in this study for a small monetary reward of twenty euro per experiment of two to two-and-a-half hours. The subjects gave their informed consent and had no known neurological condition or any physical limitation. Two subjects were left-handed, but because the primary target of our study were non-lateralized cognitive functions, the subjects were not excluded. One participant was removed, because the behavioral data of one of the scenarios was not saved. A further two participants were removed due to an excessive amount of noise in the EEG data (removal of almost 25% of either SPAM and PMV trials), which is described in the section 'EEG pre-processing and analysis', resulting in 21 remaining participants.

Another participant stopped after one scenario, but returned the day after to finish the second scenario. Because it was believed that the participant had no apparent benefit, apart from some additional rest and the opportunity to look at the task instructions that was already provided (see the experimental setup), it was decided to not exclude the participant. Another participant's screen shots were not saved, which made it difficult to check the correctness of the SPAM queries (see Measurements). Because the answer to many queries were not determined by the actions of the participant, that is, some answer are fixed between simulations, the answers were checked based on a prototypical scenario.

Task

The students were asked to perform two scenarios in a high-fidelity train traffic control (TTC) simulator with simplified controls (see Figure 1). The scenarios took place around Nijmegen, the Netherlands, and were originally developed for research to workload. Because of the dynamic nature of the task, the scenarios were deemed appropriate for the current research (See also Lo, Sehic, & Meijer, 2014).

The first scenario starts with a freight train that is overloaded and cannot drive at a high speed. Therefore, the road is blocked for subsequent trains, which will end up behind schedule. In order to have the freight train affect not all traffic, the participant must manually manage the train through the area. After five minutes the participants is asked to let the freight train wait at a train station for a while, so other trains can depart before it. Six minutes later, before the train has arrived at the station, the freight train driver asks whether he may continue driving, to maintain his momentum. At this moment the participant must decide for either option. If the participant decides to have the freight train wait, the train will block parts of the track from the 20th minute onward, because of its (unexpectedly) large length. This will block multiple trains that causes a bigger delay. After 30 minutes, this scenario is finished.

The second scenario goes as follows: In the first eleven minutes, trains will arrive and depart with minor delays, which should not cause any major troubles. After this period, a freight train driver will notify the operator that it has no traction and is, therefore, unable to move, thereby blocking the passage way for multiple trains. After ten minutes, the freight train can move again, but a level-crossing failure will occur, caused by the blocked trains. A level-crossing failure means that the crossing barrier remains closed for a longer period of time, caused by trains standing still nearby. Because pedestrians and bicyclists may grow impatient and pass the closed level-crossing, the train drivers must be informed to drive slowly to avoid any incidents, according to standard protocol. After 30 minutes in total, the scenario is finished.

Communication with the train drivers, among others, is a vital part of the task of a TTC and normally occurs via telephone. To reduce muscular artifacts in the EEG data from talking, a chat-bot was implemented. It must be noted that, although both scenarios follow a particular script, the scenarios can develop slightly differently, depending on the decisions the subjects make.

During these scenarios two additional tasks had to be performed. First of all, SPAM queries were to be answered as a measure of SA. The SPAM queries were presented at predefined moments in the scenarios. The moments of the questions were distributed in such a way that questions were asked during both high- and low engagement moments. This was inferred from the amount of workload the TTC task would impose on the participants. It was expected that the participants were more engaged

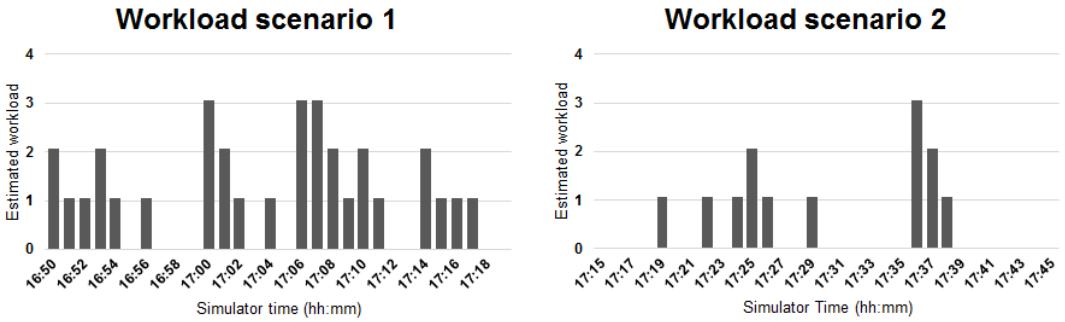


Figure 2: Minute-by-minute graphical representation of the estimated workload distribution throughout the two scenarios. Workload is shown on an arbitrary scale of 0 to 4, where 0 corresponds to no workload and 4 to maximum amount of workload. On the x-axis, simulator time is depicted, that is, the local clock within the simulator. The moments SPAM queries were presented can be found in Appendix B.

and would experience a higher workload in the task when manual control was necessary, and would be less engaged during periods of monitoring. Based on unpublished workload research, and the subjective ratings of two human factors experts, the amount workload throughout the scenarios was estimated (See Figure 2). The content of the SPAM queries were based on a SAGAT queries as used in earlier research (Lo et al., 2014), and on the situation at hand and the relevant decisions to be made. The queries only cover the perceptual component of SA (i.e. Level 1 SA in Endsley's terms), meaning that that only low level relevant information, which is always present on the screen, is asked for. This way, the focus lies solely on attention. A query could ask, for instance, for the planned departure time of train X at station Y. Appendix B displays the timing and content of the queries used for the subjects.

Because the amount of SPAM queries is limited to only one question every two to three minutes (Pierce, 2012), a psychomotor vigilance (PMV) task (Van Dongen, Maislin, Mullington, & Dinges, 2003) was added every 30 ± 3 seconds, unless it interfered with a SPAM query. A stimulus was to be responded to as quickly as possible by pressing the left CTRL button on the keyboard. This allowed us to inspect how attentiveness varied more continuously throughout the task and how this would be expressed behaviorally, that is, in response time.

Experimental setup

After having registered for the experiment, the participants received a nine-page document with basic and necessary information about the task of TTCs and how the simulator works. There is a fair amount of information that the participants need to know and having read the document before the experiment allowed the information to "sink in". Therefore, the participants had a little more context when they arrived at the experiment, which would make understanding the instructions at the time of the experiment easier.

After the procedure for EEG measurements was described and the participant signed the informed consent, they viewed an instructional five-minute video with the same background information on the task as in the document. Following, a practice scenario of 30 minutes was started during which the participants received some additional verbal instructions and they could get used to the interface. During this practice scenario, impedance of the EEG channels was inspected to ensure high quality recordings. If the participants were confident enough about their understanding or when the scenario was finished, they continued to the two experimental scenarios. The order of the scenarios was counterbalanced between participants to account for any order effect.

Measurements

Before each SPAM query was presented, a gray box was shown on the screen. After clicking the box with the cursor, the query was shown with four possible answers. The quality of SA was measured by the response time (RT) to the correctly answered questions, starting from clicking a gray box. This ensures that the RT is only based on the time to find the answer, and excludes the time to perceive

and click the box. At the moment the gray box was presented, a screen shot was made which allowed the researchers afterwards to check whether the response was in fact correct. If the gray box was not clicked after fifteen seconds, the box disappeared and the question was registered as a miss and excluded from further analyses; after fifteen seconds (plus the time to answer the question), the situation could have changed in such a degree, that the screen shot would not accurately represent the current situation anymore.

The correct answer of a SPAM query could always be found on the screen and could, in theory, always be answered correctly. If a query was answered incorrectly, the query was removed from further analyses. There are many different reasons why a question would be answered incorrectly which are unrelated to attention: people may have accidentally clicked the wrong button or they might have had a wrong understanding of the task or question. The lack of SA is then not limited to the attentional related issues, which is the focus of this research, but may primarily be associated with knowledge and skill. Comparing the correct with the incorrect trials would thus little insightful information about the relation between attention and SA.

RT to the PMV task was recorded as a measure of attentiveness more continuously throughout the task. If the alarm was not responded to when the next alarm was planned, then the first alarm would be registered as a miss and removed from further analyses.

Equipment

The simulator was run on a HP desktop computer, connected to two Philips 220BW Brilliance monitors. The participants' brain activity was recorded with the BioSemi ActiveTwo 128-channel EEG system, a medical grade EEG system, in combination with the ActiView software. Before starting the measurement, impedance values were kept below 40 $k\Omega$ to ensure high quality data. Data were collected with a sample frequency of 512 Hz and were filtered online with a 0.16 Hz high-pass filter, accompanied with a 100 Hz low-pass filter.

EEG pre-processing and analysis

The open-source Matlab toolbox, FieldTrip, was used to process the acquired data (Oostenveld, Fries, Maris, & Schoffelen, 2011). The data were segmented into trials of five seconds – four seconds pre-stimulus and one second post-stimulus – after which a low-pass filter (50 Hz), notch-filter (49-51 Hz) and a high-pass filter (1 Hz) were applied to correct for high-frequency noise, line noise and drift respectively. Several range thresholds were tested to allow for automatic artifact recognition, after which the recognized trials were compared with the authors visual judgment for two participants. This resulted in a range threshold of 450 μV. Trials were removed if this threshold was exceeded within the three second pre-stimulus to one second post-stimulus period and were not caused by an eye blink, which was determined based on visual inspection. Trials were also removed if the data was not properly transmitted, which happened when an over/under-current was detected. If more than fifteen percent of the trials had to be removed, the participant was excluded from the analysis, which happened for one participant. For some participants, a couple of channels were marked as noisy channels when there was a continuous high-frequency noise are large drifts (i.e. low frequency fluctuations) throughout the measurement that persisted to exist after filtering. This was likely caused by salt-bridges or poor connection with the scalp. These channels were therefore not considered during artifact correction.

Following, the data was subjected to an independent component analysis (ICA; Bell & Sejnowski, 1995) for artifact correction. The ICA identifies similar patterns of data over all the channels, and combines these patterns into as many components as there are channels. All of these components are orthogonal to each other, therefore ensuring complete independence of the components. Component, that showed clear signs of blinks, eye movements, high frequency (muscular) activity, EKG components or drift, were corrected for by removing the components. If more than 25 percent of the components had to be removed, the participant was removed from the analysis. We decided for this liberal value, because the task (i.e. monitoring two screens) is by nature accompanied with some head movements and, therefore, a fair amount of muscular artifacts. One participant was removed because of this, resulting in 21 remaining participants. The individual channels that were marked as noisy channels during visual inspection were left out of the ICA. Only after the ICA, were the data of these channels replaced by the average activity measured at neighboring channels.

Finally, the segments were subjected to a time-frequency analysis with wavelet convolution. This analysis calculates the power changes over time at the highest resolution possible (1/frequency) in the different frequency bands (delta = 0-4 Hz, theta = 4-7 Hz, alpha = 8-14 Hz, beta = 15-30 Hz). In this study, we looked at alpha-power and the Task-Engagement Index as metrics for attentiveness.

Statistical analysis

The idea is that when participants are very alert, they respond more quickly to the PMV probes, whereas when they are less attentive, they respond more slowly. We thus split the trials based on the median value; trials with RTs smaller than the participant‘s median RT are assigned in the low RT group, and trials with RTs equal or higher than the participant‘s median RT in the high RT group. Following, we performed a cluster-based permutation test to identify any particular channel or cluster of channels with a significant difference in alpha-power. To correct for the multiple-comparison problem, a Monte-Carlo randomization procedure was applied (Maris & Oostenveld, 2007), which compares the observed statistical results to those achieved in data sets with randomly permuted labels.

