User Experience with Adaptive User Interfaces: Comparing Performance and Preferences

Daniel Gaspar-Figueiredo[a,b,*], Jean Vanderdonckt[c], Silvia Abrahao[a], Emilio Insfran[a]

[a] *Universitat Politècnica de València, Valencia, Spain*
[b] *Instituto Tecnológico de Informática, Valencia, Spain*
[c] *Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium*

Abstract

Adaptive user interfaces dynamically change their content, presentation, and behavior to optimize the user experience, which has been primarily evaluated using classic usability measures but to a lesser extent by using neurological measures. While the perceived preference of specific user interface elements, such as graphical adaptive menus, has already been studied, no consensus exists regarding their performance and how to substitute a static menu with an adaptive one. To gain insights into how graphical adaptive menus could influence the user experience and to identify any correlation between users' performance and their preferences, we conducted an experiment in which forty participants used twenty graphical adaptive menus while their brain activity was captured by employing electroencephalography to derive four measures (*i.e.*, cognitive load, engagement, attraction, and memorization). User performance was measured using task completion time, specifically the time to select menu items. Statistical analysis suggested which graphical adaptive menus were significantly better or worse than the static menu, our baseline. These results are used as the basis to suggest implications for software developers and researchers to design more effective adaptive user interfaces.

Keywords: Adaptive Systems, User Interfaces, User Experience, Electroencephalogram, Experiment

1. Introduction

Adapting user interfaces (UI) is an essential practice in the modern development of interactive applications, to create intuitive, efficient, and user-friendly experiences [1]. *Adaptive User Interfaces* (AUIs) [2] have been introduced as a response to users' changing needs over time and to preserve the overall user experience (UX) depending on these changes. AUIs adjust the contents (*e.g.*, text, images), presentation (*e.g.*, layout of elements and their selection), and behavior (*e.g.*, navigation) to user's preferences and abilities to provide a more personalized experience. AUIs can be classified into two categories: *rule-based* and *data-driven*. While the former relies on a set of rules and meta-rules to determine how the interface should be adapted, the latter uses Machine Learning (ML) algorithms to predict these adaptations on the basis of user's data [1].

*Corresponding author
Email address: dagasfi@epsa.upv.es (Daniel Gaspar-Figueiredo)

AUIs effectively impact objective measures, such as task completion time and error rate, and subjective measures, such as user satisfaction [2]. However, the potential advantages of AUIs are mitigated by several factors, such as personality traits and cognitive load [3]. There are also some potential drawbacks [4]. For instance, rule-based adaptive systems may be inflexible and unable to account for all possible user scenarios, and data-driven approaches can be resource-intensive and require a large amount of user data to train the algorithms.

Although considerable progress has been made as regards how to develop adaptive systems in general and AUI in particular, there is a need to better understand how people interact with these systems in order to inform the design of future software systems. While traditional methods such as surveys, interviews and preference rankings can provide valuable insights into user preferences, they do not always capture the full UX spectrum, as opposed to neurological data, such as electroencephalography (EEG), which provide a more objective and insightful way in which to measure UX, and allow a more comprehensive understanding of how users interact with AUIs. Unlike traditional questionnaire-based methods, which rely on subjective self-reporting and are susceptible to biases such as social desirability or memory recall errors, EEG data provide objective, real-time insights into users' cognitive and emotional states. By directly measuring brain activity, EEG captures subtle and dynamic responses to adaptive user interfaces that are often overlooked in self-reported measures. This reliability makes EEG a valuable tool for understanding the user experience in a quantifiable and reproducible manner. Among various biometric methods, such as galvanic skin response or eye-tracking, EEG stands out for its ability to measure multiple dimensions of UX, including cognitive load, engagement, attraction, and memorization, all of which are used for evaluating the impact of adaptive user interfaces. EEG provides higher temporal resolution compared to other modalities, allowing us to capture rapid changes in mental states as users interact with adaptive menus.

The comparison of the performance measures obtained from EEG data, such as cognitive load, with user preference, makes it possible to gain further insights into how adaptive user interfaces affect UX. However, the debate regarding performance *vs.* preference [5] is far from being resolved. This refers to the question of whether it is more important to develop software systems that users prefer, even if they do not perform as well, or systems that perform well in terms of certain measures (*e.g.*, task completion time and rate), even if they are not very much appreciated by end users. Indeed, an end user may prefer a design option for the interface that is under-performing. In this case, preference takes precedence over performance: as long as users are satisfied, they do not consider the performance to be so important. Furthermore, end users may also wish to be efficient when using the software, even if the design choices do not match their preferences. Indeed, preferences vary widely from one user to another and taking them all into account is very complicated. The question of preference *vs.* performance, therefore, remains open.

Since studying the effect of a wide range of UI elements on UX may involve significant challenges, in this paper, we focus on a specific UI element: graphical adaptive menus [6]. Many variations of these menus have been proposed and tested, often against the static menu used as a baseline, and against a few other adaptive menus. Studies often report the advantages [4] and limits [7] of an adaptive menu when compared to a few others, signifying that no overall conclusion that is valid for all adaptive menus can be inferred. Jointly testing so many different adaptive menus is impossible because it represents a prohibitively expensive development cost.

In order to address this problem, we conducted an experiment with which to compare the user experience and performance of forty participants when selecting adaptive menu items while their brain activity was captured by means of electroencephalography with the preference to use these adaptive menus as expressed by a large pool of participants. User experience, which was obtained through the use of an EEG device, was measured in terms of four measures (cognitive load, engagement, attraction and memorization), while performance was measured in terms of completion time, specifically the time taken to select items in the menus. The experiment was conducted with a diverse group of users with different backgrounds (*e.g.*, software engineers, mechanical engineers, financial experts, and healthcare professionals) and a wide range of ages. Our goal was to understand how these menus differ in terms of UX and performance and to identify the strengths and weaknesses of each type of menu. The research questions used as a basis were:

RQ_1 =*Do different designs for graphical adaptive menus have different effects on EEG measures for UX?*
RQ_2 =*Do users' preferences for graphical adaptive menu designs impact task performance and EEG measures for UX?*

This paper builds upon a conference contribution by the same authors [8], in which a differentiated replication study was recently carried out to provide further evidence for RQ1 (*i.e.*, whether graphical adaptive menus have a different influence on UX) with a more homogeneous group of users (see Section 3 for more details). This allowed us to understand whether human factors (*e.g.*, gender, age, background) affect user experience with different menu types. In particular, this work presents the results obtained for the first research question (RQ1) for a diverse group of users and compares these results with those obtained in the replication study. Moreover, we contribute to the ongoing debate regarding preference *vs.* performance in UI design (RQ2) by providing empirical data on the UX of these adaptive menus and by highlighting any trade-off that may exist between UX/performance and preference. Understanding the influence of graphical adaptive menus on UX and the relationship between their performance and the preference for them will make it possible to better inform design decisions concerning adaptive UIs.

The results of our experiment have the potential to inform the design of future graphical adaptive menus, particularly in terms of usability and UX. By understanding which menu type is more effective in terms of cognitive load, engagement, attraction, and memorization, software application developers can make informed decisions about which menu type to use when building their systems. We additionally wished to gather empirical evidence that could be used by adaptive systems to make better decisions on which elements to adapt and why to make certain adaptations.

This paper is organized as follows. Section 2 provides the background regarding graphical adaptive user interfaces, UX, and EEG, while Section 3 introduces the rationale behind the research questions by discussing existing empirical studies on the UX of adaptive user interfaces and the use of neurological data in Software Engineering (SE) and Human Computer Interaction (HCI). Section 4 then presents the design and execution of the empirical study, and Section 5 presents the results obtained and discusses the findings. We also discuss the implications of the results for user interface design. Section 7 discusses the threats that could affect the validity of the results obtained, and finally, Section 8 provides a summary of results and presents future avenues for this work.

2. Background

The development of AUIs has attracted a great deal of attention, as they provide a more personalized and efficient UX. Aspects of user experience can be measured by analyzing neurological measures that represent different human affective states (e.g. pleasure, stress, relaxation) when performing a particular task. This kind of measure has received increasing attention from the SE community in the last few years. This section introduces the graphical adaptive user interfaces used as a subject in this experiment, how UX can be measured, and how brain activity can be measured and interpreted in terms of UX.

2.1. Graphical Adaptive User Interfaces

Adaptive user interfaces are adaptive systems that are capable of adjusting the display, thus adapting the organization or presentation of the UI functionality in response to certain characteristics of the user's context. Various strategies with which 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 Graphical User Interface (GUI).

