Measuring the User Experience of Adaptive Menus using EEG: A Replication Study

Anonymous Author(s)

ABSTRACT

Background: Adaptive user interfaces benefit from their ability to dynamically change their aspect and/or behaviour depending on the characteristics of the context of use, namely to improve the user experience. The user experience is an important quality factor that has been primarily evaluated with classical measures (e.g. effectiveness, efficiency, satisfaction), but not so much with physiological measures, such as emotion recognition, skin response, or brain activity. **Aim**: In a previous exploratory experiment involving users with different profiles and a wide range of ages, we analysed the user experience of twenty graphical adaptive menus via an Electroencephalogram (EEG) device. The results indicated that no statistically significant differences existed for valence (also known as attraction), memorisation, engagement and cognitive load. Therefore, we consider that it is necessary to confirm or reject these findings using a more homogeneous group of users. **Method**: We conducted an internal strict replication study at the *[eliminated for double blind review]* with an in-depth analysis. We also investigated the potential correlation between EEG signals and the participant's user experience ratings, such as preferences. **Results**: The results of this experiment confirm that there are no statistically significant differences between the EEG variables. However, there is a high correlation among the participants' user experience ratings and the EEG signals, and that a trend on performance emerges from our analysis. **Conclusions**: These findings suggest that EEG signals can effectively be used to evaluate the user experience. Regarding the menus within the scope of this study, our results suggest that graphical menus which include a temporal dimension enhance the user experience and, on the contrary, graphical menus with different formats (e.g., font types, typography) degrade it. Several insights for improving the user experience of graphical adaptive menus are outlined.

CCS CONCEPTS

• **Human-centered computing** → **Graphical user interfaces**; **Ubiquitous and mobile devices**; *User studies*; *Empirical studies in interaction design*; • **Applied computing** → **Personal computers and PC applications**.

KEYWORDS

Adaptive User Interfaces, User Experience, Electroencephalography

Submitted to ESEM '22, September 19–23, 2022, Helsinki, Finland
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Anonymous Author(s). 2022. Measuring the User Experience of Adaptive Menus using EEG: A Replication Study. In *Proceedings of Submitted to International Symposium on Empirical Software Engineering and Measurement (Submitted to ESEM '22).* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

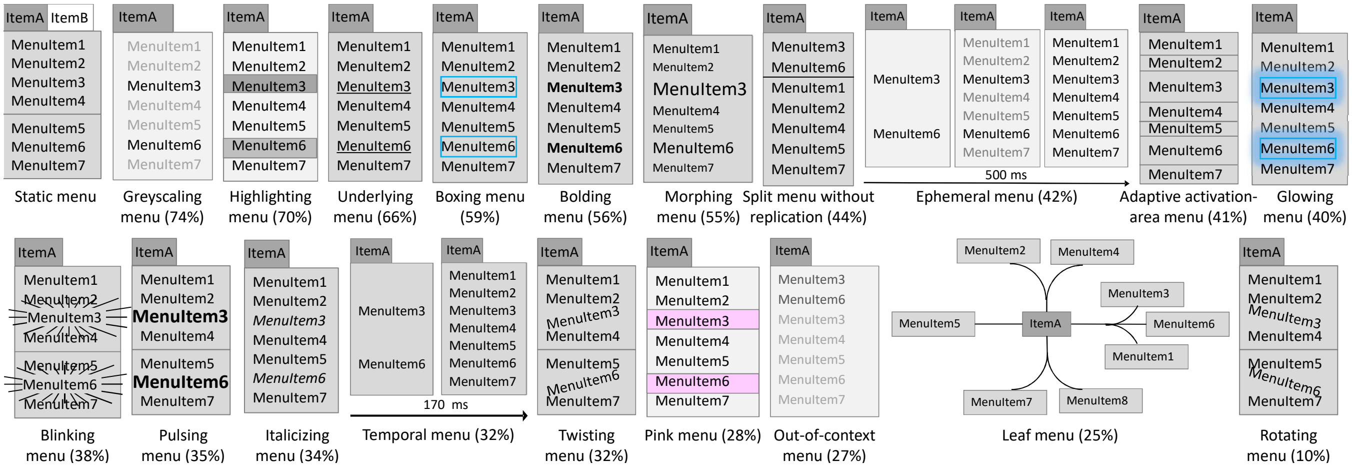
1 INTRODUCTION

Physiological measures obtained from different parts of the human body, including eyes, brain, heart, skin, among others, can be obtained using biometric-related sensors. These measurements can be analysed to represent different human affective states (e.g., pleasure, stressed, relaxed) when performing any task. Brain nervous system activity measurements have, over the past years, received increasing attention in Software Engineering (SE). Understanding how humans react during SE tasks and being able to predict what will make end-users frustrated, engaged, or pleased, will eventually increase software quality and/or user experience. A recent systematic literature review on the use of physiological measures in SE by Weber et al. [53] revealed that the focus to date has been on code comprehension, while other applications of these measurements for code inspection, programming, and bug fixing have been less frequent. The studies analysed mainly used methods related to Electroencephalography (EEG) and functional Magnetic Resonance Imaging (fMRI).

Several empirical studies have also been conducted to measure cognitive load in SE activities through brain and autonomic nervous system activity data, among which are [12, 13, 19, 20, 26, 33, 39, 41, 44]. Of these studies, only a few have reported experiments to capture end-user emotions and feelings through using EEG signals [26], [44]. In addition to new experiments, replications are needed to increase the body of knowledge about the usefulness of these physiological measures in supporting SE tasks.

In a previous study, we performed an exploratory experiment to assess the impact of Adaptive User Interfaces on the cognitive load and user experience, both measured using EEG signals [4]. The goal of this experiment was to assess if there are significant differences in the physiological response of a group of participants when using twenty different types of graphical adaptive menus selected from the catalogue reported in [52]. The results of the experiment showed the feasibility of measuring cognitive load and emotions using EEG, but the factor analysis showed that there were no significant differences between the graphical menus in terms of cognitive load, engagement, memorisation, and attraction, which is unprecedented.

In this paper, we present an internal strict replication of this experiment with a group of Computer Science students who were enrolled in a Master's degree program at the [eliminated for double-blind review]. The goal of this replication is to verify the findings of the baseline experiment in a different context. The remainder of this paper is structured as follows. Section 2 gives a background to



Static menu
Greyscaling menu (74%)
Highlighting menu (70%)
Underlying menu (66%)
Boxing menu (59%)
Bolding menu (56%)
Morphing menu (55%)
Split menu without replication (44%)
Ephemeral menu (42%)
Adaptive activation-area menu (41%)
Glowing menu (40%)
Blinking menu (38%)
Pulsing menu (35%)
Italicizing menu (34%)
Temporal menu (32%)
Twisting menu (32%)
Pink menu (28%)
Out-of-context menu (27%)
Leaf menu (25%)
Rotating menu (10%)

Figure 1: The twenty conditions (graphical adaptive menus) used in the experiment. The percentage reports the preference rating computed by the Bradley-Terry-Luce (BTL) model [8], as reported in [52].

adaptive user interfaces, user experience, and brain activity measurement using EEG. Section 3 discusses existing empirical studies that make use of neurophysiological data to support SE activities and studies that compare adaptive user interfaces. Section 4 describes the design and execution of the internal replication. This section also provides an overview of the results of the baseline experiment. Section 5 presents the data analysis and the interpretation of the results. It also analyses possible threats to the validity of the results obtained. Finally, Section 6 presents our conclusions and some future avenues for this work.

2 BACKGROUND

2.1 Adaptive User Interfaces

User needs, wishes, and preferences may change over time, and the system should be able to identify those circumstances and adapt itself to meet the new needs or preferences. Adaptive systems are systems that can modify aspects of their structure or functionalities to satisfy different users' needs and track their changes over time. Adaptive user interfaces are adaptive systems that are capable of adjusting the display, adapting the organisation or presentation of the User Interface (UI) functionality in response to some characteristic of the user context [21, 24].

Various strategies to support the adaptation of UIs have been proposed in the literature. These strategies cover the UI as a whole or certain elements of the user interface. In a recent study [52], a set of 49 graphical adaptive menus was collected which provided a new exploration of a design space based on Bertin's eight visual variables (i.e., position, size, shape, value, colour, orientation, texture, and motion) [7]. This design space allowed the authors to describe a set of graphical adaptive menus that can be expanded and compared. Graphical adaptive menus are graphical user interface menus whose predicted items of immediate use can be automatically rendered in a prediction window (via a subset of menu items resulting from a prediction scheme). The different graphical adaptive menus highlight the prediction window with the purpose of improving users' efficiency when using menus. However, experiments are needed to assess how different design choices affect the user experience.