In order to find a relation between attention and SA, we applied a linear mixed-effect regression (LMER; Bates, Mächler, Bolker, & Walker, 2014). This method allows the construction of linear regressions with fixed effects, that is, independent variables, and the easy addition of random factors. The addition of random factors, such as between subject or between gender variability, allows the explanation of variance that would otherwise be part of the error term of a typical linear regression. The benefit of this model over, for example, an ANOVA is that it allows a lot of flexibility and takes the full data set into account, as opposed to averaging over groups or trials. A LMER is typically constructed as follows:

$$y \sim x + (1|z) \qquad \text{(Model 0)}$$

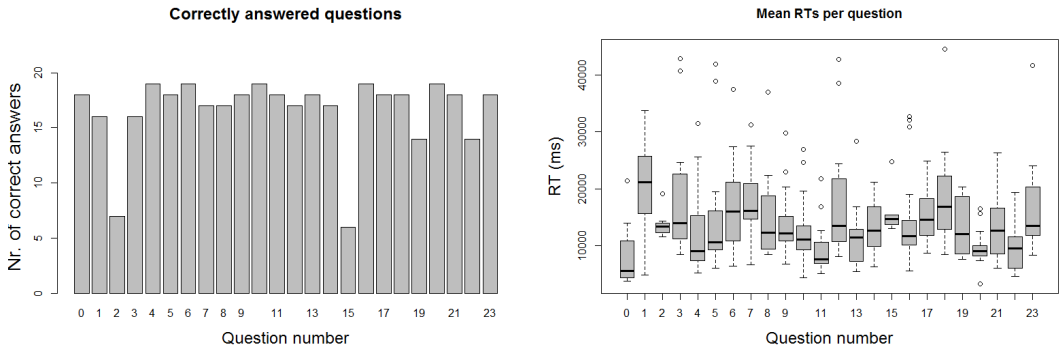
This model states that variable y is explained by fixed factor x and random factor $(1|z)$. The notation of the random factor states that for each value in factor z, e.g. male and female in factor gender, a different intercept is given. In this study, we consider RTs on the SPAM as the dependent variable and EEG metrics of attention as the fixed factor.

Results

Behavioral data

The goal of this research was to find a relation between attentiveness, as measured with EEG, and the quality of SA, as measured by RT on the questions. To do this, we intended to construct LMER in which EEG metrics of attention are used as fixed factor and RTs as dependent variable. A possible confound in this analysis is question difficulty, which may also lead to higher RTs. To rule out that differences in RTs are the result of question difficulty, we examined whether there were significant differences in the RTs to different questions. For each question we calculated the amount of times a question was answered correctly and the median RT. RTs lower than 500 ms were removed from the analysis, for they were likely guesses or a quick accidental click. The 500 ms may appear to be a long period, but the SPAM questions also require the participants to think of the answer, instead of merely clicking a button.

Figure 3a shows the amount of times that a question was answered correctly and Figure 3b the median and distribution of the RTs on each question. Questions 0-11 and 12-23 respond to the first and second scenario, respectively. As you can see from Figure 3a , participants had difficulties with answering question 2 and 15 correctly. These questions both correspond to: “How many platform tracks are available in Nijmegen?”. One explanation for the fact that this question is so often answered incorrectly, is that participants have misunderstood the instructions or the question. The small variability in RTs of these two questions compared to other questions (see Figure 3b) does suggest that the participants were confident of their answer and found the answer with relative ease. Because the actual reason for the incorrect answers is unclear, however, the questions were removed from further analysis. A one-way ANOVA on the remaining question showed that there was a significant main effect on RTs ($F(21,362) = 3.904$, $p < 0.001$). In other words, some questions took significantly more time to answer than others. Therefore, when investigating the relationship between EEG metrics of attention and SA, we must correct the SPAM RTs for question difficulty.



(a) (b)

Figure 3: **a.** Number of times each SPAM query was answered correctly. **b.** RT distribution per query in milliseconds. Queries 0 - 11 correspond to the first scenario, and queries 12-23 correspond to the second scenario.

Another factor that may affect the strength of the relation between alpha-power and SA quality is whether attentiveness actually varies throughout the task. If attention is constantly high or constantly low, then the chances of finding a relation between SPAM RTs and alpha-power are slim. Therefore, the next step is to see how attentiveness varies throughout the task, as measured with the SRT task. In Figure 4 , the median and distribution of the RTs on each PMV alarm are shown. Alarm 1 to 47 are part of the first scenario, and alarm 48 to 94 are part of the second scenario.

For depiction purposes, only the RTs below five seconds are shown. In reality, RTs go up to 22.3 seconds. As you can see, the median RTs appear to vary a fair amount between the different probes, suggesting that at some points during the task the participants required more time to perceive and respond to the probes. This is confirmed by a one-way ANOVA, which showed a significant main effect of PMV probe (F(93,1788) = 3.140, p <0.001). The ANOVA was corrected for unequal variance, because a Levene's test for homogeneity of variance confirmed that equal variance could not be assumed (F(93,1788) = 1.612, p <0.001). The error-bars and the amount of outlier data-points gave us reason to believe that an assumption of homogeneity of variance could be violated. This unequal variance may be explained by two reasons: First, the job takes place in a dynamic environment in which different decisions may be taken at different moments in time. The scenarios may, therefore, develop slightly differently between participants. The experienced workload throughout the task may thus also vary between participants. Secondly, participants may be inherently quicker or slower than others. An inspection of the distributions of RTs per participant (See Figure 5) suggested that RTs indeed vary between participants, which is confirmed by a one-way ANOVA, which showed a significant main effect of participant (F(20,1861) = 4.129, p <0.001). Again, only RTs below five seconds were included in the figure for depiction purposes. Both within- and between subject, there is thus a reasonable amount of variability in RT, suggesting a reasonable amount of variability in attentiveness. These individual differences must also be considered in subsequent analysis.

Attention and EEG

As discussed in the method, we split the PMV trials into a high- and low RT group, based on the median RT of each subject. Following the alpha power was calculated over the three seconds prior to presenting each PMV alarm and subsequently averaged over each RT group. Figure 6 shows the difference in alpha power (high RT trials – low RT trials) plotted over the scalp, averaged over all participants. Yellow areas indicate channels in which alpha power was more or less equal in both groups of trials, whereas green and blue areas indicate channels in which alpha power was lower in high RT trials. A cluster-based permutation test was performed to examine whether any within subject differences were significant, while being corrected for the multiple comparison problem with a Monte-Carlo randomization procedure. This showed that, in the posterior regions, alpha power was not significantly higher in the high RT trials, as was hypothesized. In frontal brain regions, on the other hand, alpha power in the

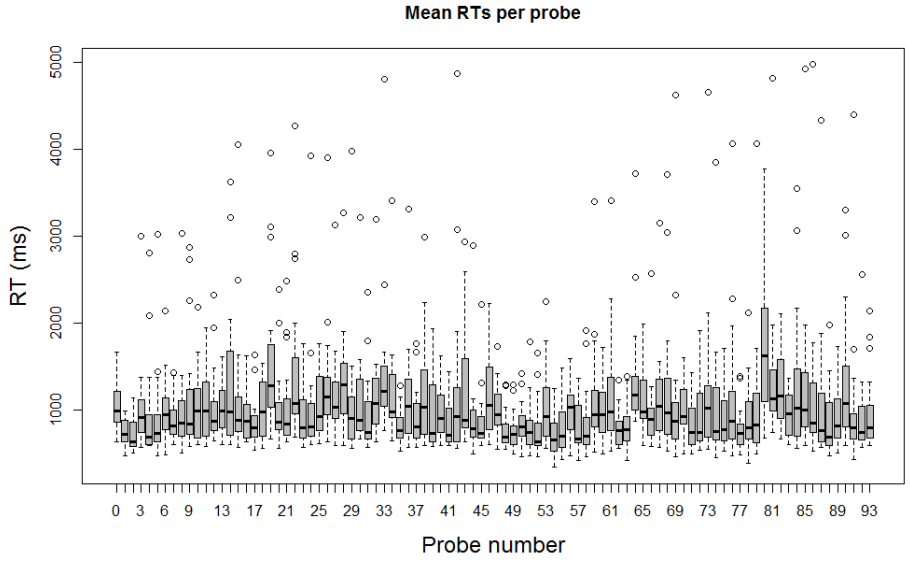


Figure 4: RT distribution per PMV probe in milliseconds. For depiction purposes, only the RTs below five seconds are shown. In reality, RTs go up to 22,300 milliseconds.

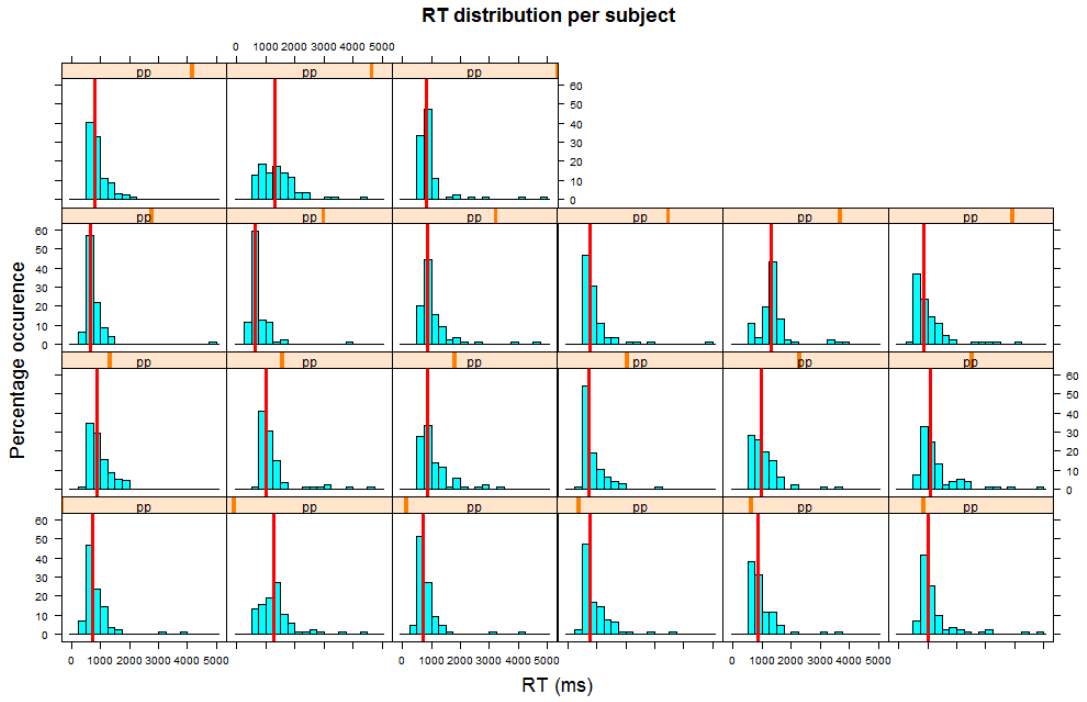


Figure 5: Histogram of RTs in milliseconds on the PMV probes, depicted for each participant separately. The red vertical lines shows the median RT for each participant.

high RT trials was significantly lower, compared to low RT trials ($p = 0.004$), which is in conflict with the hypothesis. The fact that a difference was found in temporal and frontal regions gave us reason to believe we should extend the analysis to global alpha power, as opposed to merely posterior regions, when looking at the relation between SA and attention.

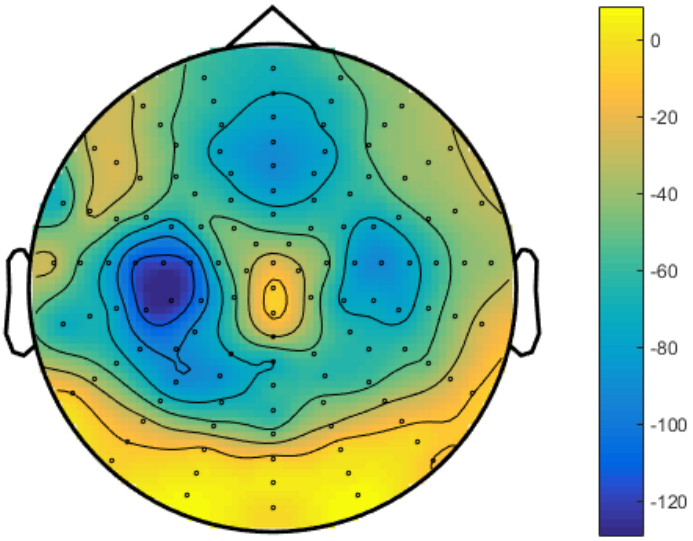


Figure 6: Grand-average scalp topography of the power differences in the alpha-frequency band (8 – 14 Hz), calculated over a three second pre-stimulus period. Green and blue areas indicate lower alpha power, consistent with higher attentiveness, in the high RT trials compared to low RT trials. Yellow areas are consistent with areas that showed little or no difference in alpha power between trials.

SA and EEG

As mentioned earlier, significant differences in RT were found between questions, suggesting that question difficulty may have played a role. To correct for question difficulty, a linear mixed-effect regression (LMER) was constructed, in which between-question and between-subject variability in RT were accounted for as random factors. The "random" LMER was defined as follows:

$$RT \sim (1|subject) + (1|question) \qquad \text{(Model 1)}$$

This model is compared with Model 2:

$$RT \sim \alpha + (1|subject) + (1|question) \qquad \text{(Model 2)}$$

, which states that RT is explained by alpha power α, averaged over channels POz, P3 and P4, as was our initial hypothesis, and the two random effects. Model 2 did not find a significant effect of alpha power (β = -0.369, SD = 0.574, t = -0.643), suggesting that no clear relation between RT and alpha power exists. Generally, for a fixed factor to be considered significant, a t-value of at least two is expected.