One important aspect of a GUI is its menu system, which provides users with a list of options and actions that they can take. In a recent study [9], a set of 49 graphical adaptive menus was collected which provided a new exploration of a UI design space on the basis of Bertin's eight visual variables (i.e., position, size, shape, value, color, orientation, texture, and motion) [10]. This design space allowed the authors to propose 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 to improve users' efficiency and satisfaction when using menus. However, experiments with which to assess how different design choices affect the UX are required.

The importance of graphical adaptive menus lies in their potential to address challenges in human-computer interaction, such as reducing cognitive load, enhancing task completion efficiency, and increasing user satisfaction. Traditional static menus may overwhelm users with too many options, making it difficult to find the desired item quickly. By leveraging adaptive mechanisms, these menus dynamically prioritize and present the most relevant options, allowing users to focus on the task at hand without unnecessary distractions. Additionally, exploring the design space of adaptive menus contributes to understanding how visual variables such as motion, color, and shape impact UX.

In this paper, we present an experiment that compares a sub-set of twenty graphical adaptive menus shown in Fig. 1. The criteria used to select these menus are discussed in Section 4.3. In particular, we selected 19 menu types proposed in [9]. For example, the *Underlying* menu is a value-changing menu (i.e., a menu that changes the font of the prediction window, underlining the menu items in order to highlight them), while a *Rotating* menu is a motion-changing menu (i.e., a menu that moves menu items to highlight them) and the *Leaf* menu is a menu that contains unusual shapes. As a baseline, we also selected the *Static* menu, which does not use the prediction window and does not highlight any specific menu items.

2.2. User Experience Measurement

User experience (UX) is defined in ISO 9241-210 [11] as *a person's perceptions and responses that result from the use and/or anticipated use of a product, system, or service*. This

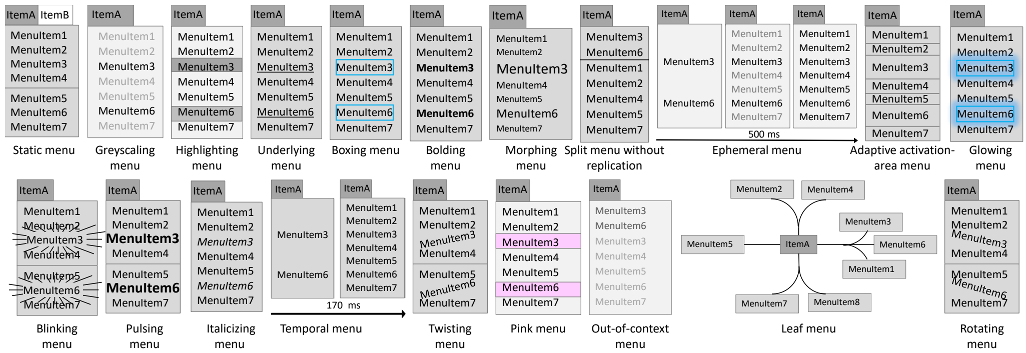

Figure 1: The twenty graphical adaptive menus used in the experiment (extracted from [6]).

includes all interface effects on the user, before, during and after use. UX evaluation methods are divided into *subjective* methods, such as interviews and surveys, and *objective* methods, such as neurological and physiological measures [12]. Subjective methods involve obtaining data concerning a user's personal opinions and assumptions, whereas objective methods involve obtaining data through observations and measurements. Each approach has its limitations. Subjective methods can provide biased information, as responses to questionnaires may be influenced by how users remember their experience rather than the experience itself [13], or by the Hawthorne effect [14], where individuals may alter their behavior due to their awareness of being observed. Similarly, objective methods have limitations, such as being potentially intrusive or failing to capture the specific aspects of user experience that subjective reports can provide. However, objective methods might be preferred in certain contexts because they can offer more consistent and quantifiable data, which can reduce the impact of personal biases and recall inaccuracies. For example, biometric data such as neurological measures, eye tracking, among others, can provide real-time data on cognitive load and emotional responses, offering insights that might not be accessible through self-reporting.

The challenges related to measuring UX led us to consider alternative methods, such as relying on biometric data and physiological reactions like blinking, sweating, eye movement, or brain activity [15]. UX can be effectively and efficiently measured using neurological and physiological measures [12]. For example, the responses of the Autonomic Nervous System (ANS) reveal multiple emotions, which are classified as negative (*e.g.*, anger, anxiety, disgust, embarrassment, fear, sadness) or positive (*e.g.*, affection, amusement, happiness, joy, pleasure, pride, relief) [16]. Electroencephalography studies have shown that different areas of the brain reveal unique UX measures. For example, frontal brain activity reveals emotion [17], and the differences between the left and right-hand sides are associated with positive emotions such as enthusiasm, or negative emotions such as anger [18]. Furthermore, EEG also provides objective measures of the users' cognitive load, attraction, engagement, and memorization. EEG frequency bands, such as alpha and theta waves linked to cognitive and memory performance [19] and beta bandwidth indicating user engagement ([20, 21]), are detailed in Section 2.3.1. The forehead brain activity, which takes place over the prefrontal cortex, is associated with various human mental states, such as attention, emotion, and cognition ([22, 23]).

The arousal-valence model of emotions, often used alongside physiological measures, categorizes emotions based on their arousal (activation) and valence (pleasantness) [24]. This model is particularly useful in understanding attractiveness, as it helps to delineate emotions in a two-dimensional space, where high arousal and positive valence indicate excitement or attractiveness, and low arousal and negative valence indicate boredom or displeasure. EEG studies can, therefore, be considered a useful and practical method for UX measurement.

2.3. Measurement of Brain Activity using EEG

The communication between cells in the brain, which are known as *neurons*, is achieved by the transfer of nerve impulses. These impulses change the electrical charge of the neurons and, when many neurons send impulses in synchronization, they produce an electrical field that can be captured using an electroencephalogram device. EEG is a technique that records changes in electrical activity, also known as *brainwaves*, generated by the neurons, by using electrodes placed on the scalp. The electrical activity measured at the electrodes is very small, in the order of microvolts (μV), thus requiring amplification that is displayed as a sequence of voltage values over time on a graph. The analysis of brainwaves using EEG has found applications in diverse fields, such as psychology, medicine, software engineering, and human-computer interaction. In particular, EEG has been explored in UX research, in which it provides insights into users' cognitive and emotional states during a particular task [25]. The measurement of brainwave patterns, allows researchers to obtain information about users' levels of engagement, workload, frustration, and other psychological states that are relevant for evaluating the UX, in particular for distinguishing a bad UI from a good UI [26]. EEG data are analyzed using signal processing techniques, such as spectral analysis, event-related potentials, and time-frequency analysis, to identify brainwave patterns associated with different cognitive and emotional processes.

2.3.1. EEG Frequencies

Brainwaves captured by EEG can be classified on the basis of multiple factors such as location, amplitude, frequency, morphology, continuity (*i.e.*, rhythmic, intermittent, or continuous), synchrony, symmetry, and reactivity. The method most commonly used to classify EEG brainwaves is frequency. Greek letters are used to refer to the different frequencies of each EEG wave, which are classified into different frequency bands, associated with mental states and activities. The frequency bands commonly analyzed in EEG are [27]:

- *Delta* band (δ: 0.5 to 4 Hz or less than 4 Hz): is mainly associated with deep sleep moments and unconsciousness, which may indicate that the participant becomes tired or less engaged. This band impacts the dependent variable Engagement (see Section 4.4.2) reported in our study: the higher the value, the less engagement.

- *Theta* band (θ: 4 to 8 Hz): is mainly associated with light sleep, relaxation, and drowsiness, which may represent a thinking period, not acting. This band impacts the variable Attention (see Section 4.4.2) reported in the study: the higher the value, the higher the cognitive strain, potentially indicating that sustained attention is ongoing.

- *Alpha* (α: 8 to 12 Hz): is mainly associated with wakeful relaxation, calmness, or mental load, which may represent a positive sign of good usability. This band impacts the variable Cognitive Load (see Section 4.4.2) reported in the study: the higher the value, the lower the cognitive load.

- *Beta* (β: 13 to 30 Hz): is mainly associated with active thinking, concentration, and focused attention, which may indicate an engagement in a complex task. This band impacts the variable Engagement reported in the study: the higher the value, the higher engagement in complex tasks. An excessively high value of this variable suggests stress or frustration in user experience.

- *Gamma* (γ: 30 to 100 Hz): is associated with sensory processing, attention [28], and high-level cognitive functions such as perception, memory, and problem-solving, which may represent an involvement in a learning activity or an awareness. This band impacts the variables Memorization (see Section 4.4.2) and Cognitive Load reported in the study: the higher the value, the higher the memorization or the more important the cognitive load.

In sum, beta and gamma bands directly impact the Cognitive Load, whereas alpha and theta impact the Engagement and the Attention. User's subjective experience is positively impacted by the relative power of Gamma band, but negatively by Beta and Theta bands, according to [25].