In this paper, we present a replication of an experiment that compares a sub-set of twenty graphical adaptive menus proposed in [52]. The selected menus cover most of Bertin's visual variables. Specifically, we selected Adaptive activation area, Blinking, Bolding, Boxing, Ephemeral, Glowing, Greyscale, Highlighting, Italicising, Leaf, Morphing, Out-of-context disappearing, Pink, Pulsing, Rotating, Split without replication, Temporal, Twisting, and Underlying. Additionally, as a baseline, we also used the Static menu, which does not use the prediction window and does not highlight any specific menu item. Figure 1 shows all the selected menus. For example, Underlying and Italicising menus are value-changing menus (i.e., menus that change the font of the prediction window, italicising or underlining the menu items to highlight them). Out-context disappearing menu is a motion-changing menu (i.e., menus that move menu items in order to highlight them). And Leaf menu is a menu with unusual shapes.

2.2 User Experience

ISO 9241-210 [17] defines the term user experience (UX) as *"a person's perceptions and responses that result from the use and/or anticipated use of a product, system, or service"*. Therefore, the user experience includes all the effects the use of an interface may have on the user, before, during, and after use. The two major categories of UX evaluation methods are termed as: *subjective*, usually measured through interviews, surveys, ratings and questionnaires, and *objective*, usually measured through user performance measure, physiological and neurological attributes' response, etc. [56]. The subjective approaches have limitations that may give biased information. The responses can be influenced by how users remember their experience and not by the real experience itself [36]).

This inaccuracy in UX measurement has led us to look for alternatives. It is already known that physiological reactions such as blinking, sweating, eye movement, or brain activity can be used to objectively measure UX [6]. As reported in a recent systematic literature review by Zaki et al. [56], UX can be measured using neurological and physiological measures. For example, EEG was used to evaluate UX through user preferences; facial expressions were recorded to detect emotions or facial decoding, and facial movements were recorded to infer emotional response. Furthermore, electrocardiograms were used to detect heart rate variability; eye tracking was used to find users' attention; galvanic skin response was measured to determine sweat gland reaction and, finally, electromyography was used to capture muscle activity.

In this replication, we used both EEG and post-experimental questionnaires to measure the UX produced when using the selected twenty graphical adaptive menus mentioned in Section 2.1. We decided to use post-experimental questionnaires, despite the disadvantages mentioned before, because we also want to study whether these subjective measures of UX correlate with the objective measures obtained by using EEG.

2.3 Brain Activity Measurement using EEG

An electroencephalogram, or EEG, is a technique that measures brain activity by recording changes in electrical activity, also called brainwaves, generated by neurons via electrodes placed on the scalp. Since the magnitude of electrical activity measured at the electrodes is extremely small (microvolts, μV), the recorded data is digitalized and sent to an amplifier to read it. The amplified data can then be displayed as a sequence of voltage values at some sampling rate on a digital graph on a monitor. Brain waves can be captured and analysed to get the user's emotional status while doing any task. This means that we can detect whenever a user is concentrated, relaxed, happy, excited, etc.

2.3.1 *EEG frequencies.*

EEG brain waves can be characterised based on multiple factors such as location, amplitude, frequency, morphology, continuity (rhythmic, intermittent, or continuous), synchrony, symmetry, and reactivity. However, the most commonly used method to classify EEG brainwaves is by frequency. Greek numerals are used to classify the different frequencies of each EEG wave. There are multiple frequencies to consider when talking about brainwaves, but the most commonly studied frequencies are delta (0.5 to 4Hz), theta (4 to 7Hz), alpha (8 to 12Hz), sigma (12 to 16Hz) and beta (13 to 30Hz) [45].

EEG records electrical activity of the brain. To get each frequency, a bandwidth filtering must be used on the EEG recordings. The potential changes in brain waves can be used as measures of specific brain activities correlated with multiple functions such as perception, movement, as well as cognitive processes related to attention, learning, and memory [5]. As reported in the literature, brain wave oscillations reflect what is happening in the participant's information processing situation, even if the participant is unaware of the changes or is unable to verbalise them ([5], [32]). In this way, it is possible to detect the cognitive load, concentration, excitement, relaxation, and several other features of the user's status by detecting brainwaves changes and oscillations.

2.3.2 *EEG data acquisition.*

Electrodes are used to measure electrical brain activity. These sensors are placed on the scalp using a cap or headset. There are many different devices that can read brain electrical activity data. There are some factors that make them different, for example, the type of electrode (dry / saline / gel), the size and shape of the cap, and the type of device (wired / wireless) [27]. The amplifier specifications may also vary (e.g., sampling rate, bandwidth, and resolution).

2.3.3 *UX measures obtained via EEG.*

The Autonomic Nervous System (ANS) is viewed as a major component of the user's emotion response in many theories of emotion. Empirical studies have shown that multiple emotions can be obtained from the analysis of ANS responses. Emotions can be divided into negative emotions (e.g., anger, anxiety, disgust, embarrassment, fear, and sadness) or positive emotions (e.g., affection, amusement, contentment, happiness, joy, pleasure, pride, and relief) [35]. In EEG studies, the analysis of different brain zones has been found to provide different UX measures. For example, several studies have been conducted to infer the role of frontal brain activity in emotion [1]. Also, the symmetry in frontal brain activity in emotional processes was analysed. Differences between brain zones were detected in these studies. For example, the study by Harmon-Jones et al. [25] suggests that differences between left and right frontal cortical activity is associated with positive emotions like enthusiasm or negative emotions like anger.

Although brain zones are important in the analysis of emotions, brain frequencies are equally important. Alpha and theta waves have been found to be correlated with cognitive and memory performance [30]. Additionally, if beta bandwidth is also considered, it is possible to measure the user engagement ([47], [18], [42]).

Finally, every person may have different brain responses during relaxation or stress situations. Therefore, a baseline sample is necessary for each participant for relaxed and stressed situations. This baseline sample is called "calibration sample". For example, the most used calibration sample for a relaxed state is the one in which the participant is asked to close their eyes to measure their brain activity without physiological impulses.

3 RELATED WORK

Over the last few years, EEG has been used in educational contexts to measure cognitive load. For example, in [43] the authors measured the cognitive load produced by the review of proposals for Master's theses. They used an EEG headset to measure the participants' attention rate while reviewing the proposals. The main goal of this work was to provide guidelines on how to plan and execute this type of experiments. Similarly, in [2] EEG has also been used to assess the cognitive load produced in the context of hypertext and multimedia learning. The main goal was to provide evidence for the feasibility of using EEG in educational research to collect and analyse the cognitive load to test the effectiveness and improve the design of learning materials. In both studies, the authors concluded that EEG is a good way to record brain activity to objectively measure the cognitive load produced by different tasks.

In addition, some studies have explored the use of psychometric data to measure cognitive load in SE tasks such as code comprehension, code readability, and error detection. For example, Siegmund et al. [49] used fMRI to determine the cognitive load registered by code comprehension tasks and locate syntax errors to contrast them. They found a different activation pattern of five regions of the brain related to working memory, attention, and language processing.

Similarly, Fakhoury et al. [14] used fNIRS to explore the effect of poor source code lexicon and readability on developers' cognitive load. Additionally, they used an eye-tracker to identify which parts of the source code produced more cognitive load. The authors concluded that the presence of linguistic anti-patterns in source code significantly increases the developers' cognitive load. In another study, Kosti et al. [33] identified brain signatures that are specific to code comprehension using EEG only. They trained a model of subjective difficulty based on the recorded brainwaves to predict the difficulty of code comprehension tasks.

Moreover, as acknowledged by Feldt et al. [15] empirical SE studies should focus not only on the developer's perception of the system, but also on the perception of the end user. The user experience is usually measured by means of interviews or questionnaires after the system use. For example, cognitive load was traditionally measured using the NASA-TLX questionnaire to correlate it to the user satisfaction [48]. However, these responses can be subject to inaccurate user recall or other subjective factors, leading to imprecise measurements ([36], [37]).