The explanatory power of the models were compared using the AIC, BIC and log-likelihood tests, which penalizes models that have more degrees of freedom, in accordance with Occam's razor. A model is considered better when the AIC- and BIC scores are lower, and the log-likelihood is higher. The results showed that Model 2 is not a significant improvement over Model 1 (see Table 1, Model 2:α-avg). The log-likelihood of both models are nearly the same and the AIC and BIC values are higher for Model 2. These values, plus the non-significant effect of alpha power, indicate that alpha power does not reliably predict RT on SPAM queries better than only random effects.

We compared the results of the averaged posterior alpha power with the the global alpha power (i.e. averaged over each channel), and to individual channels POz and Oz, to see if those channels provided better explanatory power when used in Model 2. For both the individual channels, the effect of alpha power was non-significant (β_{POz} = -0.122, SD_{POz} = 0.4687, t_{POz} = -0.261; β_{Oz}= -0.163, SD_{Oz} = 1.222, t_{Oz} = -0.133) and Model 2 was not a significant improvement over Model 1 (See Table 1). The same goes for global alpha power; no significant effect of alpha power was found (β_{global} = -0.122,

Table 1: Linear Mixed Effect Model comparison results. Each Model 2 is compared against the random Model 1.

	DF	AIC	BIC	log-lik	χ^2	Df	p
Model 1	4	7901.2	7917.0	-3946.6			
Model 2: α-avg	5	7902.8	7922.5	-3946.4	0.414	1	**0.520**
Model 2: α-POz	5	7903.1	9722.9	-3946.5	0.069	1	**0.793**
Model 2: α-Oz	5	7903.2	7922.9	-3946.2	0.018	1	**0.893**
Model 2: α-global	5	7902.8	7922.5	-3946.4	0.390	1	**0.533**

SD_{global} = 0.4687, t_{global} = -0.635), nor was the model an improvement over Model 1. None of the models were thus able to distinguish itself from another. However, because the analysis on the PMV-task data showed the largest deviations in alpha power in frontal regions, the subsequent analyses were focused on the global alpha power.

Subsequently, we verified whether the three second pre-stimulus interval of EEG data was chosen well, by comparing the explanatory power of the global alpha power with a two-second interval. Again, there was no significant effect of alpha power (β = -0.552, SD = 0.591, t = -0.933), and Model 2 was no improvement over Model 1 (see Table 2), suggesting that a random model is equally well at predicting RT on SPAM queries. The two-second model is thus not able to distinguish itself from the three-second model, and was subsequently dropped from further analysis.

Table 2: Linear Mixed Effect Model comparison results. Model 2, constructed with global alpha power calculated over a two-second period, is compared against the random Model 1.

	DF	AIC	BIC	log-lik	χ^2	Df	p
Model 1	4	7901.2	7917.0	-3946.6			
Model 2: α-global-2s	5	7902.8	7922.5	-3946.4	0.389	1	**0.533**

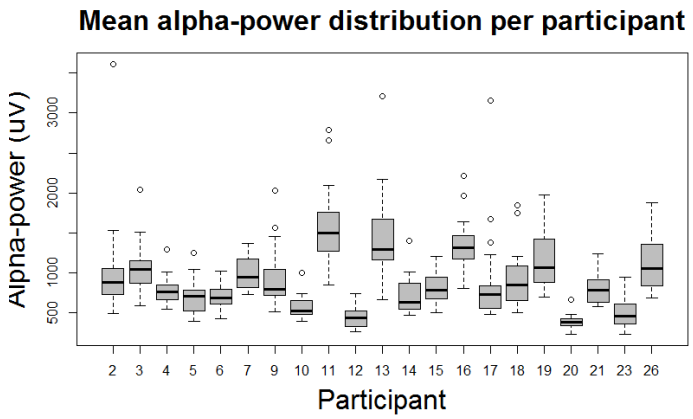


Figure 7: For each participant, the distribution of global alpha power, calculated over a three-second pre-stimulus intervals. Actual participant numbers are used, and are thus not incremental from 1 to 21.

Afterwards, a z-transformation of the alpha power, such that the mean power is zero and one standard deviation equals to one, made the data better suitable for finding an effect in a LMER. As can be seen in Figure 7, the distribution of alpha power differs to a large extent. Where 1000 μV is already quite high for participant 10 and 12, it is considered as low activity for participants 11 and 16. The LMER may, therefore, have had difficulties finding a trend. The z-transformed alpha power was put in a the LMER, which was defined as follows:

$$RT \sim \alpha Z + (1|subject) + (1|question) \qquad \text{(Model 3)}$$

, in which αZ is the z-transformed global alpha power. The model found no significant effect for αZ (β = -158.2, SD = 344.0, t = -0.460), and Model 3 was no improvement over the "random" Model 1 (See Table 3).

Perhaps alpha power alone is not sensitive enough to measure EEG activity. Therefore, we repeated the analysis for the task-engagement index (TEI), earlier defined as the ratio of beta power to alpha plus theta power, and therefore includes a wider variety of brainwaves to explain attentiveness. The model was specified as follows:

$$RT \sim TEI + (1|subject) + (1|question) \quad \text{(Model 4)}$$

, in which TEI is the Task-Engagement Index. However, Model 4 also was unable to find a significant effect of TEI (β = 15649, SD = 10492, t = 1.491) of which the direction is also in conflict with the hypothesis; the model suggests that highly engaged people have worse SA, as measured by SPAM RT. Table 3 also confirms that Model 4 is not a significantly better model than Model 1. The high β and SD value is the result of the small range of TEI values, which are in the order of 0.1, whereas RTs are in the order of thousands. Again, including TEI as a fixed factor only increases the number of degrees of freedom, without the necessary improvement.

Table 3: Model comparison of "random" Model 1 with Model 3, constructed with z-transformed alpha power as fixed-factor, and Model 4, constructed with the TEI as fixed factor.

	DF	AIC	BIC	log-lik	χ^2	Df	p
Model 1	4	7901.2	7917.0	-3946.6			
Model 3: αZ-global	5	7903.0	7922.7	-3946.5	0.210	1	**0.647**
Model 4: TEI	5	7900.9	7920.7	-3945.5	2.232	1	**0.135**

SA and longer-term EEG

The analyses so far have failed to show the hypothesized relation between SA and attention. One possible explanation is the fact that a period of three seconds is too short and does not capture SA that was tested with that particular query. Possibly, EEG data over a longer period of time is required to gain a global indication of the participants' attentiveness. This was done by taking EEG-data segments that were already available: the three-second intervals before each attended PMV stimulus. Over a period of two to three minutes before each SPAM query, three to five PMV probes were presented. The average EEG data from these probes served as a sample of attentiveness over that two-to-three minute pre-SPAM period (See Figure 8). This data, from now on referred to as blocked data, was then subjected to a similar analysis as described above.

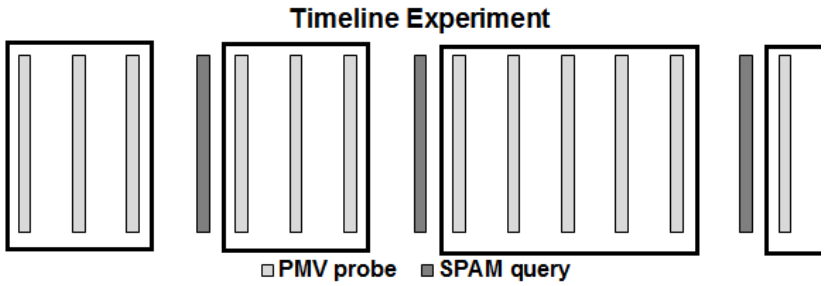


Figure 8: Graphical representation of a part of the task. The black boxes encapsulate the PMV probes of which the three-second intervals of EEG data were taken to calculated the blocked EEG data.

The analysis was limited to the global alpha power, the z-transformed global alpha power and the TEI, calculated over a three-second pre-stimulus period. This resulted in three LMER models:

$$RT \sim \alpha_{block} + (1|subject) + (1|question) \quad \text{(Model 5)}$$

$$RT \sim \alpha Z_{block} + (1|subject) + (1|question) \quad \text{(Model 6)}$$

$$RT \sim TEI_{block} + (1|subject) + (1|question) \quad \text{(Model 7)}$$

Model 5, Model 6 and Model 7 were not able to find a significant effect of blocked mean alpha power (β = 0.028, SD = 1.298, t = 0.022), z-transformed alpha power (β = 63.42, SD = 346.10, t = 0.183) and blocked TEI (β = 18432, SD = 14987, t = 1.230) respectively. The results of the comparisons of these models are shown in Table 4 , which show that the EEG metrics of attentiveness do not improve the fit of the models. Therefore, the previous comment stating that a longer period could provide better data appeared to be incorrect. When comparing results of the blocked models to the results of the original three-second-interval models, no substantial differences were found for the alpha power and TEI models. We did find a noticeable difference between Model 3 and Model 6, the models for z-transformed alpha power. Whereas Model 3 showed a non-significant inverse relation between z-transformed alpha power, a non-significant positive relation was found in the blocked Model 6, the latter being in correspondence with our hypothesis. Because both effects were non-significant, care must taken when interpreting the results, but the switch may suggest that it may be of interest to investigate attention over longer periods of time.

Table 4: Linear Mixed Effect Model comparison results. Model 4, 5 and 6 are compared to the random Model 1.

	DF	AIC	BIC	log-lik	χ^2	Df	p
Model 1	4	7901.2	7917.0	-3946.6			
Model 5: blocked α-global	5	7903.2	7922.9	-3946.6	0.001	1	**0.974**
Model 6: blocked αZ-global	5	7903.1	7922.9	-3946.6	0.031	1	**0.861**
Model 7: blocked TEI	5	7901.7	7921.4	-3945.8	1.497	1	**0.221**

Discussion

On the PMV task we found a relation between EEG metrics of attention and RT that was in conflict with our hypothesis: an increase in RTs is accompanied with a decrease in alpha power. In other words, people have more difficulty seeing the stimuli, when they are more attentive. It has indeed been shown that under higher levels of workload, people are likely to miss peripheral stimuli as an effect of attentional narrowing (Sheridan, 1981; Lavie, Beck, & Konstantinou, 2014). Given the fact that attention may also be modulated by global alpha oscillations (Sauseng et al., 2005; Laufs et al., 2003), being highly focused on the TTC task could explain the inverted relation between alpha power and RT on the PMV task, found in the frontal brain areas. Because workload is modulated by experience (Patten, Kircher, Östlund, Nilsson, & Svenson, 2006), it is possible that in Experiment 2, when the participants are more experienced, workload could have moderated and, in turn, the expected relationship between attention and RT on the PMV probes may be found.

We have been unable to find any effect in the expected direction between attentiveness, as measured by EEG, and SA. Due to inexperience, the students may not have been able to recognize pieces of information as relevant or irrelevant. In other words, regardless of how attentive a subject was, no SA was gained due to the lack of appreciation of the importance of pieces of information. Therefore, it seems likely that the SPAM queries probed the students to search for that information, instead of the TTC task itself. If the students were more experienced and had a general idea of what the task-demands are, it may be possible that they could distinguish the relevant and irrelevant information, and, therefore, are capable of gaining SA. Both issues will be addressed in Experiment 2.

Experiment 2

Method

Participants

To inspect the applicability of EEG in the field, eleven participants (age = 21.82 $\pm$ 2.04 years; 6 female) were asked to return for a second measurement with a wearable, more usable EEG system such that the EEG results from a medical grade system and the wearable system could be compared. Three participants were removed because they failed to finish the experiment, due to technological difficulties of various sources. For one participant the screen shots of one scenario was not saved. The correctness of the answers for that scenario were again checked based on a prototypical scenario.

Task and experimental setup

A brief version of the verbal instructions was given to refresh the participants' memory, and the participants were allowed to practice for a few minutes with the practice scenario. The order of the scenarios were reversed for each subject, to account for order effects.

Measurements

The quality of SA was measured by the RT on the correctly answered questions, starting from clicking the gray box. If after fifteen seconds the button was not clicked, the query was marked as a miss. If a question was answered incorrectly, the query was also removed from further analyses, because the lack of SA is then not limited to attention, but may primarily be associated with the users' understanding.