Tasks and mental states correspond to various activity patterns in specific frequency bands that are traced by employing bandwidth filtering of EEG recordings. The changes in potential in brainwaves measure specific brain activities correlated with multiple functions, such as perception and movement, along with cognitive processes related to attention, learning, and memory [29]. Brainwave oscillations reflect what is happening in the participant's information processing situation, even if the participant is unaware of changes or is unable to verbalize them [29, 30]. Cognitive load, concentration, excitement, relaxation, and several other features of the user's state are, therefore, measured by detecting changes and oscillations in brainwaves.

2.3.2. EEG Data Acquisition

Electrical activity in the brain is measured using electrodes attached to the scalp by means of a cap or headset. These electrodes detect the electrical signals generated by neurons in the brain. The number and placement of electrodes used vary depending on the specific research. There is a standard for electrode placement (10-10 and 10-20 systems) [31], which involves measuring the distance between specific points on the scalp and using this information to place electrodes accurately. Electrodes can be dry, saline, or gel-based, and each method has its advantages and disadvantages. The design of the cap, such as its size and shape, also varies according to the application [32].

In this study, we have used the BitBrain Diadem[1] headset which captures a raw EEG signal with a resolution of 24 bits at 256 Hz. This portable non-intrusive dry EEG headset is used in neuroscience research (*e.g.*, emotion recognition [33], neuroprosthese [34]), Brain-Computer Interfaces (BCIs) [35] , motor imagery and device control [36]), and HCI [37] (*e.g.*, for measuring the cognitive load [38] of mobile tasks =[39], tactile interaction [40], detection of driver's distraction [41]). This device has twelve dry electrodes (see Fig. 2) placed at positions AF7, Fp1, Fp2, AF8, F3, F4, P3, P4, PO7, O1, O2, and PO8 according to the international 10-10 standard [31] and uses the left earlobe (A1, REF) as a reference electrode and the center of the head (Fpz, GND) as the ground. EEG data is captured using an amplifier, which increases and digitizes the electrical signals from the electrodes. The specifications of the amplifier also vary, including the sampling rate, bandwidth, and resolution. The sampling rate determines how frequently the electrical signals are sampled, with higher sampling rates providing more detailed information about the electrical activity in the brain. The bandwidth determines the range of frequencies that can be captured, while the resolution determines the smallest signal change that can be detected. The EEG data captured are then analyzed to extract features that provide insight into the user's cognitive load, emotion, etc.

[1] https://www.bitbrain.com/neurotechnology-products/dry-eeg/diadem

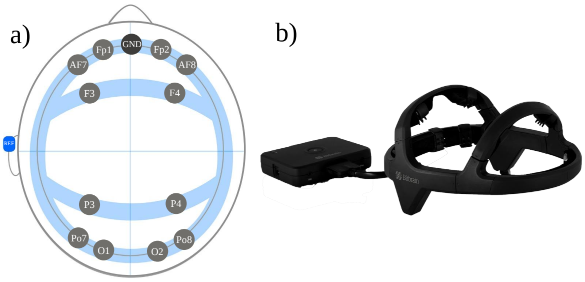

Figure 2: The BitBrain Diadem: a) the location map of the 12 electrodes based on the international 10-10 system [31], including REF and GND; b) the dry headset and amplifier records an EEG signal at a rate of 256 Hz.

2.3.3. EEG Data Interpretation

Once EEG data have been acquired, an adequate interpretation is required in order to draw meaningful conclusions about the users' emotions and feelings. Each individual has different brain responses in relaxed or stressed situations, which are addressed through the collection of a *calibration sample*. This sample serves as a baseline and is obtained when individuals are in a relaxed state with their eyes closed, thus allowing accurate comparisons to be made with subsequent measurements. By taking individual differences into account during interpretation, EEG data provide valuable insights into cognitive processes and mental states which reflect the UX of different tasks.

3. Related Work

This section discusses previous work on adaptive user interfaces and UX, and the use of neurological measures in Software Engineering (SE) and Human-Computer Interaction (HCI) as a means to assess UX. By examining the use of neurological signals in SE and UX research, we aim to provide a comprehensive overview of the state of the art and highlight the potential for using neurological data to gain greater insights into the UX of AUIs.

3.1. Empirical Studies on the Performance and UX of Adaptive User Interfaces

Many studies have assessed the influence of AUIs on UX regarding various factors, such as subjective preference, efficiency, and task completion time [1, 42]. For example, [43] compared 3 AUIs with a static, non-adaptive control condition. They found that the effect of AUIs on users varied according to their specific features, and that AUIs that duplicated frequently (rather than relocating) tended to result in a better UX and subjective satisfaction.

Studying a complete AUI can be challenging owing to the complex interplay of various components and the combination of their adaptation capabilities, either taken in isolation or combined together. As a result, several studies have focused on a specific AUI element, such as graphical adaptive menus considered as frequent elements [44], in order to precisely delineate the borders of the impact on UX.

In their work, Bouzit et al. [45] compared two graphical adaptive menus for smartphones so as to improve their hierarchical navigation, in terms of item selection time and error rate. Although this study provides insights into performance, it lacks a comprehensive analysis of user preferences and the results may not be applicable to other menus. More recently, Vanderdonckt et al. [6] structured a design space according to Bertin's eight visual variables (*i.e.*, position, size, shape, value, color, orientation, texture, and motion) [10], which enabled a systematical exploration of how to combine the variables. The analysis show that participants expressed their preference for graphical adaptive menus: there was a high preference for those menus that preserved spatial stability (position and orientation) and tolerated value-changing capabilities, and a low preference for those menus that changed color or motion or had unusual shapes. While this study provides valuable insights into participants' perceived preferences as regards a large number of adaptive menus, it does not consider UX or the performance achieved by the participants.

In another study, Park et al. [46] showed that an adaptable menu outperformed two adaptive menus and the static menu as regards performance and user satisfaction. The adaptive split menu confronted challenges when the frequency of selection changed, while the adaptive highlight menu, which performed similarly to the static menu, was preferred owing to its ability to help select frequent items. This study considered only item selection time and the number of errors. Moreover, Findlater and McGrenere [7] revealed that the static menu was the fastest as regards efficiency, followed by the adaptable menu, while the adaptive menu was the slowest, although by very little. Despite being less efficient, the majority of the participants preferred the adaptable menu to the other two menu types. In another work, Findlater et al. [47] found that two AUIs consistently led to positive outcomes in performance and user satisfaction. Finally, Jiang and Chen [48] evaluated the navigation efficiency of three menu types (*i.e.*, full, collapsible, cascading) for mobile devices. With regard to large menu breadths, collapsible menus were more efficient on smartphones, while cascading menus were more effective for smartwatches. The number of preview items significantly impacts on navigation efficiency, particularly in the case of smartwatches. While the authors considered both preference and performance for the first time, the performance was evaluated using a questionnaire, which is considered a static, self-reported evaluation method.

The aforementioned studies suggest that AUIs, and particularly, graphical adaptive menus, contribute to UX by providing users with an interface that adapts to their requirements as regards their context of use. While these studies assessed performance and/or preference via self-report measures [48], task-based evaluations [49], and A/B testing [9], none of them considered neurological data in order to evaluate the UX.

3.2. Neurological and physiological measures in SE and HCI

The use of neurological and physiological measures is a growing field of research in software engineering. Neurological data obtained using EEG and functional Magnetic Resonance Imaging (fMRI), along with physiological data obtained by means of eye tracking and Galvanic Skin Response (GSR), effectively measure aspects of UX during software engineering tasks such as code comprehension, code readability, and error detection. For example, Siegmund et al. [50] used fMRI to determine the cognitive load registered by code comprehension tasks and locate syntax errors with which to contrast them. They found a different activation pattern in five regions of the brain related to working memory, attention, and language processing. In another study, Fakhoury et al. [51] used fNIRS (functional Near-Infrared Spectroscopy) combined with eye-tracking to explore the effect of poor source code lexicon and readability on developers' cognitive load. The authors concluded that the presence of linguistic anti-patterns in source code significantly increases the developers' cognitive load.

A recent systematic literature review on the use of neurological and physiological measures in SE [52] revealed that several empirical studies have been conducted in order to measure the cognitive load in SE activities, such as: code comprehension, code inspection, programming, and bug fixing. However, only a few have reported experiments with which to capture end-user emotions and feelings. Empirical SE studies should focus on both the developer's perception and emotions during the construction of the system and the end user's perceived use [53]. However, this perceived use is typically assessed using retrospective methods (e.g., interviews and questionnaires), thus increasing the bias of self-report [54] and the Hawthorne effect [14]. For example, questionnaires with scales decomposed into sub-scales and items [55], sometimes with a level of importance, assess user satisfaction, ease of learning, or compliance with expectations, but these continue to be self-reported measures.