Some experiments are starting to include physical and physiological data in their data collection. For example, [38] used fMRI and EEG to describe how the brain reacts when users see different designs. The authors concluded that fMRI showed that the perception of different feelings towards designs is associated with the frontal and occipital lobes, and the EEG showed that the human brain responds sooner and stronger in its perception of bad feelings. In another study, Hou et al. [26] evaluated an Air Traffic Control system using EEG-based tools to monitor and record the brain states of air traffic controllers. The authors also analysed the relationship between the mental workload calculated using the traditional NASA-TLX method and that calculated using the method for labelling EEG data. They found that in most of the simulations the data were highly correlated.

Some studies that compare adaptive user interfaces have also been performed. For example, Gajos et al. [22] implemented and evaluated three graphical adaptive user interfaces in two experiments together with a non-adaptive baseline. The authors concluded that adaptive UIs have an impact on users depending on their particular properties. The interfaces which frequently duplicate (rather than moving) tend to improve users' performance and satisfaction. Other researchers [16], focusing on adaptive graphical menus, compared two adaptive interface designs and found that these provided more consistently positive results in terms of performance and user satisfaction.

The experiments mentioned above focus on detecting differences between brainwave frequencies and the cognitive load. As mentioned in [26], emotions can also change and this is also important in UX evaluations. In this study, data from EEG and questionnaires are used to measure not only the cognitive load, but also the valence (i.e. attraction), the memorisation effort, and the user engagement produced by the interaction with different user interfaces.

4 EXPERIMENTATION

This section first presents an overview of the baseline experiment followed by the design and execution of the internal replication. We followed the guidelines proposed by Wohlin et al. [55].

4.1 Overview of the baseline experiment

Different user interfaces may lead to different emotions, mental effort, and, overall, a different user experience among users. There are several types of user interface elements that could be considered. In order to focus on one type of element, we considered the different types of graphical menus proposed in [52] (see Section 2.1). This set of graphical adaptive menus has been subject to a preference analysis, but no empirical study was performed to compare the impact of the different menu types on the user experience. Accordingly, we performed an exploratory experiment to analyse this impact [4], and our findings are briefly introduced in the following sections.

4.1.1 Experiment context, goal and design. The experiment was conducted with 21 volunteers with diverse backgrounds (i.e., software engineers, mechanical engineers, financial experts, healthcare professionals) and a wide range of ages (from 18 to 63 years old with an average of 35 years old). The volunteers were recruited through an open call for participation at the [eliminated for double-blind review] with a snowball procedure.

The goal of the experiment was to **analyse** a set of adaptive graphical menus **for the purpose of** comparing them with respect to the user experience produced in terms of cognitive load, engagement, memorisation and valence **from the point of view of** both researchers and novice user interface designers **in the context of** an end-user group at the [eliminated for double-blind review] and their contacts. To this end, we investigated the effect of using menus with different properties (e.g., position-changing, orientation-changing, size-changing, shape-changing, value-changing, colour changing, texture-changing, and motion-changing). We selected a sub-set of twenty menus covering most of the properties. We used an EEG headset to obtain the UX measures produced. By means of this device, we can obtain four variables related to UX: workload, engagement, memorisation and valence. Participants brainwaves were registered while using twenty different menus to obtain these UX metrics. The alternative hypotheses tested in the experiment aimed to prove whether there were differences in the metrics obtained between the used menus.

4.1.2 Experiment Results. The results of the experiment showed that using different types of graphical adaptive menus can cause some differences in the UX, but the analysis of factors showed that there were no statistically significant differences in engagement, valence, memorisation, or cognitive load registered by the EEG device (i.e., workload p-value=0.707, engagement p-value=0.901, memorisation p-value=0.730, valence p-value=0.991). One possible reason for this finding is that it could be due to the different backgrounds and range of ages of the participants. For more details about the baseline experiment please refer to [4].

4.2 Internal replication

We carried out a strict internal replication of the baseline experiment. The same experimental protocol was applied, but to a different population, meaning that we changed only the participants, while the site, experimenters, design, variables, and instrumentation remained the same. The purpose of this replication is to test the extent to which the study results could be generalized to other populations (i.e, a more homogeneous group of users). Strict replications are needed to increase confidence in the validity of the experiment's conclusion. An additional motivation is that the sample size of the baseline experiment can affect the magnitude of

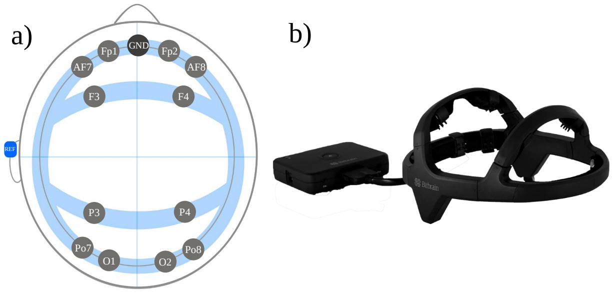


Figure 2: The Diadem brain computer interface. a): The location map of the 12 electrodes based on international 10-10 system, including REF and GND. b): Bitbrain's Diadem headset and amplifier records EEG signal at sampling rate 256 Hz.

the effect (i.e. the strength of the relationship between the type of graphical menu and the UX measures of that population). Therefore, it is convenient to run internal replications to increase the sample size, thus improving the effectiveness of confirming the experiment hypotheses.

4.2.1 Research questions. The goal of this study is to determine how graphical adaptive menus impact the user experience, measured using an EEG device. Furthermore, we would like to assess whether the UX measures gathered using EEG are correlated with the cognitive data that are obtained through traditional UX questionnaires. The perspective is that of novice UI designers interested in obtaining empirical evidence about the effect of different graphical adaptive menus on the emotions and cognitive load of end users. Hence, the study aims to answer the following research questions:

- RQ_1: Do the 20 graphical adaptive menus have a different influence on the user experience?
- RQ_2: Does the user experience measured using EEG signals correlate with the subjective ratings obtained using traditional questionnaires?

4.2.2 Selected BCI device, variables and questionnaires. There is an independent variable, which is defined by the type of graphical adaptive menu being compared. This is a nominal variable that could assume twenty possible values (see Figure 1). There are also two types of dependent variables: emotion-based and perception-based variables. Emotion-based variables assess the emotions and feelings of participants when performing the experimental task (i.e, selecting graphical menu items). In this experiment, a Diadem headset [50] was used to capture the users' EEG signals. It is a portable non-intrusive EEG headset widely used for neuroscience research and brain computer interfaces (BCI). It is optimized to estimate emotional and cognitive states (prefrontal, frontal, parietal, and occipital brain areas).

The Diadem headset has twelve dry electrodes placed at positions AF7, Fp1, Fp2, AF8, F3, F4, P3, P4, PO7, O1, O2 and PO8 according to the international 10-10 standard [11], as shown in Figure 2. This headset also uses the left earlobe (A1, REF) as a reference electrode and the centre of the head (Fpz, GND) as the ground. The headset is wired to an amplifier that communicates via Bluetooth with the researchers' computer.

Table 1: Selected variables

Variable	Values	Description
Menu type	Nominal	20 graphical adaptive menus reproduced in Figure 1
Workload	Continuous	Subject's mental effort, as measured using EEG
Engagement	Continuous	Subject's engagement, as measured using EEG
Memorisation	Continuous	Subject's mental effort to memorise, as measured using EEG
Valence	Continuous	Subject's attraction, as measured using EEG
Perceived Workload (PW)	Ordinal	Subject's mental effort, as measured using the NASA-TLX questionnaire
Perceived Valence (PV)	Ordinal	Subject's attraction, as measured using the UEQ questionnaire

Specifically, this device allowed us to measure the following emotion-based variables:

- **Workload**: measures the concentration of a participant when presented with stimuli or during experiences. It is expressed as a percentage. Values close to 0% indicate that the participant is very distracted while values close to 100% indicate that they are very attentive to the stimulus.
- **Engagement:** measures the degree of involvement or connection between the participant and the stimulus or task. This is a more complex indicator than workload, as a participant can be attentive to a task even if the information presented is not of interest. It is expressed as a percentage. A value close to 0% indicates that there is no connection to the stimuli while a value close to 100% indicates high engagement with the stimuli or task.
- **Memorisation**: measures the intensity of cognitive processes related to the formation of future memories during the presentation of stimuli or during an experience. Captures the degree of memory storage, encoding, and retention. It is expressed as a percentage. A value of 0% indicates that the probability that the stimulus will be remembered is low while a value close to 100% indicates a high possibility that the stimulus will be retained in the participants' memory.
- **Valence**: measures the degree of attraction experienced in response to a stimulus or a situation, from a positive (pleasant) reaction to a negative (unpleasant) reaction. It is expressed as a percentage. A value close to 100% positive (pleasant) or 100% negative (unpleasant) is equivalent to the response measured as baseline during the calibration phase. A valence level exceeding 100% (positive or negative) is possible if the calculated reaction exceeds that measured during calibration.