RT to the PMV task was recorded as a measure of attentiveness throughout the task. If the alarm was not responded to when the next alarm was planned, then the first alarm would be registered as a miss and removed from further analyses.

Equipment

The simulator was again run on the same HP desktop computer as in Experiment 1, connected to two Philips 220BW Brilliance monitors. The participants' brain activity was recorded with the ABM B-Alert X-10 wireless 9-channel EEG system, in combination with the B-Alert Live software. A consequence of the wearable B-Alert X10 is that it is more difficult to ensure these high quality connections with the skin. Therefore, impedance values below 70 $k\Omega$ were deemed sufficient. Data were recorded with a sample frequency of 256 Hz.

EEG preprocessing and analysis

Some channels of the B-Alert, that were not used in this study, were sampled at different rates. In order to match the sample rate of each channel, including the EEG channels, each channel was up-sampled to 1024 Hz; this was done automatically by the FieldTrip software when the data was imported. Following, the data were subjected to an identical preprocessing and analysis as in Experiment 1. That is, EEG segments from four seconds pre-stimulus to one second post-stimulus were time-locked to the presentation of the grey box for the SPAM-queries and to the presentation of the alarm of the PMV task. These segments were filtered with a low-pass filter (50 Hz), a notch filter (49-51 Hz) and a high-pass filter (1 Hz). Trials of which the range exceeded 450 μV were marked as potential artifacts, which was followed by visual inspection; trials were removed if this threshold was exceeded within the three second pre-stimulus to one second post-stimulus period and were not caused by an eyeblink. Trials were also removed when the data was not transmitted properly, caused by connectivity issues. If more than fifteen percent of the trials had to be removed, participants would have been excluded - however, this did not occur. Because of the few channels of the system and, therefore, the few components of the ICA, only components with EOG artifacts were removed. Then the segments were subjected to a time-frequency analysis with wavelet convolution. This analysis calculates the power changes over time at the highest resolution possible (1/Frequency) in the different frequency bands.

Statistical analysis

The statistical analysis is identical to Experiment 1. For the PMV task, trials were separated based on the median RT. Following, we performed a cluster-based permutation test to identify any particular channel or cluster of channels with a significant difference in alpha power, which is corrected for the multiple-comparison problem using a Monte-Carlo randomization procedure (Maris & Oostenveld, 2007). LMERs are calculated to establish the relation between EEG metrics of attention and SPAM RTs (Bates et al., 2014).

Results

Behavioral data

In Experiment 1, effects in the expected direction between attentiveness and SA were not found. This was attributed to the participants' inexperience and subsequent lack of appreciation of the importance of pieces of information. If the student subjects were more experienced, it is possible that they could distinguish the relevant and irrelevant information, and, therefore, are capable of gaining SA. If this is indeed the case, this should be represented in the behavioral data. More specifically, it is expected that the questions are answered correctly more often and that RTs are generally lower.

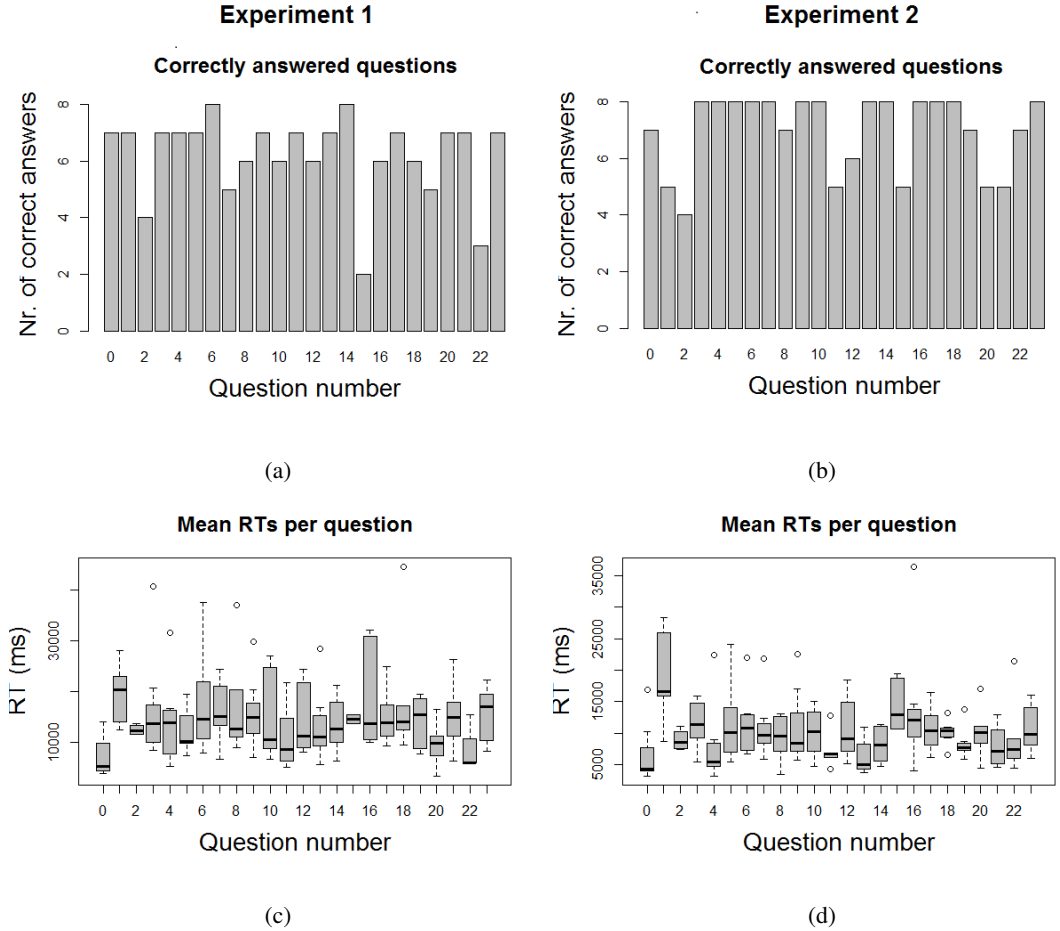


Figure 9: **a./b.** Number of times each SPAM query was answered correctly by the eight participants in Experiment 1 and Experiment 2, respectively. **c./d.** RT distribution per query in milliseconds of the eight participants in Experiment 1 and Experiment 2 respectively. Queries 0 - 11 correspond to the first scenario, and queries 12-23 correspond to the second scenario.

Figure 9 shows the amount of correct answers for each question and the RT distribution for Exper-

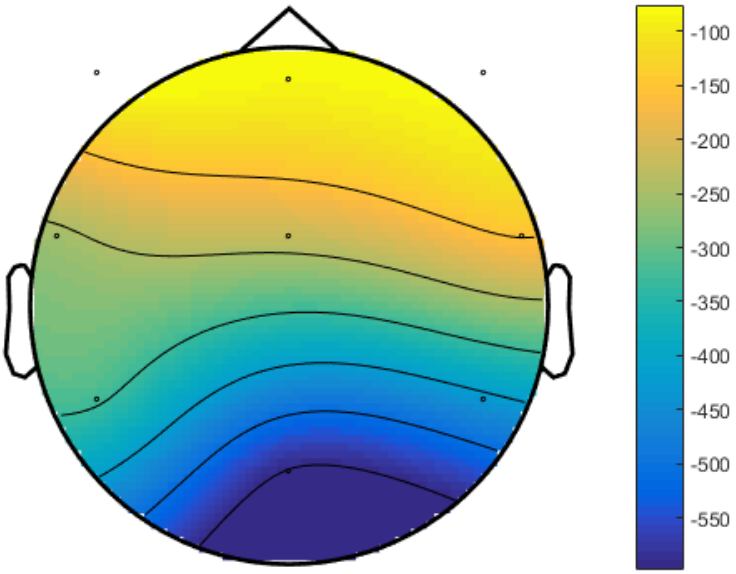


Figure 10: Grand-average scalp topography of the power differences in the alpha-frequency band (8 – 14 Hz) found in Experiment 2. Green and blue areas indicate lower alpha power in the high RT trials compared to low RT trials.

iment 1 (Figure 9a and Figure 9c) and Experiment 2 (Figure 9b and Figure 9d). To see if there was a learning effect, we performed a paired t-test comparing the amount of correct answers in each experiment, for each participant. The analyses were performed with question 2 and 15 left out, due to the few correct responses during Experiment 1. This showed that the participants answered the questions correctly significantly more often during Experiment 2 compared to Experiment 1 (t(7) = 2.707, p = 0.015). Another paired t-test, comparing the mean RTs on each question, confirmed that the questions were also answered significantly faster in Experiment 2, compared to Experiment 1 (t(21) = -7.840, p <0.001), suggesting that the participants were indeed more experienced and had a higher quality SA.

Attention and EEG

In Experiment 1, we found that with decreasing alpha power, RTs on the PMV task increased. This was attributed to attentional tunneling as an effect of high workload, which might have been the result of inexperience. Now that the behavioral data suggests that the participants were more experienced, we would expect that the workload would moderate, such that the effect of attentional tunneling, represented in the relation between alpha power and RT on the PMV task, would diminish or invert. As you can see in Figure 10 , this does not appear to be the case. The entire scalp shows a negative difference in alpha power, between high RT trials and low RT trials. No significant clusters were found (p = 1.000), but this may be attributed to having only eight participants.

SA and EEG

Because of the low number of participants that returned, it was not likely that significant effect would be found between EEG metrics of attention and SPAM RT. Instead, we focused on finding trends in the LMER and how they related to the results in Experiment 1. For a more formal comparison of the two EEG systems, please see Appendix A.

Based on the results of Experiment 1, we decided to limit this analysis to the following four models:

$$RT \sim (1|subject) + (1|question) \qquad \text{(Model 1B)}$$

$$RT \sim \alpha + (1|subject) + (1|question) \quad \text{(Model 2B)}$$

$$RT \sim \alpha Z + (1|subject) + (1|question) \quad \text{(Model 3B)}$$

$$RT \sim TEI + (1|subject) + (1|question) \quad \text{(Model 4B)}$$

, in which α is the global alpha power, αZ is the z-transformed global alpha power, and TEI is the task-engagement index, calculated over channels Cz, POz, P3 and P4. In Experiment 1, TEI was calculated using Pz instead of POz (See also Pope et al., 1995). However, because the B-Alert X10 has no Pz channel, POz was chosen as closest approximation. The models are defined exactly the same as in Experiment 1, but because the models were constructed with a different data set, and to prevent confusion, we added the affix "B" to the model names.

The effect of the EEG metric failed to reach significance for alpha power (β = -0.109, SD = 0.199, t = -0.547), z-transformed alpha power (β = 422.9, SD = 370.3, t = 1.142), and TEI (β = 7087, SD = 4842, t = 1.464), suggesting that there is no clear relation between attention and SA. Moreover, Table 5 shows that neither Model 2B, nor Model 3B or Model 4B is a significant improvement over a random model.

Table 5: Linear Mixed Effect Model comparison results. Model 2B, 3B and 4B are compared to the "random" Model 1B.

	DF	AIC	BIC	log-lik	χ^2	Df	p
Model 1B	4	3135.6	3147.8	-1563.8			
Model 2B: α-global	5	3135.6	3145.8	-1563.8	0.260	1	**0.610**
Model 3B: αZ-global	5	3136.3	3151.6	-1563.1	1.271	1	**0.260**
Model 4B: TEI	5	3135.5	3150.8	-1562.8	2.056	1	**0.152**

SA and longer-term EEG

The following models were defined with the blocked EEG data. These models are the same as in Experiment 1, but again received the "B" affix to prevent confusion. For one participant, there were no PMV-probes that could be blocked for the first two SPAM-queries; these blocked data points were subsequently removed. Model 1B for the blocked data will thus be slightly different from Model 1B from the original data.

$$RT \sim \alpha_{block} + (1|subject) + (1|question) \quad \text{(Model 5B)}$$

$$RT \sim \alpha Z_{block} + (1|subject) + (1|question) \quad \text{(Model 6B)}$$

$$RT \sim TEI_{block} + (1|subject) + (1|question) \quad \text{(Model 7B)}$$

Table 6: Linear Mixed Effect Model comparison results. Model 4, 5 and 6 are compared to the random Model 1.