Moreover, cognitive load is traditionally measured using the empirically validated NASA-TLX questionnaire [56] to correlate it with user satisfaction [57], but the participant's answers are subject to inaccurate user recall or other subjective factors, thus leading to imprecise measurements ([13, 58]). Since these methods do not capture the UX during the real interaction, neurological and physiological data can be used to effectively measure UX in real-time. Although these measures are complex to interpret, they are objective [15], thus explaining why the SE field is paying more attention to them [59].

In the HCI field, fMRI and EEG are being used to understand how the brain reacts when users compare alternative designs [60]: the perception of different feelings toward alternative designs is associated with the frontal and occipital lobes, and the human brain responds sooner and stronger as regards its perception of bad feelings. The brain states of air traffic controllers have also been monitored using EEG [61] to discover a high correlation between the cognitive load calculated using the NASA-TLX and that computed by means of EEG data. Furthermore, the behavioral, subjective and neural responses of people using two smartphones to complete three tasks were compared on the basis of EEG signals, and the conclusion was reached that the higher the UX was, the stronger the activity that was recorded in α, δ, and γ, but the weaker the activity in β and θ [62].

The aforementioned experiments focused mostly on detecting differences between brainwave frequencies and cognitive load. While the detection of emotion could also be used to assess UX in real-time, it is recognized as being less representative [61]. In this study, we have, therefore, chosen EEG in order to measure not only the end users' cognitive load but also their memorization effort, attraction, and the engagement produced by the interaction with different adaptive user interfaces.

To provide a structured overview of the existing studies, we summarize their key contributions and limitations concerning the research questions. Table 1 presents an evaluation matrix that maps the studies to relevant quality criteria: Cognitive Load (CL), User Satisfaction (US), Biometric Data (BD), Adaptive Design (AD), Performance Metrics (PM), and Real-Time Analysis (RT).

As illustrated in the table, no existing study combines performance metrics with real-time analysis, indicating a significant gap in the integration of these aspects. Additionally, only one study [48] addresses cognitive load, performance metrics, and user satisfaction simultaneously. However, this study does not incorporate biometric data or real-time analysis, further underscoring the fragmented nature of prior research. These observations highlight the necessity of a holistic approach that integrates real-time, biometric, and performance-based evaluations to comprehensively assess the impact of adaptive user interfaces on user experience. This work seeks to address these gaps and advance the state of the art in this domain.

Table 1: Evaluation Matrix for Related Work

Study	CL	US	BD	AD	PM	RT
[43]	✗	✓	✗	✓	✓	✗
[45]	✗	✗	✗	✓	✓	✗
[46]	✗	✓	✗	✓	✓	✗
[7]	✗	✓	✗	✓	✓	✗
[47]	✗	✓	✗	✓	✓	✗
[48]	✓	✓	✗	✓	✓	✗
[50]	✓	✗	✓	✗	✗	✗
[51]	✓	✗	✓	✗	✗	✗
[61]	✓	✗	✓	✗	✗	✓
[62]	✓	✓	✓	✗	✗	✓

4. Empirical study

This section presents the empirical study, which was structured according to Ko *et al.*'s [63] recommendations: the research questions to be addressed and their related hypotheses to be tested, the context and the subject selection, the experimental objects and stimuli, and the independent and dependent variables, the instrumentation, the experimental procedure, and how data analysis is performed.

4.1. Research Questions

The goal of this study is, according to the Goal-Question-Metric (GQM) framework [64], to **analyze** a set of graphical adaptive menus **for the purpose of** comparing them with respect to the user experience produced **from the point of view of** both researchers and software designers **in the context of** an end-user group at the *Universitat Politècnica de València* and their contacts.

The objective of the study was, therefore, to address the following research questions (RQ):

- RQ_1: Do different designs for graphical adaptive menus have different effects on EEG measures for UX?
- RQ_2: Do users' preferences for graphical adaptive menu designs impact task performance and EEG measures for UX?

4.2. Context, Participant Recruitment and Selection

The experimental study was conducted in June-July 2022 with 40 volunteers (20 female, 20 male, and 0 identified as gender variant/non-conforming) with diverse backgrounds (*e.g.*, software engineers, mechanical engineers, financial experts, and healthcare professionals). They were aged from 18 to 64 (M=39.55, SD=15.26, Mdn=41). The volunteers were recruited by means of an open call for participation at the *Universitat Politècnica de València* using a snowball procedure.

We sent an announcement concerning the study to an email list and also placed flyers in buildings frequented by students and administrative staff. In order to ensure the voluntary nature of participation, we emphasized that the participants could leave the experiment at any time. No compensation was offered. A pre-survey questionnaire was employed in order to verify that each participant had some experience in using interactive mobile and desktop applications (at least two years) and did not suffer from any disability. Once the participants had been selected, we gave them the informed consent document.

The study protocol was approved by the Institutional Review Board (IRB) of the *Universitat Politècnica de València* to ensure ethical compliance. All participants provided informed consent before participating in the experiment. The consent form emphasized that no personal data would be collected during the study. In case any data was collected, it would be anonymized, not published, and securely destroyed after the experiment's completion. The experiment followed the university Institutional Review Board protocol.

Note that the number of participants coincide with previous UX studies that have used an EEG device. It is not, as yet, possible to determine a sample size for UX tests performed with neurological monitoring, but it has been shown that the median sample size in EEG studies is 18 participants [65].

4.3. Experimental Objects

To the best of our knowledge, the largest list of graphical adaptive menus reported in literature is composed of 49 types (Table 5, [9]), from which we extracted 19 menus plus the static menu as the baseline on the basis of the following criteria: their preference percentages belonged to various intervals and were uniformly distributed (see Fig. 3), they covered the four major categories (*i.e.*, highly preferred, preferred, not preferred, rejected), they covered Bertin's variables in different combinations, and they represented candidates that had already been recognized as valuable conditions([47, 49, 44, 9, 6]).

The menus selected consisted of the following adaptive menus shown in Fig. 1: (1) *Static menu*, defined as the baseline condition [7], (2) *Pulsing menu* [66], (3) *Bolding menu* [46], (4) *Temporal menu* [67], (5) *Morphing menu* [68], (6) *Ephemeral menu* [49], (7) *Underlining menu* [6], (8) *Out-of-context disappearing menu* [69], (9) *Split without replication menu* [70], (10) *Italicizing menu* [6], (11) *Highlighting menu* [46], (12) *Grayscale menu* [9], (13) *Rotating menu* [6], (14) *Leaf menu* [71], (15) *Twisting menu* [66], (16) *Glowing menu* [72], (17) *Pink menu* [43], (18) *Blinking menu* [6], (19) *Boxing menu* [6], and (20) *Adaptive activation area menu* [73].

These menus have different properties (see Table 2), and we, therefore, expected that this spectrum would produce differences in the participants' performance and UX (measured by employing EEG variables). The graphical adaptive menus were initially presented in an abstract manner, *e.g.*, with *menuItem 1*, *menuItem 2*, that was independent of any context of use (Fig. 1). In order to make them realistic and induce a real interaction, we instantiated the menus to two popular contexts of use (Fig. 4): a mail manager menu and a web browser settings menu. For both contexts of use, we designed the menus in such a way that they all contained the same number of items (19), and incorporated them into an interactive presentation such that the participants could use them as if they were real.

4.4. Variables

4.4.1. Independent Variable

There was one independent variable in this study:

- Menu-Type: nominal variable with 20 conditions, representing the aforementioned 20 graphical adaptive menus with which to execute commands in the UI of a software system, along with the end users' preference percentage as regards using these menus (see Fig. 3). Note that these preferences were obtained in a previous study with a larger pool of participants [6].

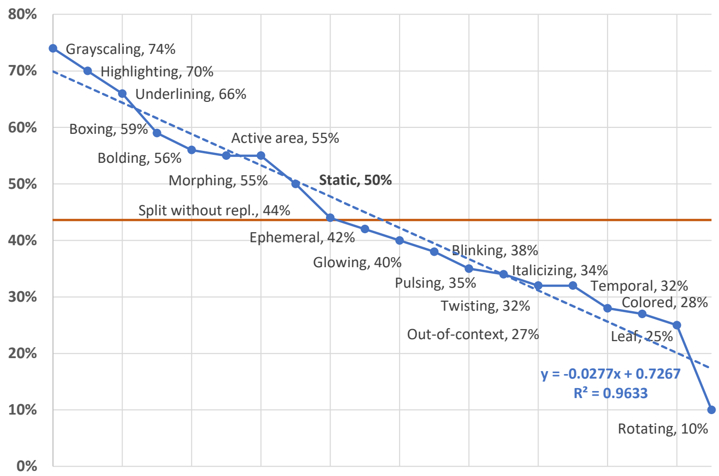

Figure 3: Distribution of our 20 graphical adaptive menus in terms of percentage obtained for preference [6].