Furthermore, two perception-based variables were used to assess the participants' perceptions of their user experience when interacting with the adaptive graphical menus: Perceived Workload and Perceived Valence. These variables were measured using the standardized NASA-Task Load Index (NASA-TLX) and the User

Experience Questionnaire - Short (UEQ-S), which are both widely applied and empirically validated survey questionnaires.

The NASA-TLX questionnaire is a tool for measuring subjective mental workload. It rates performance across six dimensions (mental demand, physical demand, temporal demand, effort, performance, and frustration) to determine an overall workload rating. This questionnaire is divided in two parts. First, each dimension is rated within a range of 100 points with five-point steps. The second part aims to create an individual weighting of these dimensions by letting the participants compare them pairwise based on their perceived importance (i.e., the participants choose which dimension is more relevant to workload).

The UEQ-S is used to obtain the hedonic quality (HQ), which is the best predictor of the valence of an experience [51], and the pragmatic quality (PQ), which is also relevant for UX measurements. It consists of eight items that are grouped into two scales (i.e., hedonic and pragmatic quality). Each item has a positive and a negative value that must be rated from 1 (being the negative value of the item) to 7 (being the positive value of the item), with 4 being the neutral value. The items used and their values, as well as the questionnaires used in this experiment (in Spanish), can be found at: https://dagasfi.github.io/.

4.2.3 Hypothesis formulation. The following null hypotheses were defined to verify the first research question, namely whether the use of twenty graphical adaptive menus have a different impact on the user experience:

- H_{n10}: There are no significant differences in the users' workload when using different graphical adaptive menus.
- H_{n11}: There are no significant differences in the users' engagement when using different graphical adaptive menus.
- H_{n12}: There are no significant differences in the users' valence when using different graphical adaptive menus.
- H_{n13}: There are no significant differences in the users' memorisation when using different graphical adaptive menus.

Similarly, the following null hypotheses were defined to verify the second research question, namely whether the user experience measured using EEG signals correlate with the subjective ratings obtained using traditional questionnaires:

- H_{n20}: There is no correlation between the workload measured using the EEG signals and the perceived workload measured using the NASA-TLX questionnaire.
- H_{n21}: There is no correlation between the valence measured using EEG signals and the perceived valence measured using the UEQ-Questionnaire.

The goal of the statistical analysis was to reject these hypotheses and possibly to accept the alternative ones (e.g., $H_{a10} = \neg H_{n10}$). All the hypotheses are two-sided because we did not postulate that any effect would occur because of the graphical adaptive menu usage.

Although we could obtain four UX measures using the selected EEG device, we nevertheless decided to correlate cognitive load and valence only. As suggested by Georges et al. [23], these cognitive and emotional states are the most important ones when measuring UX. Similarly, Tuch et al. [51] suggested that valence, since it is an emotional state, is a very important factor that should be taken into account when modeling user experience.

In addition, to mitigate a potential maturation threat (i.e., subject tiredness) we limited the number of questionnaires that the participants had to complete. Instead of giving them two questionnaires (NASA-TLX and UEQ-short) for each set of the twenty graphical adaptive menus, we asked them only to fill out these two questionnaires for just six of the graphical adaptive menus (i.e., the three menus that the participants liked the most and the three menus that the participants mostly disliked, according to the results of the baseline experiment).

4.2.4 Context and subject selection. The participants included in this study were selected by means of convenience sampling. They were 23 students enrolled on a Master's degree in Computer Science at the *[eliminated for double blind]*. Specifically, seventeen of them were male and six were female. Their ages ranged from 20 to 40 years old. They all had experience using user interfaces either from their studies, professional occupations or hobbies. The number of participants is in line with previous UX studies. As suggested by Apraiz et al. [3], a sample size for UX tests performed with physiological monitoring cannot yet be determined, but these authors indicate that the median sample size in EEG studies is eighteen participants. All the participants were volunteers and were made aware of the practical and pedagogical purposes of the experiment, but the research questions were not disclosed to them. The participants were not rewarded for their effort.

4.2.5 Design of the replication. The experiment is a one-factor measures design. We carried out one EEG session per participant. We represented the different graphical adaptive menus using an interactive slide presentation. Our menu representations were based on two different domains: a mail manager software and a web browser. We selected menus with almost twenty menu items to make it harder to look for a specific menu item. Each participant was asked to use all types of menus and in a random order so as to ensure that the results of a menu were not influenced by the order in which they were used. Specifically, we prepared an interactive slide presentation for each participant where the twenty menu types and the two domains were randomised (i.e., ten menu types with each domain). Between each stimuli (i.e., use of a particular menu type), there was a short period for rest, also known as the "washout phase". During this period the participants were instructed to close their eyes and enter a resting state. The washout block is meant to de-couple the emotional response from one phase to the next one, and to serve as a break where the participant clears their emotional responses.

4.2.6 Preparation and Data analysis. Prior to the experiment, each participant signed a consent form. Next, they were given a set of instructions to inform them about the experiment protocol. An experimenter was also present to answer any questions. Each EEG session started with the headset placement and calibration phase that lasted about eight minutes. It is mandatory to establish a baseline brain activity for each participant to compute the UX metrics. To this end, the calibration phase is composed of five steps:

1. Calibration task familiarisation (1 min.): The participant learns to perform the calibration task.
2. Closed eyes baseline (2 min.): The goal is to register the brain response of the participant in a relaxed state with no other

physiological responses like blinking, eye movements, etc. It is used to establish a null baseline for EEG data.

(3) Open eyes baseline (2 min.): The goal is to register the brain response of the participant in a relaxed state also considering other physiological responses like blinking or eye movements. It is used to establish a baseline measurement in EEG.

(4) Calibration baseline (1 min.): The goal is to register the brain response of the participant in a controlled stress state. The participant will have to count backwards in intervals of seven, out loud, starting with a number generated randomly.

(5) Closed eyes washout (2 min.): The goal is to clear every participant reaction before starting the following steps.

To familiarise participants with the menu types, participants performed a practice trial with a different graphical adaptive menu. This trial could be repeated as many times as needed. Once familiarised, and after the resting phase, they could start the experimental task: selecting a menu item for each one of the twenty graphical adaptive menus. The experimental task is mainly composed of three repeated steps: 1) Use the graphical adaptive menu with no time limit; 2) Resting phase (ten seconds) and 3) Open eyes.

In step 1, one of the graphical adaptive menu type is randomly shown on the screen. First, the menu is hidden, and we asked the participant to select a specific menu item which is in the predicted window that will be highlighted by the menu. Then, after expanding the menu, all the menu options are displayed, and the participants should select the item asked previously. In step 2, after selecting the menu item, participants are asked to close their eyes, relax, and do nothing for ten seconds. This resting phase was used to washo ut emotions (e.g., to separate the frustration from one menu to another). In step 3, the participant is asked to open their eyes again, and repeat the process from step 1. During the entire process, all mouse input and brainwaves are recorded.

Finally, participants filled in the two questionnaires to assess the perceived workload and perceived valence when using six out of the twenty menu types. As explained in Section 4.2.3, we selected only the six menus that were found to have the best/worst performance according to the baseline experiment [4].

The results of the experiment were collected using the Bitbrain's software "SennsLab" and the questionnaires. SennsLab records the EEG signals in real-time (i.e., brainwaves of each participant) which are then processed using SennsCloud to compute the UX measures in relation to the brain zones, brainwave bandwidth, and calibration samples from the EEG data. We then used SennsMetrics[1], Excel and R Studio to analyse the data collected.

We used descriptive statistics, histogram plots, and statistical tests to analyse the data collected. As is usual, in all the tests, we accepted a probability of 5% of committing a Type-I Error, i.e., rejecting the null hypothesis when it is in fact true. The data analysis was carried out as follows:

(1) We first carried out a descriptive study of the measures for the dependent variables.