	DF	AIC	BIC	log-lik	χ^2	Df	p
Model 1B	4	3085.0	3097.2	-1538.5			
Model 5B: blocked α-global	5	3086.9	3102.2	-1538.5	0.069	1	**0.793**
Model 6B: blocked αZ-global	5	3086.9	3102.2	-1538.5	0.060	1	**0.807**
Model 7B: blocked TEI	5	3086.8	3102.0	-1538.4	0.204	1	**0.652**

Model 5B and Model 6B were not able to find a significant effect for alpha power (β = -0.095, SD = 0.458, t = -0.209) and z-transformed alpha power (β = 91.46, SD = 367.32, t = 0.249) respectively.

Although Model 7B is the first model with an effect of TEI in the correct direction, that is, with increasing engagement RT decreases, the effect was not found to be significant (β = -4586, SD = 10147, t = -0.452). None of these models were a significant improvement over the "random" Model 1B (Table 6).

Discussion

The fact that only few students participated in Experiment 2 makes interpretation of the results difficult. Comparison of alpha power between low and high RT PMV probes showed no significant clusters, but a trend was found that showed that alpha power was lower in the high RT trials. This suggests that the attention tunneling effect, as described in Experiment 1, persisted. It appears that the task demands were similar in both experiments and that experience likely did not play a large role.

Comparison of the LMERs of both experiments is difficult, given that each model had a non-significant effect for EEG metrics of attention. Although many of the models are numerically similar, no conclusions can be drawn about the similarity of data. A more formal comparison of the two systems' data was performed, by comparing event-related potentials found after PMV probes, which is discussed in Appendix A.

Experiment 3

Method

Participant

Four male TTCs (age = 44.25 $\pm$ 5.11 years; experience = 11.75 $\pm$ 4.82 years) participated in this study. One TTC was familiar with the Nijmegen workplace - he received training in this area more than five years ago - but none of the participants knew the scenarios that was be played. This way it was ensured that SA had to be gained throughout the task, and that the SPAM queries could not be answered solely based on memory.

Task and Experimental Setup

The TTCs performed the simulated task with the same scenarios that were used in the student experiments, with the same simulator. Whereas the students only used a subset of the controls to make the task easier to understand, the TTCs were allowed to use the full functionality of the simulator, such that the simulator would approximate reality as much as possible. Not all functionalities that are used in the real world are implemented in the simulator, however.

New SPAM queries were constructed with a former TTC (Boris de Groot, 18 years experience), to have the SPAM queries match the decision making process of the TTCs better. Along with the TTC, a more complete minute-by-minute description of the scenarios was constructed. This description included when particular events occurred and how a TTC would subsequently act. For a full overview of the SPAM queries, please see Appendix C.

Brief verbal instructions were given about the experiment, EEG measurements and the possibilities and limitations of the simulator. After verbal consent was given, the practice scenario was run for the duration of the subjects' liking, such that they could become familiar with the SPAM queries, PMV task and the chat. During the practice scenario, the subject was being equipped to the EEG sensor and impedance values were checked. The participants were allowed to stop with the experiment without having to state any reason. The order of the scenarios were counterbalanced to prevent any order effects. During debriefing, we asked the subject about the realism of the simulator and about their experience with EEG measurements.

Measurements

Due to some technical issues, the behavioral data were not stored for one participant, making it impossible to establish the relation between EEG data and RTs on either the SPAM queries or PMV probes. Due to a lack of resources, we were unable to analyze the EEG data of the remaining participants. Only qualitative results from the debriefing were thus gathered. The following questions were asked, not including questions asking for further clarification:

- What are your thoughts about the scenarios?
  - Did you think the scenario was realistic
  - Did you think the scenarios went well?
- What are your thoughts about the EEG system?
  - Did you think it was comfortable?
  - Was the system intrusive?
  - Would you participate in these kinds of studies more often?
  - Would your coworkers want to participate in these kind of studies?
- On a scale from 1 to 10, to what degree were the functionalities of the simulator sufficient for completing the task successfully?
- On a scale from 1 to 10, how relevant were the SPAM queries?

The conversation was not recorded, but briefly noted and paraphrased instead. No transcript exists thus.

Equipment

The simulator was run on a HP desktop computer, connected to a Philips LA2306x and a Philips LA2205wg monitor. The participant's brain activity was recorded with the ABM B-Alert X-10.

Results

The simulator

The TTCs' experience with the experiment in general was quite positive. The scenarios were considered to be realistic, that is, the scenarios were close to what could happen in a real situation and the trains behaved accordingly. However, as mentioned earlier, not all functions that a TTC would use in real life were implemented in this version of the simulator. Each TTC, therefore, indicated a serious impact on the realism of the simulator. Another issue that came forward was the fact that each TTC, because they knew it was a simulation, behaved differently then in a normal situation. One participant indicated that he made less predictions about the future state and, instead, took a more reactive approach. Moreover, it was noted that he did not feel actual stress, because he knew it was "just a game". Others indicated that they knew something was going to happen and were, consequently, more attentive than normal.

The lack of functionalities especially came forward during the first scenario, which required quite some manual control. This did affect the realism of the scenario. Moreover, some of the functionalities did not always work as expected due to some minor bugs, which gave one TTC the idea of having no control of the situation. The second scenario, which required less manual control, was less affected by the lack of functionalities, although one TTC did mention that the protocol for a level-crossing failure is normally more extensive then in the simulator. Overall, the TTCs rated the simulator with a 6.75 ($\pm$ 0.43).

The SPAM queries

The SPAM queries were also received positively, and received a 7.75 ($\pm$ 0.83) out of 10 on relevance. This suggests that the new list of SPAM queries were constructed well, despite not using a goal-directed task analysis (Endsley, 2013). Two TTCs also noted that in some cases the query covered a topic that was just attended to seconds ago and could be instantly answered, and in other cases required more deliberate search. The other TTCs, however, saw the SPAM as an additional test and saw no resemblance between the questions and their thought processes.

The TTCs did mention that they were primed to be more attentive because of the both the PMV task and the SPAM queries. To paraphrase one of them: "Normally, in a simulator you know that something will happen, and you simply wait for it to happen. Only then I start actively searching for information, whereas I only scan the screen for the remainder of the time. Now, I was constantly attentive, in the expectation of an alarm (i.e., a PMV probe) or a query." This corroborates our idea that the SPAM is still rather intrusive, and a new measure of SA is necessary.

The EEG system

The TTCs found the experience with the EEG system interesting. None of the participants minded the time to setup the system, which roughly took fifteen minutes to complete. This may in part be explained by the fact that the participants were able to play the practice scenario. The TTCs also did not mind that they had to wash their hair after the experiment.

Three out of the four TTCs also indicated that they did not find the system bothersome during the execution of the task. Although they noticed the system initially, after a while they almost forgot that they were connected to the B-Alert. They also said they could easily wear the system for a couple of hours during work. The fourth TTC did find the system annoying, but mentioned that he was overly sensitive when objects are placed on or around his head. He would rather not wear the system longer than an hour.

The participants all indicated that they would like to participate more often in these types of research, although one TTC explicitly mentioned not on a daily basis. There were doubts whether other TTCs will be as positive about EEG research. Each TTC argued that about half of their colleagues would be willing to participate and the other half would not. One TTC predicts that some TTCs will especially worry

about privacy. Another TTC thinks the younger generation is more willing to participate, compared to the older generation, because the younger generation is more used to new technologies.

Discussion

The goal of this pilot study was to assess how TTCs experience doing task simulations and EEG research, and to gain more insights about SA research with TTCs. Because of some issues with data collection, and the small number of participants, we decided to only focus on qualitative data from a brief informal interview after the experiment. This showed that the TTCs were in general positive about the simulator, EEG system and the SA queries.

Given the response of the TTC, it is necessary to have a properly functioning simulator. As long as the basic functionalities are present and one could perform his/her job as he/she would in real life, it is not necessary to implement all functionalities. In other words, the simulator must match the demands of the scenario. It is advisable to use TTCs in the simulator who are familiar with the area. Being familiar with the area appears to influence how one performs the task. This, however, does not change the fact that participants know they are doing a simulation, an issue inherent to using the simulator, and subsequently behave differently. This issue can only be solved by measuring on the workplace.

The SPAM also caused the TTC to pay more attention on the task, compared to in a normal situation where TTCs would only scan the screens more superficially. This must be considered when you generalize the results of studies using the SPAM. Another issue is that the decision making processes appears to vary between TTCs and, therefore, it is difficult to construct SPAM queries that assess SA for each individual. These issues highlight the fact that the SPAM is not an optimal tool for assessing SA.

The attitude towards the B-Alert was surprisingly positive; the majority of TTCs did not mind wearing the system and would not mind participating more often. However, it was argued that not all TTCs would be that positive. Introducing the TTCs to similar research with biosensors, such as eye tracker research (e.g. Moore & Gugerty, 2010; van de Merwe, van Dijk, & Zon, 2012) or heart rate monitoring for workload (e.g. Brookhuis & de Waard, 2010), could help them to get used to EEG research.

General discussion

Being aware of the situation is incredibly important to ensure high operator performance and system safety, not only for TTCs, but in any work domain. Measuring one's SA, however, has thus far been a cumbersome process. In this exploratory research we attempted to find electrophysiological measures that could predict an operator's SA in real time, without interruptions. More specifically, we attempted to find a relation between EEG metrics of attention and SA, as measured by RT on SPAM queries. The results, however, were not able to show a consistent relationship.

In many studies it has been shown that an increase in RTs to a PMV task is associated with decrease in attentiveness (e.g. Van Dongen, Maislin, Mullington, & Dinges, 2003). Nevertheless, the results in this paper did not confirm that relationship. Instead, we found an effect in the opposite direction. This may be explained by the fact that the subjects were focused more on the TTC task and considered the PMV task as distracting; under higher levels of workload, people are likely to miss peripheral stimuli as an effect of attentional narrowing (Sheridan, 1981; Lavie, Beck, & Konstantinou, 2014). Responding late to a PMV probe may then not only be caused by inattentiveness, but also by high attentiveness to the TTC task. Despite the clear learning effect that was found in the behavioral data in Experiment 2, which was though to be accompanied by a decrease in workload (Patten et al., 2006), the proposed "attentional-tunneling effect" persisted (although this effect was not significant). It is likely that the increased experience was not accompanied by a decrease in workload, but in an increase in performance instead. Research indeed found that there exists a trade-off between workload and task performance (Granholm et al., 1996).

If the participants were indeed focused on the TTC task, we would still expect to see the positive relation between alpha power and RT on the SPAM queries. The behavioral results on the SPAM indicated that some question were harder to answer than others, given varying distributions in RTs. Therefore, in order to find a reliable trend, question difficulty must always be accounted for. Moreover, the quality of SA may differ between subjects. When the goal is to focus on within-subject differences instead of between-subject differences, these differences must also be accounted for by use of paired comparisons or, in the case of this research, by the addition of a random effect in the LMER.

Despite these precautions that were taken, we did not find any statistical significant relation between attention and SA in either Experiment 1, or Experiment 2. The data did appear to show a trend in the expected direction between z-transformed alpha power and RT on the SPAM queries, but this could not be statistically confirmed. This trend is surprising given the fact that fewer data were available for Experiment 2; only eight participant finished the task successfully in Experiment 2, compared to 21 participants in Experiment 1. Moreover, although the data quality of the B-Alert X10 is considered to be very good for a wearable EEG system, it is not perfectly equivalent to medical grade EEG systems (See Appendix A; see also Ries et al., 2014)). It is thus not unlikely that this near significant trend is also the result of random variations in the data.

In the following section, some factors are discussed that may have influenced the outcome of the experiments. A central theme is the subjects' experience in the task, which is argued to be vital for gaining SA (Berka et al., 2006, p. 4).

SPAM, gaining SA, and the relation with attention

First of all, we only had 21 participants, and only few SPAM queries could be asked in the scenarios (Pierce, 2012). This made it very difficult to find reliable trends in the data. Increasing the scenario length and the number of participants would provide more reliable results. However, increasing the amount of data will not be the only solution.

As discussed by Endsley (2013), SPAM queries need to be constructed based on a goal-directed task analysis (GDTA). A GDTA focuses on the goals and decisions of an operator and is technology free, meaning that it is independent of the tools that are being used. Such a task analysis was not available, however, and due to a lack of time it was not possible to construct one. Instead, the scenarios were investigated on a minute-by-minute basis, after which the amount of workload throughout the task was estimated. Based on these estimates and the specific events in the task, questions were formulated on specific moments in time (See Appendix B). The type of questions were discussed with a former TTC (Boris de Groot, 18 years of experience), but the moments when the questions were presented were not discussed for the student SPAM queries (Experiment 1 and 2). It was argued that the decision making

and information seeking process of inexperienced TTCs and, by extension, of students would vary to a large extent (Boris de Groot, personal communication, April 4, 2016). Therefore, it is impossible to construct the SPAM queries to suit all participants, and it was deemed justified to not verify the SPAM queries any further.