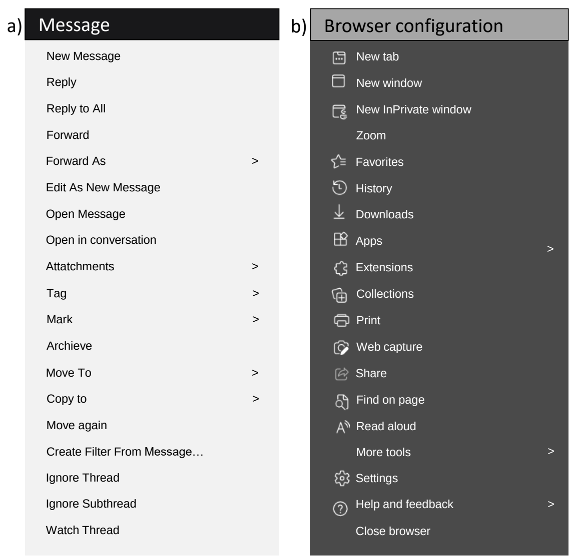

Figure 4: The two contexts of use for menu instantiation: (a) the mail manager message menu, and (b) the web browser settings menu.

Menu-Type	Property							
	Spatial		Physical		Format			Temporal
	Position	Orient.	Size	Shape	Value	Color	Text.	Motion
Static	Constant							
Split without repl.	var.							
Bolding					var.			
Highlighting					var.			
Morphing	var.		var.					
Adaptive area	var.		var.	var.				
Temporal					var.			var.
Ephemeral					var.			var.
Out-of-context				var.				var.
Leaf	var.		var.	var.		var.		var.
Twisting	var.	var.						var.
Rotating	var.	var.						var.
Pulsing			var.	var.				var.
Colored					var.	var.		
Grayscaling					var.			
Italicizing					var.			
Underlining					var.			
Boxing			var.		var.			
Blinking					var.		var.	
Glowing				var.	var.	var.		

Table 2: Graphical adaptive menus selected, sorted by stability property. Light yellow lines depict slightly changing menus. Peach lines depict menus with motion. Light blue lines depict spatially-stable menus. Green is a color typically used to represent glowing.

4.4.2. Dependent Variables

The dependent variables can be divided into two types: experience-based and performance-based variables. Experience-based variables assess participants' emotional and cognitive responses objectively. In particular, we used four EEG quantitative measures [74] to assess UX:

- Cognitive-Load: A continuous variable that measures the participant's neurological focus or concentration when presented with stimuli or during experiences calculated based on Antonenko et al. [75]. It represents the use of cognitive resources to carry out a task or visualize a stimulus. Cognitive load is expressed as a percentage: a value close to 0% indicates that participants are very distracted, while a value close to 100% indicates that they are very attentive to the stimulus.

- Engagement: A continuous variable that measures the degree of involvement or connection between the participant and the stimulus or task, calculated based on Lehmann et al. [76]. Engagement is expressed as a percentage: a value close to 0% indicates no connection or link to the stimuli, while a value close to 100% indicates high engagement with the stimuli or task.

- Attraction (also known as Valence): A continuous variable that measures the affective quality referring to the intrinsic attractiveness or adverseness experienced in response to stimuli or a situation, from a positive/pleasant reaction to a negative/unpleasant reaction, calculated based on Laming and Mason [77]. Attraction is expressed as a percentage: negative values (from −100% to 0%) indicate that participants had negative emotions while positive values (from 0% to 100%) indicate that participants had positive emotions.

- Memorization: a continuous variable that measures the intensity of cognitive processes related to the formation of future memories during the presentation of stimuli or during an experience, calculated based on Noh et al. [78]. This variable captures the degree of storage, encoding, and retention in memory. Memorization is expressed as a percentage: a value of 0% indicates that the chance 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 participant's memory.

These measures were obtained through the monitoring of brain activity in different areas (*i.e.*, pre-frontal, frontal, parietal, and occipital) and at frequencies (*i.e.*, δ, θ, α, γ, and β). This process was automated using the Bitbrain service [79] as follows: data were captured with SennsLab, decoded with SennsCloud, and analyzed with SennsMetrics, thus covering the entire workflow of the study. This service used calibration data and brain activity data to compute the four measures, on the basis of prior research into EEG data analysis (e.g., cognitive load ([19, 80]), engagement [21], attraction ([17, 18]), and memorization([81, 82])).

Encephalographs can vary greatly based on numerous factors such as task, stimulus, fatigue state, expertise, and directionality of attention. The standard approach to encephalographic analysis is as follows:

1. Visualization of all channels across the recording
2. Removal of dead or excessively noisy channels
3. Independent component analysis (minimum 20 components) to remove physiological and other noise artifacts (*e.g.*, blinks or signal transmission noise).
4. Filtering (1-40 Hz will leave you with the classic bands of δ, θ, α, β, and low γ).
5. Epoching (always epoch with at least one second before and after your measurement window of interest to permit time-frequency decomposition)
6. Visual inspection of all epochs and removal of those which are noisy (this is highly subjective, but 100-120 microvolts is often used as a threshold.
7. If event-related potentials are targeted, average across epochs.
8. If event-related perturbation is targeted, one method is to perform Hilbert transform, then average the resulting envelopes across epochs and standardize according to baseline activity. Often final averaging over time is appropriate.
9. If connectivity is targeted, compute on each epoch and average across epochs.
10. Source level activity is also possible to estimate, but we believe that this is not feasible with the device used.

As for cognitive load, engagement, attention, memory processes, etc. Generally, these are interpreted based on the combination of experimental design/context and the results of the above processing in comparison with findings in the extant literature.

Finally, we added a fifth dependent variable with which to measure user performance:

- Completion-Time: a continuous variable that measures (in seconds) the time that has elapsed between the initial state of the task and the final state, when the task is correctly completed.

Completion-Time is an important usability measure that reflects the amount of time that a user takes to complete a task or achieve a goal using a product or service [58]. This efficiency measure has a significant impact on UX: if the completion time is long (*i.e.*, it takes too much time for users to select an item from a condition menu) or short, the users may become frustrated or annoyed, which negatively impacts on their subjective satisfaction. However, users may feel that the menu is efficient and easy to use, which positively impacts on their subjective satisfaction.

4.5. Hypotheses

The following null hypotheses were proposed in order to verify the first research question RQ_1, namely whether the use of twenty adaptive graphical menus has a different influence on UX:

- H_{01}: There are no significant differences in the users' Cognitive-Load when using a different Menu-Type. H_{11}=$\neg H_{01}$.
- H_{02}: There are no significant differences in the users' Engagement when using a different Menu-Type. H_{12}=$\neg H_{02}$
- H_{03}: There are no significant differences in the users' Attraction when using a different Menu-Type. H_{13}=$\neg H_{03}$
- H_{04}: There are no significant differences in the users' Memorization when using a different Menu-Type. H_{14}=$\neg H_{04}$
- H_{05}: There are no significant differences in the users' Completion-Time when using a different Menu-Type. H_{15}=$\neg H_{05}$

We hypothesized that the Menu-Type would have an effect on the four EEG variables and on the Completion-Time, as some menus have different configurations, presentations and behavior: the more complex the menu type is, the higher the value of the Cognitive-Load, Memorization and Completion-Time measures should be.

In order to get a better understanding on these differences, we computed the Dynamic Time Warping (DTW) distance [83], a method with which to compare two time series that allows non-uniform time intervals between measurements. This works by stretching and compressing one of the time series as required in order to align it with the other time series, and then calculates the distance between the aligned time series. The computation of this distance makes it possible to compare time series that may have different lengths, maxima, minima and varying rates of measurement [84]. The DTW can be computed for two series so as to assess the similarity or difference between them. A small DTW distance suggests that the two time series are similar, while a large DTW distance suggests that the time series are different. The DTW distance is not constrained to the time series scale, and we consequently, proposed the following null hypotheses:

- H_{0-11}: There is no significant difference in the DTW distance between the participants' Cognitive-Load when using the Menu-Type. H_{1-11}=$\neg H_{0-11}$.
- H_{0-12}: There is no significant difference in the DTW distance between the participants' Engagement when using the Menu-Type. H_{1-12}=$\neg H_{0-12}$.
- H_{0-13}: There is no significant difference in the DTW distance between the participants' Attraction when using the Menu-Type. H_{1-13}=$\neg H_{0-13}$.
- H_{0-14}: There is no significant difference in the DTW distance between the participants' Memorization when using the Menu-Type. H_{1-14}=$\neg H_{0-14}$.