(2) We analysed the characteristics of the data to determine which test would be most appropriate to test our hypotheses. Since the sample size was less than 50, we applied the Shapiro–Wilk test, so as to test the normality of the data, and the Brown-Forsythe Levene-type test to determine the homogeneity of any variances.

(3) The results of the tests were then employed as a basis on which to test the null hypotheses formulated. When the data were normally distributed and the variances were homogeneous, we used one-way Analysis of Variance (ANOVA) to analyse the data by considering the menu type as a main factor [54]. When the ANOVA assumptions could not be satisfied, we used the Kruskal–Wallis test) to compare the mean averages of the twenty treatments.

(4) We analysed both the NASA-TLX and the UEQ-S questionnaires. With the NASA-TLX we used an Excel file which computes the subjective workload based on the users' responses. The UEQ-S data was analysed using the UEQ Data Analysis Tool which is available free of charge on the UEQ homepage[2]. This tool is also an Excel file which facilitates data analysis. First, we entered the data from the UEQ-S into the Data worksheet. Then, the tool calculated all the statistics necessary to interpret the results and automatically created diagrams.

(5) We analysed the correlation of the UX measures obtained using the EEG signals with the perceived UX measures collected using the questionnaires. We used Pearson's correlation coefficient to analyse the data by comparing the means of the objective measure with the corresponding subjective measure (i.e., Valence and Perceived Valence). When the normality assumption did not hold, we used the Spearman's correlation coefficient.

(6) The statistical significances of the experiment were complemented with the magnitude of their effects. For this purpose, Cliff's δ estimates [10] were obtained with a confidence interval of 95%. These measures are recommended when dealing with ordinal scale data [29]. Moreover, the non-parametric nature of Cliff's δ estimates serves to reduce the influence of distribution shape, differences in dispersion, and extreme values. The magnitude of the effect was assessed using the thresholds provided by Kraemer and Kupfer [34], i.e., |d| < 0.112 "negligible ", |d| < 0.276 "small" |d| < 0.428 "medium", otherwise "large".

5 RESULTS

Figure 3 visually presents the descriptive statistics of the emotion-based variables obtained for each menu type. Error bars show 95% confidence intervals. We used the *Static* menu as a baseline because its properties are constant.

At a glance, we can observe that the participants had to make less mental effort (workload) when using the *Pink* (min: 0.934; max: 73.391; mean: 30.443; SD: 15.423) and *Greyscale* (min: 1.883; max: 60.596; mean: 31.376; SD: 13.572) menus, and that they had to concentrate more when using the *Underlying* (min: 1.671; max: 98.541; mean: 41.177; SD: 22.348), *Bolding* (min: 1.132; max: 132.598; mean: 40.404; SD: 26.063) and the *Leaf* (min: 22.551; max: 105.147; mean: 40.018; SD: 16.621) menus.

The participants felt more engaged with the *Underlying* (min: 16.571; max: 79.181; mean: 40.904; SD: 14.826) and *Temporal* (min:

[1] https://www.bitbrain.com/neurotechnology-products/software/sennsmetrics

[2] https://www.ueq-online.org/

17.220; max: 57.889; mean: 39.943; SD: 10.285) menus, and they felt less engaged with the *Greyscale* (min: 9.507; max: 73.338; mean: 34.263; SD: 14.505) and *Out of context disappearing* (min: 9.819; max: 54.857; mean: 34.731; SD: 11.047) menus.

We can also observe that the intensity of cognitive processes to create future memories was low for the *Pink* (min: 2.823; max: 39.037; mean: 19.798; SD: 8.975) and *Ephemeral* (min: 7.016; max: 43.491; mean: 22.260; SD: 8.804) menus. In contrast, it was high for the *Twisting* (min: 6.632; max: 70.564; mean: 29.835; SD: 14.130) and *Leaf* (min: 0.451; max: 47.162; mean: 28.449; SD: 11.969) menus.

Finally, we can see that the *Rotating* (min: -40.364; max: 58.165; mean: 10.775; SD: 23.430), *Underlying* (min: -19.150; max: 48.422; mean: 9.428; SD: 15.759) and *Ephemeral* (min: -61.598; max: 47.794; mean: 9.292; SD: 23.673) menus registered a high level of attraction while the *Italicising* (min: -70.668; max: 33.191; mean: -0.364; SD: 24.967), *Bolding* (min: -81.614; max: 75.372; mean: 0.678; SD: 28.551) and *Adaptive-activation area* (min: -27.567; max: 53.314; mean: 0.972; SD: 17.718) menus registered the lowest values. Note that all menu types registered a higher value than the *Static* menu, suggesting that participants were pleased to see something different rather than the classic menu which did not change any property.

Overall, these results suggest that the graphical adaptive menus cause different levels of workload, memorisation and valence in the participants, while their engagement remains almost unchanged when compared to the *Static* menu.

5.1 RQ1: Influence of menus

The Shapiro-Wilk test revealed that only the Engagement variable followed a normal distribution for all the menu types (i.e., p-value > 0.05). The Brown-Forsythe Levene-type test revealed that the four variables are homogeneous for all the menu types. Table 2 shows the statistical tests that were applied to each emotion-based variable and their respective significance level. The Kruskal-Wallis test revealed that the null hypotheses H_{n10}, H_{n12}, H_{n13} could not be rejected (i.e., workload p-value = 0.47618, memorisation p-value = 0.50563, valence p-value = 0.53301). Also, the One-way ANOVA test revealed that the null hypothesis H_{n11} could not be rejected (i.e., engagement p-value = 0.97755). These results indicate that there are no statistically significant differences in the participants' workload, memorisation, valence and engagement when using the twenty graphical adaptive menus.

Nevertheless, we found practical significance for hypothesis H_{n12} using Cliff's δ. As shown in Table 2, valence has a medium effect size while workload, engagement and memorisation showed a negligible effect size. These results suggest that the adaptive menus studied have a different impact on the attractiveness of user interfaces for the participants.

5.2 RQ2: Survey results and correlation analysis

Figure 4 visually summarises the answers of the UEQ-S questionnaires. The items 1 to 4 concern the pragmatic quality while the items 5 to 8 concern the hedonic quality.

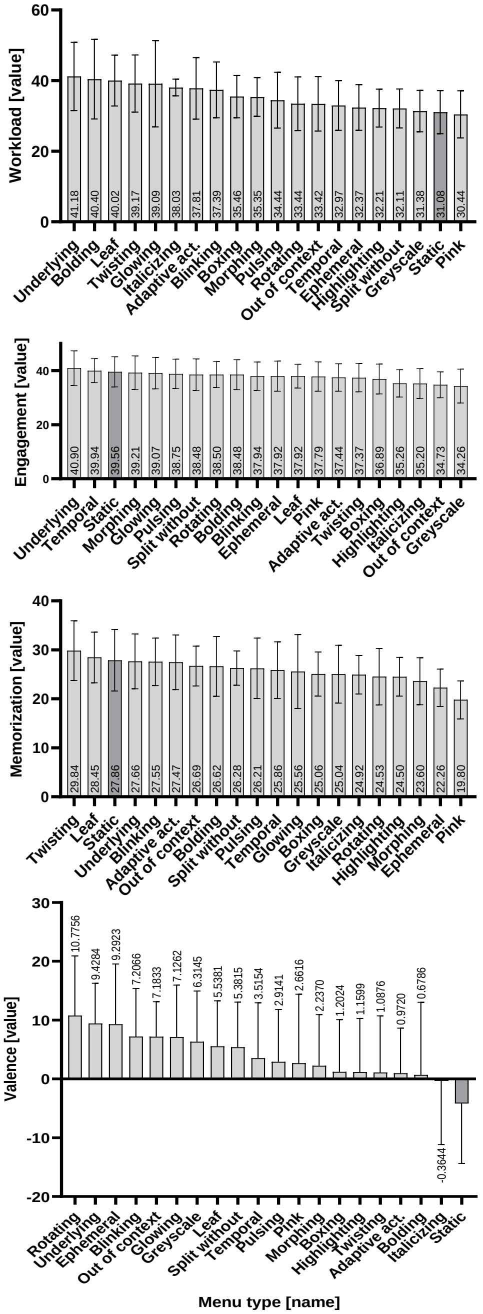


Figure 3: Emotion-based variables obtained for each menu type, in decreasing order of their mean value, in comparison to the static menu (darker colour) as a baseline.