The fact that information seeking processes of the student participants is varying, exposed another issue. In the data-analysis we took three-second pre-stimulus intervals to infer attentiveness with EEG and related that to RT on the SPAM query. However, it is not sure whether the participants had attended the information relevant for that SPAM query within those three seconds. It is possible that the information was attended in the minutes beforehand or that the participant had not even attended the information at all. This means that the brain activity measured in the three second period, does not at all correspond to the awareness of the content of that particular query. This issue was tackled by looking at the blocked data, but the blocked data were only based on few thee-second snapshots over a period of two to three minutes and may, therefore, still not reflect awareness of the information asked about. In the future, either (1) more careful time frames must be selected to match the SPAM queries (e.g. by using eye tracking to see when the information was attended), (2) a different method than the SPAM must be used, that is better suited to be matched with attentional metrics of EEG, or (3) the study must be performed with experienced TTCs for which appropriate SPAM queries are constructed.

The latter point was done in a pilot study described in Experiment 3. However, only qualitative data could be assessed and analyzed. Therefore, we could not compare how well the TTCs performed on the SPAM. The TTCs did report that the they found the queries relevant to the task, suggesting that the newly constructed SPAM queries matched the decision making process of a TTC better. Still, only two TCCs explicitly mentioned that some SPAM queries covered information that they were attending to just seconds before. The other TTCs did not have this experience. It seems thus that, even among professionals, it is difficult to plan the SPAM queries to match each TTC's decision making process. An even better understanding of the TTCs' decision making process and the process of gaining SA, by use of a task analysis for instance, may thus be necessary to construct better queries. The difficulty of constructing the queries clearly is a major shortcoming of using the SPAM.

PMV Task

The PMV task was rather intrusive and may have fundamentally changed how the primary tasked was performed. The PMV task was not well integrated in the TTC task, and required additional monitoring, which may have affected general attentiveness throughout the task (Gould, Brumby, & Cox, 2013; Monk, Trafton, & Boehm-Davis, 2008). TTCs indeed noted that they were more attentive during the experiment compared to in real life, because of the inclusion of the PMV task (and SPAM). In this study, however, the inclusion of the PMV task did provide interesting information about how attention was allocated by the student participants, arguably caused by high workload as an effect of inexperience with the TTC task. Using the PMV task may thus be an informative addition, but, if ecological validity is important, it is advisable to refrain from including any additional tasks.

EEG metrics of attention

Beside a trend in z-transformed alpha power in the LMER explaining SPAM RT, the data also suggested there was a trend for the TEI. Interestingly, however, the trend was in the direction opposite of our hypothesis. That is, with increasing engagement, SA appeared to become worse. This is counter-intuitive and there is no apparent explanation for it. This effect cannot be attributed to the attentional tunneling effect found in the PMV task, because, according to that explanation, the focus was on the TTC task and should be accompanied with a higher quality SA. What may have played a role is the amount of high-frequency muscular noise in the data, as a result of the head movements during monitoring, which may have interfered with beta-power. Moreover, the TEI is a 20 year old tool. Over the years, new, more sophisticated measures of attention and engagement have been developed that may be more sensitive (e.g. (Berka et al., 2007; Mantini et al., 2007; Paiva, 2014).

More sensitive metrics will not solve the biggest shortcoming of EEG: diagnosticity. With EEG it is possible to measure to what extent one is attentive, but it is not possible to see where attention is focused. In dynamic workplaces with many sources of information, attentional metrics of EEG alone will not provide enough insights to predict one's SA accurately. Instead, a multi-measure approach is

necessary to capture complex constructs like SA (Salmon et al., 2006). Eye-tracking, a camera based tool that is able to follow the movements of the eye, for example, could be used to see where attention is focused. Moreover, different metrics have been developed to infer workload and SA in real time (van de Merwe et al., 2012; Moore & Gugerty, 2010; Yu et al., 2014). A first attempt to combine the two for an accurate measure of attentional lapses is already made (Paiva, 2014). Future research should shine some light on how EEG and eye-tracker may complement each other for measuring SA.

Technical issues

Two technical issues were encountered during the experiment that also have influenced the results. First of all, the simulator was implemented in Java, which made it necessary to send event-markers to the parallel port to the EEG systems, via Java. Many different applications were tried and tested, but none of them appeared to work. As a workaround, we had Java call a Python script via the command line, in order to send the event marker. This extra step may have introduced not only additional latency between presentation of the stimulus and recording of the event marker, but will likely also have caused more variability due to threading of Windows processes. Because the EEG segments were averaged over three seconds, the consequences will have been marginal. However, for ERP analyses (See Appendix A), this may seriously affect the results.

Secondly, we encountered some issues with the monitors. From time to time, the screens froze, while the simulator continued. The participants were asked to notify the experimenter as soon as this happened. In some cases, however, it took the participant over one minute to notice the freeze, making the participant miss multiple PMV probes or SPAM queries. Whenever the freeze was noticed, the experimenter paused the experiment and fixed the monitors by unplugging them and plugging them back in. As a result of this, some windows moved to different locations which could not be moved back (e.g. the alarm stimuli of the PMV task and the chat window). The mix-up of the windows forced the participant to reorient and to get used to the new orientation, which most likely is accompanied different brain activity.

Future directions

As is apparent from this discussion, many issues need to be resolved before we can better understand the dynamics between attention and SA, and subsequently develop an EEG metric of SA. Other steps can be taken in parallel, however. This study focused solely on attention and how it may affect SA, but SA is not only modulated by attention; an understanding of the task and the ability to make predictions are also vital parts of SA (Smith & Hancock, 1995), and in fact modulate how and where attention is focused (Catherwood et al., 2014). Understanding how knowledge and predictions can be assessed with EEG and understanding their dynamics with attention, may bring forth novel ways to asses SA with EEG.

Conclusion

Having high quality SA is vital for high operator performance and safety. To be able to understand the dynamics of SA, it is necessary that it is measured reliably and continuously. In this paper, we tried to establish the relationship between SA and EEG metrics of attention, as a set-up for a new online objective measure of SA. All in all, no consistent evidence was found for the relation between attention and SA in this study, but many valuable insights were gained. Most importantly, the role of experience in the task must not be underestimated, given its importance not only for gaining SA by the subjects, but also for assessing SA by the researchers.

Acknowledgments

I would like to express my deep gratitude to my supervisors, Marieke van Vugt and Julia Lo, for their continuous support, great ideas and feedback on the project. Furthermore, I would like to thank my colleagues and co-interns at ProRail Innovation and Development for their support and for testing the experiment. Arie van der Berg deserves a special mention, because without his help we would not have

been able to perform the experiment. Also, special thanks should be given to Ursula Beer, researcher at the Delft University of Technology, for her contribution to the task instructions. Many thanks also go out to the participants, both students and train traffic controllers, for without them there would be few results to report on. Finally, I want to show my appreciation to Melcher Zeilstra, partner at Intergo, human factors & ergonomics consultancy, for his valuable input in the project. More importantly, without his enthusiasm, this project would not have existed.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal Of Statistical Software*, *67*(1), 1–48.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, *7*(6), 1129–1159.

Berka, C., Levendowski, D. J., Davis, G., Whitmoyer, M., Hale, K., & Fuchs, K. (2006). Objective measures of situational awareness using neurophysiology technology. *Augmented Cognition: Past, Present and Future*, 145–154.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., ... Craven, P. L. (2007). Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, *78*(Supplement 1), B231–B244.

Brookhuis, K. A., & de Waard, D. (2010). Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident Analysis & Prevention*, *42*(3), 898–903.

Catherwood, D., Edgar, G. K., Nikolla, D., Alford, C., Brookes, D., Baker, S., & White, S. (2014). Mapping brain activity during loss of situation awareness an eeg investigation of a basis for top-down influence on perception. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*(8), 1428–1452.

Croft, D. G., Banbury, S., Butler, L. T., & Berry, D. C. (2004). The role of awareness in situation awareness. *A cognitive approach to situation awareness: Theory and application*, 82–103.

Crundall, D., Crundall, E., Clarke, D., & Shahar, A. (2012). Why do car drivers fail to give way to motorcycles at t-junctions? *Accident Analysis & Prevention*, *44*(1), 88–96.

Donald, F. M., & Donald, C. H. (2015). Task disengagement and implications for vigilance performance in cctv surveillance. *Cognition, Technology & Work*, *17*(1), 121–130.

Durso, F. T., Dattel, A. R., Banbury, S., & Tremblay, S. (2004). Spam: The real-time assessment of sa. *A cognitive approach to situation awareness: Theory and application*, *1*, 137–154.

Durso, F. T., & Sethumadhavan, A. (2008). Situation awareness: Understanding dynamic environments. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 442–448.

Endsley, M. R. (1988). Situation awareness global assessment technique (sagat). In *Aerospace and electronics conference, 1988. naecon 1988., proceedings of the ieee 1988 national* (pp. 789–795).

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 32–64.

Endsley, M. R. (2013). Situation awareness-oriented design. *The Oxford Handbook of Cognitive Engineering*, 272.

Endsley, M. R., & Garland, D. (2000b). Theoretical underpinnings of situation awareness: A critical review. *Situation awareness analysis and measurement*, 3–32.

Endsley, M. R., & Garland, D. J. (2000a). Pilot situation awareness training in general aviation. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 44, pp. 357–360).

French, H. T., Clarke, E., Pomeroy, D., Seymour, M., & Clark, C. R. (2007). Psycho-physiological measures of situation awareness. *Decision making in complex environments*, *291*.

Georgiev, S., Lalova, Y., Ivanova, I., & Philipova, D. (2006). Attention scores and erp components in sensomotor task. *Homeostatis*, *44*(3), 119–125.

Golightly, D., Wilson, J. R., Lowe, E., & Sharples, S. (2010). The role of situation awareness for understanding signalling and control in rail operations. *Theoretical Issues in Ergonomics Science*, *11*(1-2), 84–98.

Gould, S. J., Brumby, D. P., & Cox, A. L. (2013). What does it mean for an interruption to be relevant? an investigation of relevance as a memory effect. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 57, pp. 149–153).

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology*, *33*(4), 457–461.

Knyazev, G. G., Slobodskoj-Plusnin, J. Y., Bocharov, A. V., & Pylkova, L. V. (2011). The default mode network and eeg alpha oscillations: an independent component analysis. *Brain research*, *1402*, 67–79.

Laufs, H., Kleinschmidt, A., Beyerle, A., Eger, E., Salek-Haddadi, A., Preibisch, C., & Krakow, K. (2003). Eeg-correlated fmri of human alpha activity. *Neuroimage*, *19*(4), 1463–1476.

Lavie, N., Beck, D. M., & Konstantinou, N. (2014). Blinded by the load: attention, awareness and the role of perceptual load. *Phil. Trans. R. Soc. B*, *369*(1641), 20130205.

Liu, S., Wanyan, X., & Zhuang, D. (2014). Modeling the situation awareness by the analysis of cognitive process. *Bio-medical materials and engineering*, *24*(6), 2311–2318.

Lo, J. C., Sehic, E., & Meijer, S. A. (2014). Explicit or implicit situation awareness? situation awareness measurements of train traffic controllers in a monitoring mode. In *International conference on engineering psychology and cognitive ergonomics* (pp. 511–521).

Mantini, D., Perrucci, M. G., Del Gratta, C., Romani, G. L., & Corbetta, M. (2007). Electrophysiological signatures of resting state networks in the human brain. *Proceedings of the National Academy of Sciences*, *104*(32), 13170–13175.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, *164*(1), 177–190.

Monk, C. A., Trafton, J. G., & Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied*, *14*(4), 299.

Moore, K., & Gugerty, L. (2010). Development of a novel measure of situation awareness: The case for eye movement analysis. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 54, pp. 1650–1654).

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology.* WH Freeman/Times Books/Henry Holt & Co.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of Neuroscience*, *35*(21), 8145–8157.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational intelligence and neuroscience*, *2011*.

Paiva, J. I. S. (2014). *Predicting lapses in attention: a study of brain oscillations, neural synchrony and eye measures*. (Unpublished master's thesis). Universidade de Coimbra, Coimbra, Portugal.