Subsequently, in order to verify the second research question RQ_2, we hypothesized that there would be a relationship between the percentage obtained for preference as regards the 20 menu types and their performance and UX expressed through the five measures. The following null hypotheses were, therefore, proposed in order to investigate whether the preference for a particular Menu-Type was correlated to their performance and UX:

- H_{06}: There is no relationship between the participants' Cognitive-Load on item selection and their preference percentage for the Menu-Type. $H_{16}=\neg H_{06}$.
- H_{07}: There is no relationship between the participants' Engagement on item selection and their preference percentage for the Menu-Type . $H_{17}=\neg H_{07}$.
- H_{08}: There is no relationship between the participants' Attraction on item selection and their preference percentage for the Menu-Type. $H_{18}=\neg H_{08}$.
- H_{09}: There is no relationship between the participants' Memorization on item selection and their preference percentage for the Menu-Type. $H_{19}=\neg H_{09}$.
- H_{0-10}: There is no relationship between the participants' Completion-Time on item selection and their preference percentage for the Menu-Type. $H_{1-10}=\neg H_{0-10}$.

4.6. Study Design and Task

We conducted one EEG session per participant. Participants were instructed to select a menu item for each of the 20 graphical adaptive menus, half of which were randomly instantiated in the first context of use and half in the second context (see Fig. 4). The menus and the item to be selected were randomly assigned to obtain a random distribution of context-menu-item. Each Menu-Type had the same item and context, and they were randomly assigned. Additionally, the order of usage was randomly assigned to each participant. Our design is a within-subjects design with 40 participants × 20 menus × 1 menu item × {10 instantiations to context #1 / 10 instantiations in context #2} = 800 trials.

4.7. Experimental Setting and Instrumentation

The experiment was performed in a quiet, controlled environment composed of (Fig. 5): a table supporting the participant's display (Apple LED Cinema Display 27', 2560 × 1440), which was connected to the experimenter's station (HP ProBook 450 G8, Intel i7-1165G, 16 GB, Windows 10) in order to control the experiment, present the stimuli and collect data; a mouse (MSI, Clutch GM08), and the headset. We used the BitBrain Diadem headset (Fig. 2) to capture a raw EEG signal with a resolution of 24 bits at 256 Hz. This device is a portable non-intrusive EEG headset that is used in neuroscience research and Brain-Computer Interfaces (BCI). Its twelve dry electrodes were placed in positions regulated by the international 10-10 standard [31], as shown in Fig. 2, so as to optimally estimate emotional and cognitive states (*i.e.*, it measures brain activity in the prefrontal, frontal, parietal, and occipital areas of the brain). The headset was wired to an amplifier that communicated with the experimenter's personal computer via Bluetooth in order to minimize the intrusiveness of the device. Possible data loss in the case of wireless connection failure or fault, was mitigated by storing raw data in a storage device located in the EEG amplifier. The experimental objects were displayed on the experimenter's personal computer on a desktop. This material, along with the scripts, are publicly accessible in our GitHub repository at https://resquelab.github.io/pages/experiment-1.html.

4.8. Experimental Procedure

Prior to the experiment, each participant was welcomed, had the process explained to them, signed a GDPR-compliant consent form on which the regulations and anonymity were explained, and filled out a demographic questionnaire. They then received instructions concerning the experimental procedure. The experimenter was also present to answer any question raised by the participant before the interactive session, but subsequently left in order to mitigate the Hawthorne effect. Each session started with the placement of the EEG headset, ensuring that the electrodes were well connected and capturing good signal quality through SennsLab software. This was followed by the calibration phase (8 min.) so as to establish a baseline brain activity for each participant in order to accurately acquire the measures:

1. *Calibration task familiarization* (1 min.): the participant learned to perform the EEG calibration.
2. *Closed eyes baseline* (2 min.): the participant's brain response was recorded in a relaxed state, with no other neurological responses such as blinking or eye movements. This was used to establish a null baseline for EEG data.
3. *Open eyes baseline* (2 min.): the participant's brain response was recorded in a relaxed state, also considering other neurological responses such as blinking or eye movements. This was used to establish a baseline measurement for EEG.
4. *Calibration baseline* (2 min.): the participant's brain response was recorded in a controlled state of stress. The participant counted backward in intervals of seven, out loud, starting with a number generated randomly.
5. *Closed eyes washout* (1 min.): in order to have some time to rest before starting the following steps, such that the calibration phase would not influence the following steps of the experiment.

Once the EEG baseline had been acquired, the participant was able to start the interactive session itself. The participant performed a practice trial with a menu that was not part of the

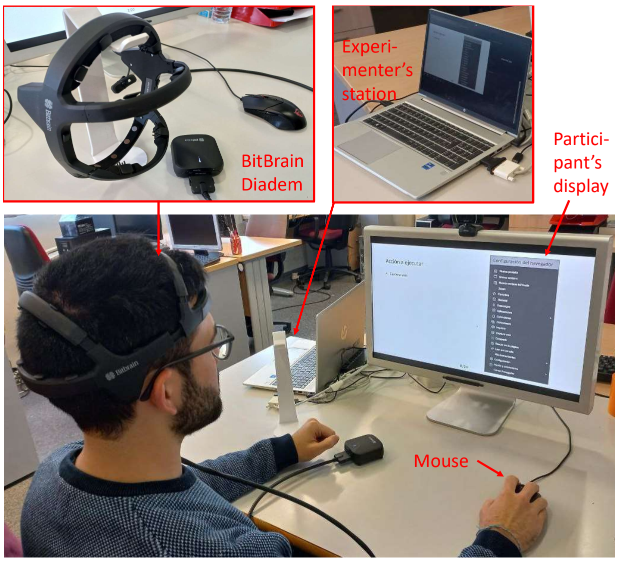

Figure 5: Setup of the experiment with its instrumentation.

subsequent evaluation so as to become familiarized with the graphical adaptive menus. This trial could be repeated as many times as necessary. Once familiarized, and after a resting phase of about ten seconds, the participant was able to start the experimental task: select a specific menu item displayed for each condition. This was composed of three parts:

1. *Item selection*: a randomly selected condition was presented. The menu was first hidden, and the participant was prompted with a menu item displayed in a prediction window [85] that would be highlighted by the menu. After expanding the menu, all menu items were displayed and the participant had to select the item prompted.
2. *Resting time*: the participant closed his/her eyes and did nothing for about ten seconds in order to maximize rest and washout emotions such as frustration, surprise, angriness, and impatience.
3. *Activation time*: the participant opened his/her eyes and went to the first step with the next condition until all conditions had been completed.

All mouse input and brainwaves were recorded during the entire process. When all the conditions had been completed, the recording stopped, the headset was removed, and the participant was allowed to go.

4.9. Data Analysis

Bitbrain SennsLab[2] recorded the EEG signals in real-time, and these were then processed using SennsCloud in order to compute the four measures regarding the brain zones, brainwave bandwidth and calibration samples from the EEG data. SennsCloud applies multiple Independent Component Analysis (ICA) techniques to process the brain signal data, including the removal of stereotyped eye movements, muscle artifacts, and line noise. We then used SennsMetrics[3] [79], a Jupyter[4] Notebook script using Python and R Studio, to analyze the data collected so as to eventually obtain the COGNITIVE-LOAD, ENGAGEMENT, ATTRACTION, and MEMORIZATION recorded during the whole experiment in the form of time series. The Dynamic Time Warping distance [83] was then computed. Since we expected to have time series with different lengths, as each participant would interact with the 20 conditions in different ways during a varying period, we proceeded as follows:

1. We grouped the time series per adaptive menu such that the participants' data could be compared.
2. We computed the distance matrix for each adaptive menu and variable (see Fig. 6 for a reduced example, in which each cell represents the distance between two participants located on the x-axis and the y-axis. For example, cell (3,1) represents the similarity between participants #3 and #1).
3. We calculated the mean DTW distance for each participant by averaging corresponding columns in the matrix. The mean value for the n column represents the mean distance of participants with ID n with respect to other participants.
4. We repeated this process for all the measures.

[2] https://ga-eyetracking.ru/bitbrain-sennslab
[3] https://ga-eyetracking.ru/bitbrain-sennsmetrics
[4] https://jupyter.org/

If a participant is recorded with a higher mean distance, this means that the participant's brain responses were very different from those of the other participants, while a lower mean distance means that the brain responses were more similar to the other participants. The results show which conditions were more different or similar.