The analysis of the answers collected from items 1 to 4 showed that the pragmatic quality of the menus can vary considerably. In particular, the *Italicising* and *Leaf* menus can be considered to be obstructive, complicated, inefficient and confusing when selecting a menu item from a prediction window. On the other hand, the *Out context disappearing* and *Blinking* menus can be considered to be very helpful to highlight the prediction window. This means that menus with different structures than usual and menus with different font types (e.g., italics, bold) are perceived to be worse in terms of ease of use rather than motion-changing or position-changing menus. The analysis of the answers collected from items

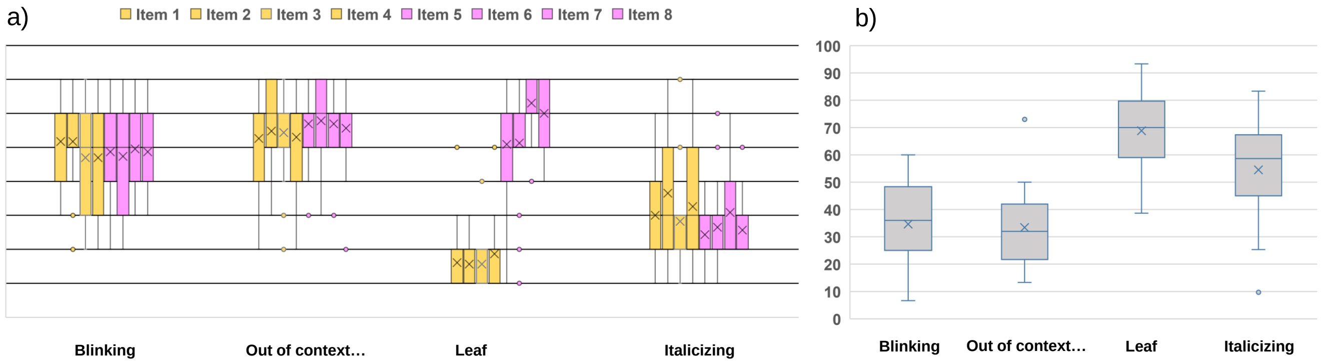


Figure 4: a) UEQ-S answers for the internal replication. Yellow-coloured are pragmatic quality item responses while Pink-coloured are hedonic quality item responses. b) NASA-TLX weighted answers

Table 2: Hypotheses testing

Variable	Method	p-value	Cliff's δ estimate	Effect size
Workload	Kruskal-Wallis	0.47618	-0.058	Negligible
Engagement	One-Way ANOVA	0.97755	-0.032	Negligible
Memorization	Kruskal-Wallis	0.50563	-0.051	Negligible
Valence	Kruskal-Wallis	0.53301	-0.285	Medium

Table 3: Correlation analysis results

Menu	Valence	PV	Workload	PW
Blinking	7.206	0.858	37.389	34.652
Split without replication	5.381	-0.271	32.112	30.695
Out of context disappearing	7.183	1.684	33.420	33.434
Morphing	2.236	0.065	35,347	38.188
Italicizing	-0.364	-1.315	38,030	54.536
Leaf	5.538	1.597	40.018	68.840
Variables			Correlation	p-value
Valence-Perceived Valence (PV)			0.824	0.043
Workload-Perceived Workload (PW)			0.87	0.02337

5 to 8 suggest that almost every menu was perceived as a visually attractive menu (e.g., *Blinking*, *Out of context disappearing* and *Leaf menus*). That is, almost every menu produced a positive valence. Only the *Italicising* menu was perceived as unattractive. Other menus such as the *Split without replication* and the *Morphing* menu were perceived as neutral menus.

Figure 4 visually summarises the weighted perceived workload obtained from the NASA-TLX questionnaire for each menu type. The results show that *Italicising* and *Leaf* menus are considered to be more cognitive demanding than the other menu types. Note that the pragmatic quality (UEQ-S) and the perceived workload are inversely correlated (i.e., when pragmatic quality increases, perceived workload decreases). These findings suggest that some menus are considered to be attractive but difficult to use (e.g., *Leaf* menu), or attractive and easy to use (e.g., *Out of context disappearing* menu) or unattractive and difficult to use (e.g., *Italicising* menu). Therefore, these different dimensions could be useful when designing UIs that provide a good perceived UX.

After analysing the questionnaire results, we performed the correlation analysis. First, the Shapiro–Wilk test performed for the six menu types revealed a normal distribution for all the variables. We thus applied Pearson's correlation coefficient to compare the means of the emotion-based measures with the corresponding perception-based measures. Table 3 summarises the correlation analysis results. These results suggest that hypotheses H_{n20} and H_{n21} can be rejected, meaning that valence and workload are highly correlated with perceived valence and perceived workload, respectively. This means that EEG signals can effectively be used to evaluate some dimensions of user experience (i.e., workload and valence).

5.3 Discussion

Regarding the first research question, the results suggest that there are no statistically significant differences among different graphical adaptive menus. Nevertheless, we discovered some patterns that may be useful when designing new graphical adaptive menus. For example, as shown in Figure 3, using graphical adaptive menus will likely improve the Valence produced on the users, but may increase the Workload. This effect may be produced by the first impression and by interpreting the animations and highlights. Perhaps, the regular use of graphical adaptive menus will reduce the Workload because users will get used to them, thus meaning that there is no additional mental workload to interpret and understand the highlights. Further experimentation with the same participants may increase the reliability on this thoughts.

Although there are no statistically significant differences, the analysis was complemented with the magnitude of their effect size. Valence was the only variable that showed a medium effect size. The other EEG variables showed negligible effect size. This suggests that, for this context of users, only Valence was important and Workload, Memorization and Engagement were not influenced by the different menus used. Nevertheless, further experiments with a larger sample size should be carried out.

Regarding the second research question, we can conclude that emotion-based valence and workload are highly correlated with perception-based valence and workload, respectively. This correlation suggests that the metrics calculated from the EEG are accurate. Thus, EEG devices can be effectively used to obtain UX measurements easier and faster than traditional questionnaires. Also, the questionnaires results suggest that the graphical adaptive menus do not always need a high pragmatic quality in order to produce a positive valence on users (e.g., *Leaf* menu). Overall, there is a trend on the graphical adaptive menus performance. For example, the results suggests that menus which include the temporal dimension improves the UX. That is, menus which first show (e.g., *Temporal*, *Ephemeral*, *Out-of-context disappearing*), move (e.g., *Rotating*) or temporally highlight (e.g., *Blinking*) the prediction window, tend to improve the user experience.

Emotion-based and perception-based variables showed that the debate of preference *vs.* performance is always ongoing [46]. In this experiment, we were able to relate that some users prefer, or find more attractive, menus which do not increase performance. For example, menus with high ratings on both, Valence and Perceived Valence, also registered high workload compared to other menus (e.g., *Leaf* menu). We consider that these findings can be used by UI designers that want to apply adaptive menus to their software. When deciding which menu to use, the ones that registered higher Valence values should be used. Also, we consider that this information could be used in adaptive systems since the bio-information on how humans react to software can improve the adaptation rules to improve the UX through different graphical menus.

5.4 Threats to validity

With regard to internal validity, learning effect, understandability of the slides used to represent the menus, and the instrumentation validity must be considered. The learning effect was mitigated by randomising the order in which the menus were used. The understandability of the slides used to represent the menus was assessed during the experiment, after the calibration phase, as described in Section 4.2.6. Finally, to avoid any possible source of bias, the experimental materials were evaluated by an experienced researcher in Human-Computer Interaction. We mitigated the instrumentation validity by using a specific EEG device from a company which research and develops high-quality EEG devices.

With regard to external validity, the representativeness of the results, and the size and complexity of the tasks that might affect the generalisation of the results must be considered. The representativeness of the results could have been affected by the number of menus that were compared. We believe that the menus selected for this study can be considered as a baseline to obtain indications as to which properties produce better results on UX. However, we are aware that they cannot represent the whole design space of graphical adaptive menus. We decided to use relatively small tasks to enable the subjects to complete the whole experimental session within 60 minutes. Replications with different and more complex graphical menus are nevertheless needed to study the effect of the graphical adaptive menus variables on the observed results.