Patten, C. J., Kircher, A., Östlund, J., Nilsson, L., & Svenson, O. (2006). Driver experience and cognitive workload in different traffic environments. *Accident Analysis & Prevention*, *38*(5), 887–894.

Pierce, R. S. (2012). The effect of spam administration during a dynamic simulation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(5), 838–848.

Pope, A. T., Bogart, E. H., & Bartolome, D. S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biological psychology*, *40*(1), 187–195.

Ratwani, R. M., McCurry, J. M., & Trafton, J. G. (2010). Single operator, multiple robots: an eye movement based theoretic model of operator situation awareness. In *Proceedings of the 5th acm/ieee international conference on human-robot interaction* (pp. 235–242).

Ries, A. J., Touryan, J., Vettel, J., McDowell, K., & Hairston, W. D. (2014). A comparison of electroencephalography signals acquired from conventional and mobile systems. *Journal of Neuroscience and Neuroengineering*, *3*(1), 10–20.

Ruby, F. J., Smallwood, J., Sackur, J., & Singer, T. (2013). Is self-generated thought a means of social problem solving? *Frontiers in psychology*, *4*.

Salmon, P., Stanton, N., Walker, G., & Green, D. (2006). Situation awareness measurement: A review of applicability for c4i environments. *Applied ergonomics*, *37*(2), 225–238.

Sauseng, P., Klimesch, W., Stadler, W., Schabus, M., Doppelmayr, M., Hanslmayr, S., ... Birbaumer, N. (2005). A shift of visual spatial attention is selectively associated with human eeg alpha activity. *European Journal of Neuroscience*, *22*(11), 2917–2926.

Scheeringa, R., Bastiaansen, M. C., Petersson, K. M., Oostenveld, R., Norris, D. G., & Hagoort, P. (2008). Frontal theta eeg activity correlates negatively with the default mode network in resting state. *International Journal of Psychophysiology*, *67*(3), 242–251.

Sheridan, T. B. (1981). Understanding human error and aiding human diagnostic behaviour in nuclear power plants. In *Human detection and diagnosis of system failures* (pp. 19–35). Springer.

Smallwood, J. (2013). Distinguishing how from why the mind wanders: a process–occurrence framework for self-generated mental activity. *Psychological Bulletin*, *139*(3), 519.

Smith, K., & Hancock, P. (1995). Situation awareness is adaptive, externally directed consciousness. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 137–148.

Steenhuisen, B. (2009). *Competing public values: Coping strategies in heavily regulated utility industries*. TU Delft, Delft University of Technology.

van de Merwe, K., van Dijk, H., & Zon, R. (2012). Eye movements as an indicator of situation awareness in a flight simulator experiment. *The International Journal of Aviation Psychology*, *22*(1), 78–95.

Van Dijk, H., Schoffelen, J.-M., Oostenveld, R., & Jensen, O. (2008). Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability. *The Journal of Neuroscience*, *28*(8), 1816–1823.

Van Dongen, H., Maislin, G., Mullington, J., & Dinges, D. (2003). The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *SLEEP*, *2*, 117–126.

Vidulich, M. A., Stratton, M., Crabtree, M., & Wilson, G. (1994). Performance-based and physiological measures of situational awareness. *Aviation, Space, and Environmental Medicine*.

Yu, C.-s., Wang, E. M.-y., Li, W.-C., & Braithwaite, G. (2014). Pilots' visual scan patterns and situation awareness in flight operations. *Aviation, space, and environmental medicine*, *85*(7), 708–714.

Appendix A: Comparison of the BioSemi and B-Alert

In this study we sought to establish a relationship between EEG metrics of attention and situation awareness. The data were acquired with two different EEG systems: the medical grade BioSemi ActiveTwo 128-channel system and the wearable Advanced Brain Monitoring B-Alert X10 9-channel system. To see whether the B-Alert was fit for field studies, the intention was to compare the data and the results of the analyses. No conclusions could be drawn, however, due to a lack of significant effects between SA and attention.

Previous research has already shown the quality of the B-Alert data recordings in a controlled experiment (Ries, Touryan, Vettel, McDowell, & Hairston, 2014). In the study of Ries et al. (2014), participants performed the visual odd-ball task, while being recorded with three different EEG systems: The BioSemi ActiveTwo 64-channel system, the B-Alert X10, and the Emotiv EPOC 14-channel commercially available headset. In the task, participants had to identify one of two stimuli (one of which is presented 88% of the time) and respond by pressing the corresponding keyboard button. Ries et al. (2014) limited their analysis to the correctly responded to trials of the frequent stimulus.

The study showed that, during pre-processing, significantly more trials had to be rejected for the EPOC data compared to both the B-Alert and BioSemi, while there was no significant difference between the latter two system. Following, individual subject event-related potentials (ERPs) that were time-locked to the event-markers, as registered by the system, were calculated and compared with ERPs that were time-locked to a third-party system, which corrects for temporal jitter. The event-timing variability was quantified using the standard deviation of the response onset peaks. This showed that the BioSemi had deviations of $\pm$ 1.7 ms and the B-Alert deviations of $\pm$ 3.3 ms. Compared to the EPOC, which showed deviations of $\pm$ 32.65 ms, the performance of the B-Alert can be considered almost equal to the BioSemi. Furthermore, they found relative high correlations between the grand averaged ERPs, calculated with the jitter-corrected data, of the BioSemi and B-Alert system (ρ = 0.596, SD = 0.230), compared to more moderate correlations between the BioSemi and the EPOC (ρ = 0.480, SD = 0.230). This again confirms that the data quality of the B-Alert X10 approaches the quality of the BioSemi. Many more different analyses were performed, but these lie outside the scope of this study.

As of yet, no replication studies have been conducted, nor have the systems been tested in a dynamic environment. In this section we will, therefore, again formally validate the quality of the B-Alert X10 recordings, to ensure that similar data can be recorded in field studies.

Method

Participants

Eight participants that successfully finished both Experiment 1 and Experiment 2 (age = 21.82 $\pm$ 2.04 years; 6 female) were included for this study.

EEG preprocessing and analysis

The pre-processed data of Experiment 1 and Experiment 2 were used for the comparison. The data in Experiment 1 was sampled with the BioSemi at 512 Hz. The data in Experiment 2 was sampled at 256 Hz with the B-Alert, but was up-sampled to 1024 Hz, to match different channels in the B-Alert data that were not used for this experiment. The B-Alert data was subsequently down-sampled to 512 Hz, such that the data matches the BioSemi data.

The comparison was done on the PMV probes, to ensure that sufficient data was available for a reliable comparison. For each subject, all trials were segmented from 0.2 seconds pre-stimulus to one second post-stimulus and underwent baseline correction based on the mean activity over the 0.2 second pre-stimulus interval. Following, individual and global average ERPs were calculated for each participant. This analysis was done for each channel available in the B-Alert. In the case that there was no exact match with the BioSemi headset, due to a different layout of channels, a nearby channel was used as closest approximation. In Table 7 the channels of the B-Alert X10 and the matched BioSemi channels are shown.

Table 7: The channels of the B-Alert X10 and the matched BioSemi channels that were compared. The names within parentheses are the channel names given by BioSemi.

System	Channels								
B-Alert	Fz	Cz	POz	F3	F4	C3	C4	P3	P4
BioSemi	Fz(C21)	Cz(A1)	POz(A21)	(D4)	(C4)	C3(D19)	C4(B22)	(A8)	(B5)

Results

Behavioral comparison

How quickly the stimuli are responded to may indicate how quickly the stimuli are seen. This may in turn determine the shape of the ERPs. Therefore, we would like to see no differences in RT. A two-tailed paired t-test showed that the difference in RTs between Experiment 1 ($m_1 = 1276.0 \pm 326.53$ ms) and Experiment 2 ($m_2 = 1028.9 \pm 133.63$ ms) is near-significant ($t(7) = 2.123$, $p = 0.07$). This suggests that response times were in fact shorter in Experiment 2. This difference should be taken into account when considering the difference between the ERPs.

ERP comparison

The ERPs for both systems are plotted in Figure 11 . When we look at the frontal and central channels, the magnitude of the ERP components stands out immediately. Whereas the BioSemi's ERPs have a range of about 6 mV, the range of the B-Alert data is in the order of 2 mV. To make visual comparison easier, we normalized the grand averaged data by calculating the standardized z-scores. The results of that comparison are shown in Figure 12

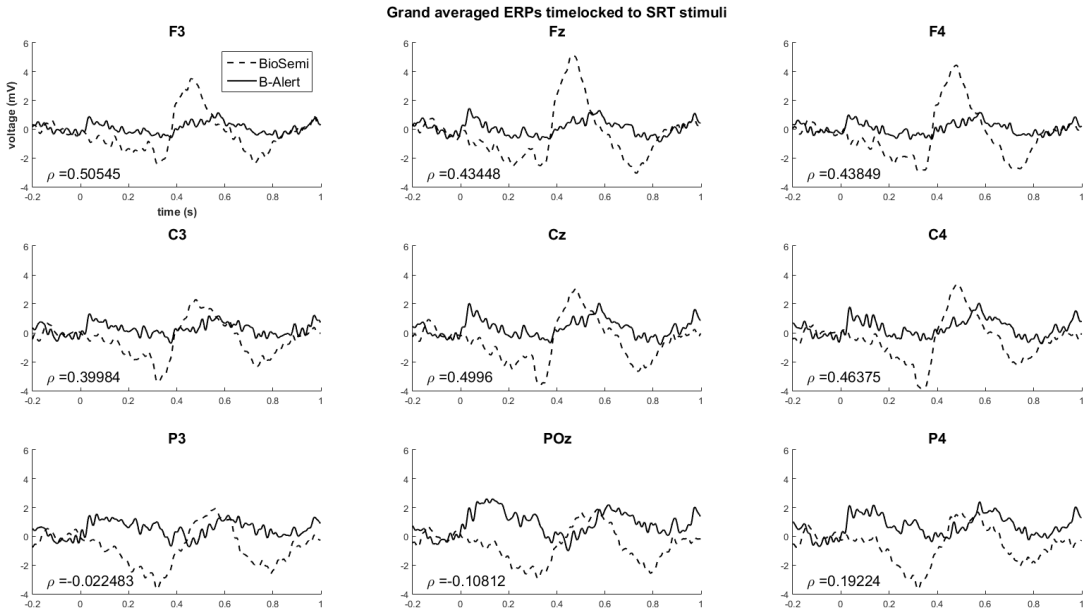


Figure 11: Event related potentials of each channel, time-locked to the presentation of the PMV probes ($t = 0$). In each subplot the correlation coefficient is shown.

If we would only look at the shape of the ERP waveforms, we see that in most cases the shape of the curve is similar, but the peaks of the data do have a mismatch in timing. More specifically, the P3, the positive deflection around 400 ms, of the B-Alert data seems to be shifted to the right. Moreover, it appears as if the peaks in the B-Alert X10 data are still smaller than the BioSemi data. A paired t-tests comparing each channel confirmed that the P3 amplitude is indeed significantly larger in Experiment 1 ($m_1 = 4.678 \pm 0.779$) compared to Experiment 2 ($m_2 = 3.416 \pm 0.666$; $t(16) = 5.687$, $p < 0.001$). Moreover, the latency of the P3 is significantly shorter for Experiment 1 ($m_1 = 0.5504 \pm 0.046$ seconds

post-stimulus) compared to Experiment 2 ($m_2 = 0.579 \pm 0.024$ seconds; $t(8) = -6.359$, $p < 0.001$). The maximum peak value and its corresponding latency were determine for each channel separately within the 0.3 to 0.7 post-stimulus time period. Finally, we also see that the B-Alert shows much more fluctuations, compared to the relatively smooth BioSemi, which suggests that the data is more noisy.

These results taken together, seem to be in accordance with the correlation coefficients, that are in the order of ρ = 0.5 in frontal and central channels. This is close to the values found in the study of (Ries et al., 2014), which can be considered reasonably well, given the dynamic nature of the task. In these posterior channels, however, the correlation coefficients much lower, centering around ρ = 0.0.