After processing the time series and computing the distance matrix, we performed descriptive statistics, histogram plots, and statistical tests to analyze the data. We accepted a probability of α=5% of committing a Type-I Error, *i.e.*, rejecting the null hypothesis when it was 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 analyzed the characteristics of the data in order to determine which test would be most appropriate. Since the sample size was less than 50, we applied the Shapiro–Wilk test [86] to test the data normality as this provides a generally superior omnibus measure of non-normality [87] and the Brown-Forsythe Levene-type test [88] in order to determine the homogeneity of variance.
3. These results informed the testing of the null hypotheses formulated. When the data were normally distributed and the variances were homogeneous, it was possible to use a one-way Analysis of Variance (ANOVA) to analyze the data by considering the Menu-Type as a main factor. When these assumptions were not satisfied, we used the Friedman non-parametric one-way analysis of variance for matched data [89], since our design was within-subjects, considered as a repeated measure of each variable for each Menu-Type. When this test returned a significant difference, we used Dunn's pair-wise comparison test with multiplicity adjusted *p*-value. In fact, Dunn's test performs pair-wise comparisons between each Menu-Type and shows which ones are statistically significantly different at a level of α=.05.
4. When there was a significant difference for the value of two or more Menu-Types, we also computed the effect size: Cohen's d coefficient when two Menu-Types had similar standard deviations and were of the same size (which is our case), and Glass's Δ, when each Menu-Type had a different standard deviation. It was, therefore, possible to compare the effect size returned by the two coefficients: if they were close to each other, their interpretation would remain the same; if they were far from each other, their interpretation would be the lowest one. The magnitude of the effect size was assessed using the thresholds provided by Cohen [90]: d<0.2="very small", 0.2≤d<0.5="Small", 0.5≤d<0.8="Medium", d≥0.8="Large".

In order to investigate any potential correlation between UX (i.e., EEG variables) and performance with the preference for the 20 adaptive menus, the data analysis was performed as follows:

1. We computed Spearman's ρ rank correlation coefficient when the data did not pass the normality test. This coefficient is the counterpart of Pearson's r coefficient for the non-parametric case. We also computed Kendall's τ coefficient, a non-parametric test that measures a relationship using ranked data, which is considered more robust and preferred, but usually gives smaller values than Spearman's ρ.
2. We computed the cosine similarity S_C [91] among vectors expressing the participants' UX and performance for the five dependent variables and the percentage obtained for preference (Fig. 3). Cosine similarity was chosen owing to its simplicity of interpretation and

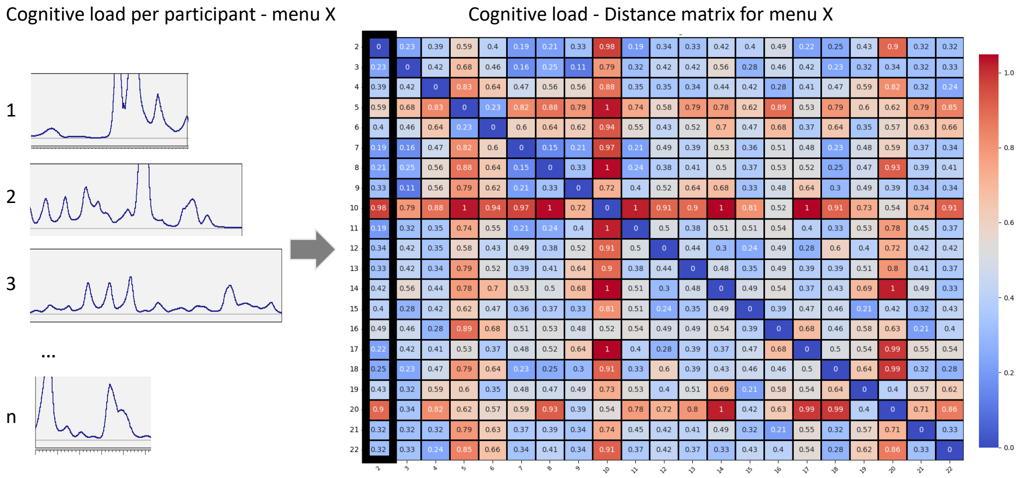

Figure 6: The distance matrix is computed from the time series data. We then calculated the mean distance for each participant with respect to the other participants (as shown in the black box in the first column).

its widespread adoption [92], along with other properties such as independence from vectors' length, which makes it independent of the layout size. The measure computes the cosine of the angle between the vector of UX and performance variables and the vector of the participants' preferences. Each vector represents a combination of EEG measures or performance metrics for each menu type, where each coordinate corresponds to a specific menu type. A value of 0 means that the two vectors are 90° from each other (orthogonal) and do not match. The closer the cosine value is to 1, the smaller the angle and the greater the match between similar vectors (positive co-linear).

3. We complemented the performance-preference analysis by performing an Importance-Perfor- mance Analysis (IPA) [93], which determined the participants' satisfaction by querying their performance with respect to their preference for a set of variables and presenting it on a plot. IPA has recently been used in software evaluation [55] in, for example, web development [94]. The variables in our case comprised Cognitive-Load, Engagement, Attraction, Memorization, with their corresponding DTW distance, and the Completion-Time.

5. Results and Discussion

This section presents and discusses the results of our study, examining the effect of Menu-Type on the five measures considered in the user experience and performance analysis in response to RQ_1 (Section 5.1), and summarizes the results of the UX and performance *vs.* preference analysis in response to RQ_2 (Section 5.2). We conclude by stating which hypothesis has been supported or rejected. Finally, we discuss the implications of our findings and their significance for the design and development of user interfaces containing graphical adaptive menus.

5.1. RQ_1: Analysis of User Experience and Performance

Fig. 7 shows the Cognitive-Load (❶ - see Table A.7 for details), Engagement (❷ - see Table A.10 for details), Attraction (❻ - see Table A.14 for details), Memorization (❹ - see Table A.18 for details), and Completion-Time (❺ - see Table A.23 for details) for each Menu-Type in decreasing order of averaged value, when compared to the *Static menu*.

We used the Coefficient of Variation (CV), a measure of relative variability calculated as the ratio of the standard deviation to the mean and expressed as a percentage, to analyze the dispersion of participants' responses. The descriptive analysis showed that the mean of the participants' Cognitive-Load (CV_{Min}=37% to CV_{Max}=55%), Engagement (CV_{Min}=32% to CV_{Max}=49%), Attraction (CV_{Min}=298% to CV_{Max} = 25614%), and Memorization (CV_{Min}=18% to CV_{Max}=38%) was largely dispersed among the different Menu-Types.

Overall, menus with atypical structures (*Leaf menu*), with different font styles (*Italicizing menu* and *Morphing menu*), menus with movement (*Rotating menu*) and menus with an increased activation area of menu items (*Morphing menu* and *Activation Area menu*) could have very different influences on users. However, the measures recorded were more similar among the participants when employing menus that use temporality (*Ephemeral menu*), color-changing properties (*Colored menu*), or both together (*Blinking menu* and *Glowing menu*). More specifically, all graphical adaptive menus induced a higher Cognitive-Load and Memorization than the *Static menu*, which is owing to the changing nature of the adaptive menu as opposed to the *Static menu*, which was never altered: i.e., changing any property of the menu to make it adaptive inevitably deteriorates these measures, but to no great extent.

On the positive side, seven menus were considered more engaging than the *Static menu*, in the following order: *Italicizing menu*, *Rotating menu*, *Boxing menu*, *Leaf menu*, *Blinking menu*, *Colored menu*, and *Twisting menu*, all of which have striking properties that increased their average Engagement. All the other menus were considered less engaging than the *Static menu*, which contradicts the preconceived notion that an adaptive menu requires more attention than a non-adaptive menu. Eight menus, including the *Static menu*, received a negative Attraction mean, thus indicating that the participants did not feel attracted by them. Menus that changed in color or temporality were considered the most attractive, specifically the *Morphing menu*, *Temporal menu*, *Colored menu*.

One interesting instance is the emblematic case of the *Ephemeral menu*, which the participants found to be faster than the *Static menu* and the *Highlighting menu* [2]. In fact, the Completion-Time confirms these results, but we also empirically demonstrate that the *Ephemeral menu* is slightly less engaging than the *Static menu* and requires slightly more Memorization than the *Static menu*, with much better Attraction. These two differences are minimal and could be considered negligible, thus promoting the Ephemeral menu as an excellent representative adaptive menu candidate when compared to the others.

Furthermore, the participants selected the items faster when using the *Glowing menu*, *Colored menu*, *Ephemeral menu*, and *Blinking menu*, whose Completion-Times were below that of the *Static menu*. The highest Completion-Time was recorded for the *Leaf menu*, *Italicizing menu*, *Activation area menu*, and *Rotating menu*. Using the *Static menu* as a baseline, the Completion-Time of every menu that uses temporal and color-changing properties or motion-properties (*e.g.*, *Twisting menu*, *Pulsing menu*) tends to be reduced. This suggests that the more time the participants need to complete the task, the more differences will emerge in the brain activity as a result of this.