With regard to construct validity, the measures used to obtain a quantitative evaluation of the subjects' UX (i.e., the ones obtained from the EEG data) and the reliability of the post-experiment questionnaires must be considered. We used the Bitbrain processing service (i.e., SennsCloud) to obtain the objective measures of the subjects' UX from the EEG. This service computes the metrics based on relevant previous research in EEG data analysis which provide details on how to compute Workload ([30], [9]), Engagement ([18], [42]), Memorisation ([31], [40]) and Valence ([1], [25]) from EEG signals. The reliability of the questionnaire as regards assessing the Perceived Valence was tested using the Cronbach's alpha test. For the Blinking questionnaire, questions related to HQ and PQ obtained a Cronbach's α coefficient of 0.87 and 0.89; for the Morphing questionnaire, the result was 0.81 and 0.91; for the Out of context disappearing questionnaire, the result was 0.91 and 0.84; for the Leaf questionnaire, the result was 0.72 and 0.74; for the Split without replication questionnaire, the result was 0.87 and 0.85; and finally, for the Italicizing questionnaire, the result was 0.73 and 0.88. All the results were higher than the threshold level (0.70) [28]. To measure Perceived Workload, we used NASA-TLX which is a widely used survey instrument to measure workload.

Finally, with regard to conclusion validity, the data collection and the validity of the statistical tests applied must be considered. With the purpose of decreasing the data collection threat, we systematically applied the same data-extraction procedure in each session with each participant. This is, before each session, and during the EEG device placement, we made sure that every electrode was in contact with the scalp and was placed correctly according to the guidelines [11]. Regarding the statistical tests, proper tests were performed to test the hypothesis [28]. To select the statistical tests, we considered the design of the experiment and the nature of the variables and their assumptions.

6 CONCLUSIONS AND FUTURE WORK

This paper has presented an internal replication of an experiment that aimed to analyse how twenty graphical adaptive menus impact the participants' UX, measured by using an EEG device. Additionally, we investigated the correlation between the EEG signals and the participants' UX ratings. The results showed that there were no statistically significant differences between the twenty graphical adaptive menus, but a medium effect size for valence was found, suggesting that, for this context of users, only valence was important and workload, memorisation and engagement were not influenced by the different menus used. Nevertheless, further experiments with a larger sample size are needed to verify these findings.

The results also showed that the emotion-based measures of UX are highly correlated with the perception-based measures of UX, suggesting that EEG signals can effectively be used to evaluate some dimensions of UX (i.e., valence and workload). We also find some trends such as that menus that include the temporal dimension, and position-changing menus tend to enhance the UX while menus that use different formats (e.g., font types, typography) degrade it. These results may be useful to UI designers interested in the effect of different graphical adaptive menus on the emotions and cognitive load of end users.

As future work, we plan to run EEG experiments with other biosensors (e.g., eye tracking, galvanic skin response) to increase the richness of the physiological data that are captured while users interact with different types of UI elements.

REFERENCES

[1] John J.B. Allen, James A. Coan, and Maria Nazarian. 2004. Issues and assumptions on the road from raw signals to metrics of frontal EEG asymmetry in emotion. *Biological Psychology* 67, 1 (2004), 183–218. https://doi.org/10.1016/j.biopsycho.2004.03.007 Frontal EEG Asymmetry, Emotion, and Psychopathology.

[2] Pavlo Antonenko, Fred Paas, Roland Grabner, and Tamara Van Gog. 2010. Using electroencephalography to measure cognitive load. *Educational psychology review* 22, 4 (2010), 425–438.

[3] Ainhoa Apraiz Iriarte, Ganix Lasa, and Maitane Mazmela. 2021. Evaluating User Experience with physiological monitoring: A Systematic Literature Review. *Dyna (Bilbao)* 8 (03 2021), 21. https://doi.org/10.6036/NT10072

[4] Anonymous author(s). -. eliminated for double blind review. (-).

[5] E Basar. 1999. Brain function and oscillations. II. Integrative brain function. *Neurophysiology and cognitive processes* (1999).

[6] Jennifer Romano Bergstrom, Sabrina Duda, David Hawkins, and Mike McGill. 2014. Physiological response measurements. In *Eye tracking in user experience design*. Elsevier, 81–108.

[7] J Bertin. 1967. Sémiologie graphique, Paris, Mouton/Gauthier-Villard. *Réédition (2005) EHESS* (1967).

[8] Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39, 3/4 (1952), 324–345. https://doi.org/10.2307/2334029

[9] Anne-Marie Brouwer, Maarten A Hogervorst, Jan B F van Erp, Tobias Heffelaar, Patrick H Zimmerman, and Robert Oostenveld. 2012. Estimating workload using EEG spectral power and ERPs in the n-back task. *Journal of Neural Engineering* 9, 4 (jul 2012), 045008. https://doi.org/10.1088/1741-2560/9/4/045008

[10] N. Cliff. 1993. Dominance statistics: ordinal analyses to answer ordinal questions. *Psychological Bulletin* 144 (1993), 494–509. https://doi.org/10.1037/0033-2909.114.3.494

[11] Electrode Position Nomenclature Committee et al. 1994. Guideline thirteen: guidelines for standard electrode position nomenclature. *J. Clin. Neurophysiol.* 11 (1994), 111–113.

[12] Ricardo Couceiro, Gonçalo Duarte, João Durães, João Castelhano, Catarina Duarte, César Teixeira, Miguel Castelo Branco, Paulo Carvalho, and Henrique Madeira. 2019. Biofeedback augmented software engineering: monitoring of programmers' mental effort. In *2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*. IEEE, 37–40.

[13] Aruna Duraisingam, Ramaswamy Palaniappan, and Samraj Andrews. 2017. Cognitive task difficulty analysis using EEG and data mining. In *2017 Conference on Emerging Devices and Smart Systems (ICEDSS)*. IEEE, 52–57.

[14] Sarah Fakhoury, Yuzhan Ma, Venera Arnaoudova, and Olusola Adesope. 2018. The effect of poor source code lexicon and readability on developers' cognitive load. In *2018 IEEE/ACM 26th International Conference on Program Comprehension (ICPC)*. IEEE, 286–28610.

[15] Robert Feldt, Richard Torkar, Lefteris Angelis, and Maria Samuelsson. 2008. Towards individualized software engineering: empirical studies should collect psychometrics. In *Proceedings of the 2008 international workshop on Cooperative and human aspects of software engineering*. 49–52.

[16] Leah Findlater and Krzysztof Z Gajos. 2009. Design space and evaluation challenges of adaptive graphical user interfaces. *AI Magazine* 30, 4 (2009), 68–68.

[17] International Organization for Standardization. 2010. *Ergonomics of Human-system Interaction: Part 210: Human-centred Design for Interactive Systems*. ISO.

[18] Frederick G Freeman, Peter J Mikulka, Lawrence J Prinzel, and Mark W Scerbo. 1999. Evaluation of an adaptive automation system using three EEG indices with a visual tracking task. *Biological Psychology* 50, 1 (1999), 61–76. https://doi.org/10.1016/S0301-0511(99)00002-2

[19] Thomas Fritz, Andrew Begel, Sebastian C Müller, Serap Yigit-Elliott, and Manuela Züger. 2014. Using psycho-physiological measures to assess task difficulty in software development. In *Proceedings of the 36th international conference on software engineering*. 402–413.

[20] Thomas Fritz and Sebastian C Müller. 2016. Leveraging biometric data to boost software developer productivity. In *2016 IEEE 23rd international conference on software analysis, evolution, and reengineering (SANER)*, Vol. 5. IEEE, 66–77.

[21] Krzysztof Z Gajos and Krysta Chauncey. 2017. The influence of personality traits and cognitive load on the use of adaptive user interfaces. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 301–306.

[22] Krzysztof Z Gajos, Mary Czerwinski, Desney S Tan, and Daniel S Weld. 2006. Exploring the design space for adaptive graphical user interfaces. In *Proceedings of the working conference on Advanced visual interfaces*. 201–208.

[23] Vanessa Georges, François Courtemanche, Sylvain Senecal, Thierry Baccino, Marc Fredette, and Pierre-Majorique Leger. 2016. UX Heatmaps: Mapping User Experience on Visual Interfaces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4850–4860. https://doi.org/10.1145/2858036.2858271

[24] Francesca Gullà, Silvia Ceccacci, Michele Germani, and Lorenzo Cavalieri. 2015. Design adaptable and adaptive user interfaces: a method to manage the information. In *Ambient Assisted Living*. Springer, 47–58.