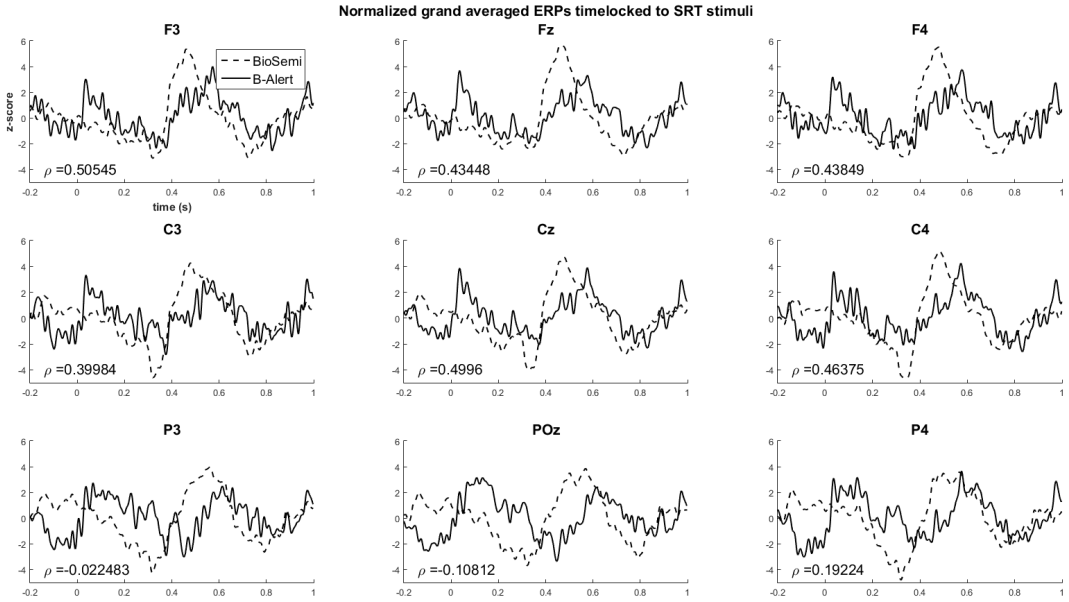


Figure 12: Normalized event related potentials of each channel, time-locked to the presentation of the PMV probes ($t = 0$). In each subplot the correlation coefficient is shown.

Discussion

With this analysis we sought to show that the B-Alert is capable of recording high quality EEG data by comparing ERPs, time-locked to the PMV probes, of the B-Alert and the BioSemi. Despite the low number of participants, promising results were found that suggest that the B-Alert indeed records high-quality EEG data. Some differences were found, however, which are discussed in the remainder of this section.

The difference in range, as seen in Figure 11 could be explained by the fact that we were more liberal with impedance with the B-Alert. Whereas impedance values of the BioSemi were maintained below 40 $k\Omega$, a value of 70 $k\Omega$ was deemed sufficient for the B-Alert. With impedance values that can almost be twice as large, it can be expected that lower power is recorded. Normalization of the data is thus justified.

Despite the normalization, we still saw that the P3 was significantly later and smaller in amplitude in the B-Alert data, compared to the BioSemi. One explanation for this is given by Georgiev, Lalova, Ivanova, & Philipova (2006), who showed that when people are more attentive, P3 latency becomes shorter and P3 amplitude becomes higher. Together with the proposed "attentional-tunneling effect" found in Experiment 1 - it was found that RT increased with increasing attentiveness - this may explain why RTs were shorter in Experiment 2 compared to Experiment 1. The behavioral results and the ERPs are thus consistent with each other.

Another explanation may be the higher variability of the timing of event-markers for the B-Alert, as shown in the study by Ries et al. (2014). This variability may move the timeline and, by extension, the P3 peak of individual ERPs both to to the left and to the right. After averaging, the peak may have smeared out, thereby having not only a lower amplitude, but also a broader P3. This effect may have

been amplified by the variability caused by the parallel-port communication used in this study. It was not possible to have Java send event markers to the B-Alert directly. Instead, it was necessary to call a Python script, which made the communication dependent on the threading of computer processes. To which extent this may have played a role cannot be tested, however, because it could not be distinguished from the inherent variability of the B-Alert or the behavioral effect caused by attentiveness.

Ries et al. (2014) used the temporally corrected ERPs to calculate the correlation coefficient between the two systems. We were not able to perform this correction, but were still able to find correlations of the same magnitude in frontal and central channels. The low correlations between the systems in the posterior regions were unexpected. This may, in part be explained by temporal differences between the ERP waveforms of the two systems. However, the effect on the correlation coefficients is not proportional in comparison to the frontal and central channels. Unless temporal difference was larger for the posterior regions than the central and frontal regions, there is no apparent explanation for the low correlations. A more careful inspection of the data is necessary to understand why these differences exist.

Finally, we also found that the B-Alert data shows more fluctuations, which may be the result of more noisy data acquisition. Moreover, the artifact correction procedure that was limited for the B-Alert data. There are few channels and, in turn, few components that could be constructed by the independent component analysis (Bell & Sejnowski, 1995). It was, therefore, decided to only remove components containing EOG artifacts, so as not to remove too much data. Muscular artifacts, for instance, are thus still present in the data.

Conclusion

In sum, the results of the two systems are in correspondence to a large extent. Some differences in the ERP waveforms came to light, such as the timing mismatch and some noise in the B-Alert data, but these could, in part, be explained by a difference in attentiveness. Any remaining differences, although not quantified, appear acceptable given the ease of use of the B-Alert system. It is thus safe to conclude that the B-Alert X10 is a good option for EEG research in which traditional systems, such as the BioSemi, are not available or not preferable for the situation at hand.

Appendix B: Student SPAM Queries

Table 8: Dutch SPAM queries, constructed for students, with possible answers specific for the first scenario. The time column shows the simulator time, starting from 16:50 and ending at 17:20. The correct answers from the prototypical scenario are printed in bold. Note that for different participants, the scenario may develop differently and, therefore, the correct answers may vary.

Time	Question	Answers			
16:52	Hoeveel minuten vertraging heeft trein 55555?	6 of minder	**7**	8	9 of meer
16:54	Op welk spoorstuk bevindt trein 3658 zich?	**Station Oss**	Oss 205	PC	GO
16:57	Hoeveel perrons zijn er beschikbaar in Nijmegen? (A en B gelden als aparte perrons)	1 of minder	2	**3**	4 of meer
16:59	Wie arriveert eerst in Nijmegen volgens planning?	4458	3661	**3157**	32356
17:02	Vanaf welk spoor rijdt trein 55555 Nijmegen binnen?	NW	**WN**	BC	AC
17:04	Trein 3060 vertrekt vanuit Nijmegen richting spoor ...?	NW	WN	BC	**AC**
17:06	Wat is het geplande aankomst spoor van trein 4458 in Nijmegen?	103B	**104B**	105B	106B
17:08	Wat is de geplande tijd 'Doorrij' tijd van trein 3060 in Elst?	17:10	17:13	17:16	**17:19**
17:11	Wat is het geplande 'Doorrij' spoor van trein 55555 in Nijmegen?	103	104	**105**	106
17:14	Welke actie neemt trein 7660 in Elst om 17:17?	A: Aankomst	V: Vertrek	**K: Korte stop**	D: Doorrijden
17:17	In welk PPLG bevindt trein 3661 zich momenteel?	**Nijmegen**	Elst	Wijchen	Ravenstein
17:19	Hoeveel perrons zijn er vrij op Wijchen?	0	1	**2**	3

Table 9: Dutch SPAM queries, constructed for students, with possible answers specific for the second scenario. The time column shows the simulator time, starting from 17:15 and ending at 17:45. The correct answers from the prototypical scenario are printed in bold. Note that for different participants, the scenario may develop differently and, therefore, the correct answers may vary.

Time	Question	Answers			
17:17	Wat is de geplande aankomsttijd van trein 3159 in Nijmegen?	17:29	17:30	17:31	**17:32**
17:19	Vanaf welk spoor rijdt trein 22222 Elst binnen?	AC	BC	AE	**BE**
17:21	Trein 4463 vertrekt vanuit Nijmegen naar... ?	AC	BC	**NW**	WN
17:24	Hoeveel perrons zijn er beschikbaar in Nijmegen (A en B gelden als aparte perrons)?	1 of minder	**2**	3	4 of meer
17:26	In welk PPLG bevindt trein 3658 zich?	**Elst**	Nijmegen	Ravenstein	Wijchen
17:28	Welke actie neemt trein 3660 in Wijchen om 17:35?	A: Aankomst	V: Vertrek	K: Korte stop	**D: Doorrijden**
17:30	Wat is het geplande aankomstspoor van trein 3663 in Nijmegen?	101A	103A	103B	**104A**
17:33	Op welk spoorstuk bevindt trein 7661 zich?	**Nijmegen 104A**	AC	Elst 97	Elst 94
17:36	Hoeveel perrons zijn er vrij op Wijchen?	0	1	**2**	3
17:39	Welke trein arriveert eerst in Nijmegen volgens planning?	3663	**3159**	3059	17661
17:41	Hoeveel minuten vertraging heeft trein 3663?	6 of minder	**7**	8	9 of meer
17:43	Wat is het geplande aankomstspoor van trein 32258?	101A	102A	**101B**	102B

Appendix C: TTC SPAM queries

Table 10: Dutch SPAM queries, constructed for the TTCs, with possible answers specific for the first scenario. The time column shows the simulator time, starting from 16:50 and ending at 17:20. The correct answers from the prototypical scenario are printed in bold. Note that for different participants, the scenario may develop differently and, therefore, the correct answers may vary.

Time	Question	Answers			
16:52	Hoeveel minuten vertraging heeft trein 55555?	6 of minder	**7**	8	9 of meer
16:54	Vanaf welk spoor komt trein 3157 Elst binnen rijden?	**BE**	AE	EO	AC
16:56	Wat is de geplande doorrij tijd van trein 3658 in Ravenstein?	17:01	17:02	**17:03**	17:04
16:59	Wat is het geplande doorrij spoor van trein 55555 in Nijmegen?	102A	104A	**105A**	106A
17:02	Vanaf welk spoor rijdt trein 55555 Nijmegen binnen?	NW	**WN**	BC	AC
17:04	Op welk spoorstuk bevindt trein 3060 zich momenteel?	**101A**	101B	103A	103B
17:06	Hoeveel minuten vertraging heeft trein 4458?	6 of minder	**7**	8	9 of meer
17:08	Wat is de geplande aankomsttijd van trein 3658 in Elst?	17:09	17:10	17:11	**17:12**
17:11	Welke actie neemt trein 3057 om 17:17 in Nijmegen?	**Aankomst**	Vertrek	Korte stop	Doorrijden
17:14	Wat is de geplande doorrij tijd van trein 31153 in Elst?	17:10	17:13	**17:16**	17:19
17:17	In welk PPLG bevindt trein 3661 zich momenteel?	**Nijmegen**	Elst	Wijchen	Ravenstein
17:19	Wat is het geplande aankomstspoor van trein 7660 in Nijmegen?	101A	103A	**104A**	105A

Table 11: Dutch SPAM queries, constructed for the TTCs, with possible answers specific for the second scenario. The time column shows the simulator time, starting from 17:15 and ending at 17:45. The correct answers from the prototypical scenario are printed in bold. Note that for different participants, the scenario may develop differently and, therefore, the correct answers may vary.

Time	Question	Answers			
17:17	Vanaf welk spoor rijdt trein 22222 Elst binnen?	AC	BC	AE	**BE**
17:19	Wat is de geplande aankomsttijd van trein 7660 in Nijmegen?	17:23	17:24	17:25	**17:26**
17:21	Wat is de geplande doorrij tijd van trein 3159 in Elst?	17:24	17:25	**17:26**	17:27
17:24	Trein 3162 rijdt om 17:27 van Nijmegen naar...?	NW	**AC**	162	UA
17:26	In welk PPLG bevindt trein 3159 zich?	**Elst**	Nijmegen	Ravenstein	Wijchen
17:28	Welke actie neemt trein 3663 in Elst om 17:35?	Aankomst	Vertrek	**Korte stop**	Doorrijden
17:30	Wat is het geplande aankomstspoor van trein 3159 in Nijmegen?	**101A**	101B	103A	103B
17:33	Wat is de geplande vertrektijd van trein 3062 uit Nijmegen?	**17:42**	17:43	17:44	17:45
17:35	Hoeveel minuten loop trein 3159 achter op schema, volgens de laatste meting?	0	**1**	2	3 of meer
17:39	Vanaf welk spoor vertrek 3663 uit Nijmegen om 17:47?	101A	103A	**104A**	105A
17:41	Hoeveel minuten vertraging heeft trein 3663?	**6 of minder**	7	8	9 of meer
17:43	Op welk spoorstuk bevind trein 17664 zich momenteel?	103B	**NW**	301	302