[25] Eddie Harmon-Jones, Philip A. Gable, and Carly K. Peterson. 2010. The role of asymmetric frontal cortical activity in emotion-related phenomena: A review and update. *Biological Psychology* 84, 3 (2010), 451–462. https://doi.org/10.1016/j.biopsycho.2009.08.010 The biopsychology of emotion: Current theoretical and empirical perspectives.

[26] Xiyuan Hou, Fitri Trapsilawati, Yisi Liu, Olga Sourina, Chun-Hsien Chen, Wolfgang Mueller-Wittig, and Wei Tech Ang. 2017. EEG-based human factors evaluation of conflict resolution aid and tactile user interface in future air traffic control systems. In *Advances in Human Aspects of Transportation*. Springer, 885–897.

[27] Nuraini Jamil, Abdelkader Nasreddine Belkacem, Sofia Ouhbi, and Abderrahmane Lakas. 2021. Noninvasive Electroencephalography Equipment for Assistive, Adaptive, and Rehabilitative Brain–Computer Interfaces: A Systematic Literature Review. *Sensors* 21, 14 (2021), 4754.

[28] MAXWELL K.D. 2002. Applied Statistics for Software Managers. *Applied Statistics for Software Managers* (2002). https://cir.nii.ac.jp/crid/1573668924056253312

[29] Barbara Kitchenham, Lech Madeyski, David Budgen, Jacky Keung, Pearl Brereton, Stuart Charters, Shirley Gibbs, and Amnart Pohthong. 2017. Robust statistical methods for empirical software engineering. *Empirical Software Engineering* 22, 2 (2017), 579–630.

[30] Wolfgang Klimesch. 1999. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews* 29, 2 (1999), 169–195. https://doi.org/10.1016/S0165-0173(98)00056-3

[31] W. KLIMESCH, M. DOPPELMAYR, H. SCHIMKE, and B. RIPPER. 1997. Theta synchronization and alpha desynchronization in a memory task. *Psychophysiology* 34, 2 (1997), 169–176. https://doi.org/10.1111/j.1469-8986.1997.tb02128.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.1997.tb02128.x

[32] Wolfgang Klimesch, Bärbel Schack, and Paul Sauseng. 2005. The functional significance of theta and upper alpha oscillations. *Experimental psychology* 52, 2 (2005), 99–108.

[33] Makrina Viola Kosti, Kostas Georgiadis, Dimitrios A Adamos, Nikos Laskaris, Diomidis Spinellis, and Lefteris Angelis. 2018. Towards an affordable brain computer interface for the assessment of programmers' mental workload. *International Journal of Human-Computer Studies* 115 (2018), 52–66.

[34] Helena Chmura Kraemer and David J. Kupfer. 2006. Size of Treatment Effects and Their Importance to Clinical Research and Practice. *Biological Psychiatry* 59, 11 (2006), 990–996. https://doi.org/10.1016/j.biopsych.2005.09.014

[35] Sylvia D. Kreibig. 2010. Autonomic nervous system activity in emotion: A review. *Biological Psychology* 84, 3 (2010), 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010 The biopsychology of emotion: Current theoretical and empirical perspectives.

[36] Sari Kujala, Virpi Roto, Kaisa Väänänen-Vainio-Mattila, Evangelos Karapanos, and Arto Sinnelä. 2011. UX Curve: A method for evaluating long-term user experience. *Interacting with computers* 23, 5 (2011), 473–483.

[37] Effie Lai-Chong Law, Paul van Schaik, and Virpi Roto. 2014. Attitudes towards user experience (UX) measurement. *International Journal of Human-Computer Studies* 72, 6 (2014), 526–541. https://doi.org/10.1016/j.ijhcs.2013.09.006 Interplay between User Experience Evaluation and System Development.

[38] Haeinn Lee, Jungtae Lee, and Ssanghee Seo. 2009. Brain response to good and bad design. In *International Conference on Human-Computer Interaction*. Springer, 111–120.

[39] Seolhwa Lee, Danial Hooshyar, Hyesung Ji, Kichun Nam, and Heuiseok Lim. 2018. Mining biometric data to predict programmer expertise and task difficulty. *Cluster Computing* 21, 1 (2018), 1097–1107.

[40] Nicole M. Long, John F. Burke, and Michael J. Kahana. 2014. Subsequent memory effect in intracranial and scalp EEG. *NeuroImage* 84 (2014), 488–494. https://doi.org/10.1016/j.neuroimage.2013.08.052

[41] Julio Medeiros, Ricardo Couceiro, João Castelhano, M Castelo Branco, Gonçalo Duarte, Catarina Duarte, João Durães, Henrique Madeira, P Carvalho, and C Teixeira. 2019. Software code complexity assessment using EEG features. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1413–1416.

[42] Peter J. Mikulka, Mark W. Scerbo, and Frederick G. Freeman. 2002. Effects of a Biocybernetic System on Vigilance Performance. *Human Factors* 44, 4 (2002), 654–664. https://doi.org/10.1518/0018720024496944 arXiv:https://doi.org/10.1518/0018720024496944 PMID: 12691372.

[43] Jefferson Seide Molléri, Indira Nurdiani, Farnaz Fotrousi, and Kai Petersen. 2019. Experiences of studying Attention through EEG in the Context of Review Tasks. In *Proceedings of the Evaluation and Assessment on Software Engineering*. 313–318.

[44] Sebastian C Müller and Thomas Fritz. 2015. Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, Vol. 1. IEEE, 688–699.

[45] Anilkumar A. C. Nayak, C. S. 2019. Eeg normal waveforms. (2019).

[46] Jakob Nielsen and Jonathan Levy. 1994. Measuring Usability: Preference vs. Performance. *Commun. ACM* 37, 4 (apr 1994), 66–75. https://doi.org/10.1145/

175276.175282

[47] Alan T Pope, Edward H Bogart, and Debbie S Bartolome. 1995. Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology* 40, 1 (1995), 187–195. https://doi.org/10.1016/0301-0511(95)05116-3 EEG in Basic and Applied Settings.

[48] Peter Schmutz, Silvia Heinz, Yolanda Métrailler, and Klaus Opwis. 2009. Cognitive load in eCommerce applications—measurement and effects on user satisfaction. *Advances in Human-Computer Interaction* 2009 (2009).

[49] Janet Siegmund, Christian Kästner, Sven Apel, Chris Parnin, Anja Bethmann, Thomas Leich, Gunter Saake, and André Brechmann. 2014. Understanding understanding source code with functional magnetic resonance imaging. In *Proceedings of the 36th international conference on software engineering*. 378–389.

[50] Bitbrain Technologies. 2020. Diadem, Bitbrain's EEG Device. https://www.bitbrain.com/neurotechnology-products/dry-eeg/diadem

[51] Alexandre N. Tuch, Paul Van Schaik, and Kasper Hornbæk. 2016. Leisure and Work, Good and Bad: The Role of Activity Domain and Valence in Modeling User Experience. *ACM Trans. Comput.-Hum. Interact.* 23, 6, Article 35 (dec 2016), 32 pages. https://doi.org/10.1145/2994147

[52] Jean Vanderdonckt, Sara Bouzit, Gaëlle Calvary, and Denis Chêne. 2019. Exploring a design space of graphical adaptive menus: normal vs. small screens. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 1 (2019), 1–40.

[53] Fischer T. Riedl R. Weber, B. 2021. Brain and autonomic nervous system activity measurement in software engineering: A systematic literature review. *Journal of Systems and Software* 178 (2021), 110946.

[54] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Analysis and Interpretation*. Springer Berlin Heidelberg, Berlin, Heidelberg, 123–151. https://doi.org/10.1007/978-3-642-29044-2_10

[55] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Empirical Strategies*. Springer Berlin Heidelberg, Berlin, Heidelberg, 9–36. https://doi.org/10.1007/978-3-642-29044-2_2

[56] Tarannum Zaki and Muhammad Nazrul Islam. 2021. Neurological and physiological measures to evaluate the usability and user-experience (UX) of information systems: A systematic literature review. *Computer Science Review* 40 (2021), 100375. https://doi.org/10.1016/j.cosrev.2021.100375