Homogeneity of group variances was assessed using Levene's test [88], which indicated that all the EEG measures were homogeneous, while Completion-Time was not. None of the dependent variables were normally distributed since they did not pass the Shapiro-Wilk test. For example, the Cognitive-Load passed the normality test for a few conditions (*e.g.*, *Activation area*: W=0.98, p=.82, *n.s.*, *Grayscaling*: W=0.96, p=.22, *n.s.*), but did not pass for the vast majority of the other conditions (*e.g.*, *Boxing*: W=0.91, p=.005**, *Colored menu*: W=0.89, p=.0009***).

Since our five dependent variables were not normally distributed, we calculated a Friedman

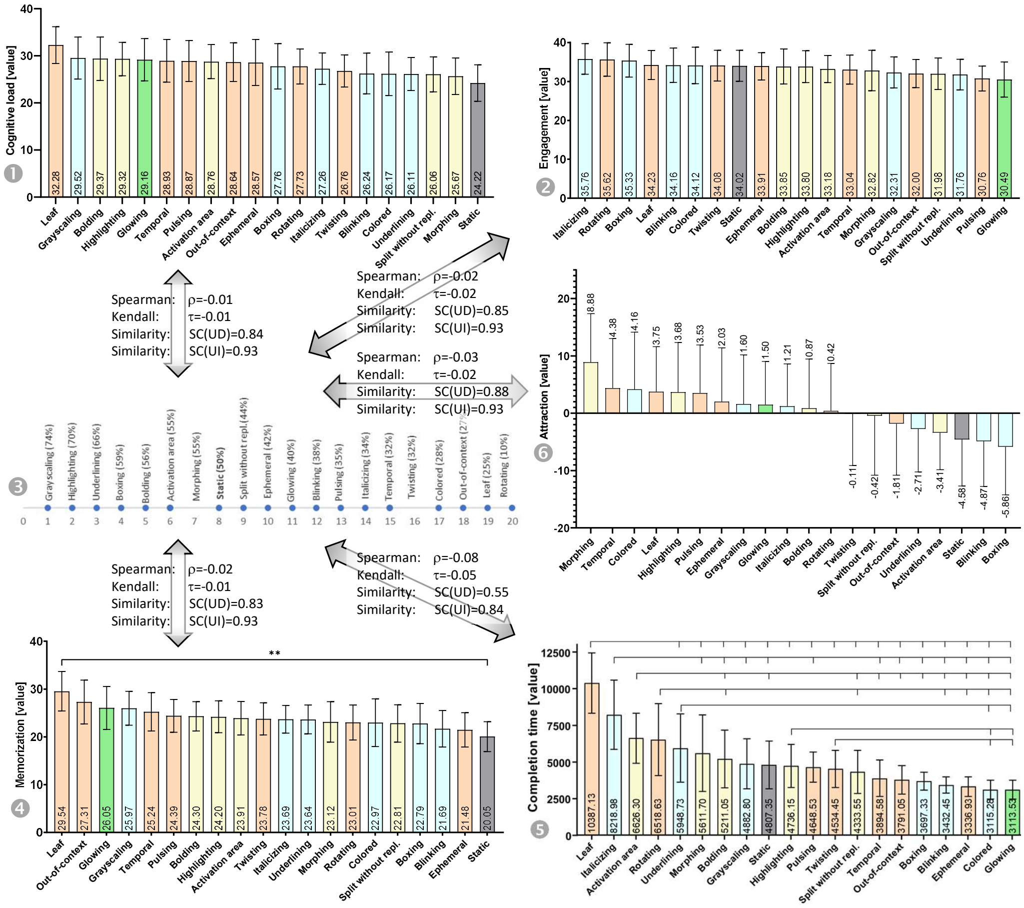

Figure 7: Results of the Performance *vs.* Preference Analysis (part 1/2): ❶ Cognitive-Load, ❷ Engagement, ❻ Attraction, ❹ Memorization, and ❺ Completion-Time for each Menu-Type in decreasing order of their mean value, in comparison to the *Static menu* (dark gray) as a baseline. Error bars show 95% confidence intervals. The preference distribution represented in ❸ reproduces the ordering of Fig. 3. Horizontal bars show a statistically significant difference among conditions, with levels omitted in order to preserve legibility.

Variable	Measure	Friedman's Q	p-value	Friedman's Q^*	p-value	Null hyp.	Conclusion
Cognitive-Load	Mean	22.00	.28, *n.s.*	1.16	.28, *n.s.*	H_{01}	Supported
	DTW	200.07	<.0001****	13.93	<.0001****	H_{0-11}	Rejected
Engagement	Mean	18.59	.48, *n.s.*	0.97	.48, *n.s.*	H_{02}	Supported
	DTW	317.56	<.0001****	27.99	<.0001****	H_{0-12}	Rejected
Attraction	Mean	19.10	.45, *n.s.*	1.00	.45, *n.s.*	H_{03}	Supported
	DTW	253.41	<.0001****	19.51	<.0001****	H_{0-13}	Rejected
Memorization	Mean	34.35	.016*	1.84	.015*	H_{04}	Rejected
	DTW	227.66	<.0001****	16.68	<.0001****	H_{0-14}	Rejected
Completion-Time	Mean	209.01	<.0001****	14.82	<.0001****	H_{05}	Rejected

Table 3: Results of Friedman's test for the mean and the DTW distance of the five dependent variables as regards 20 Menu-Types, n=40 participants, $df1$=19, $df2$=741.

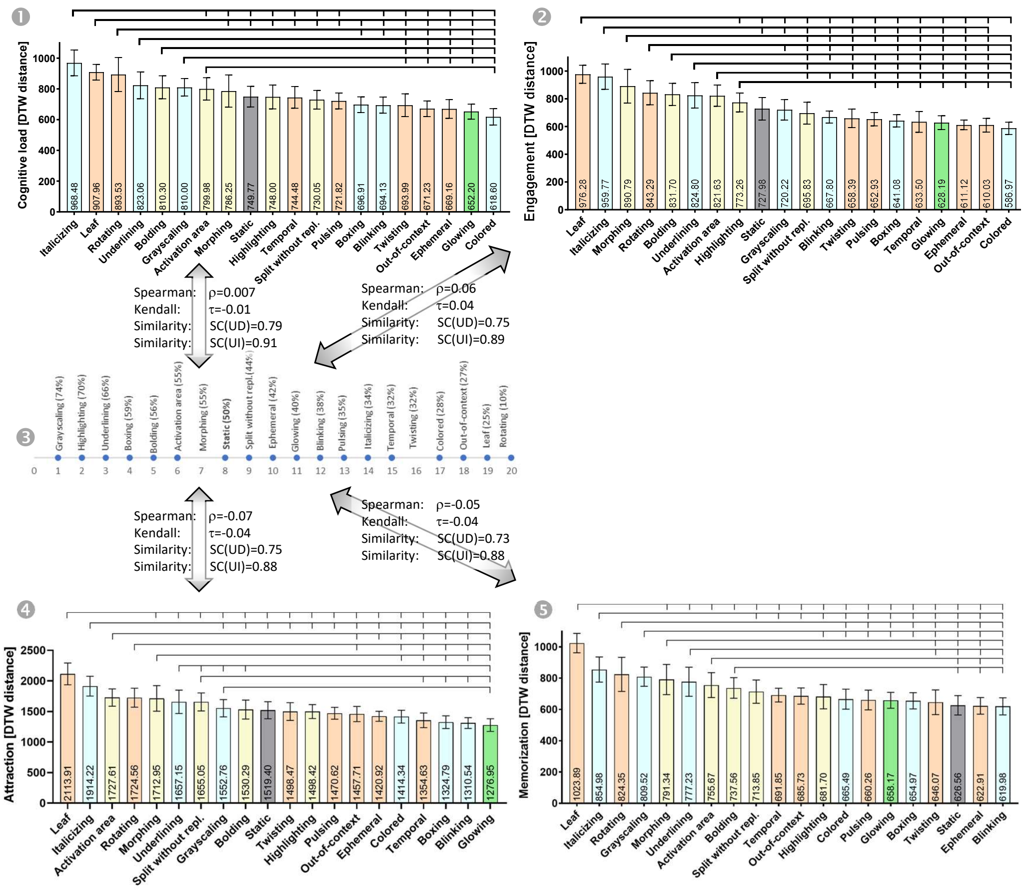

Figure 8: Results of the UX/Performance *vs.* Preference Analysis (part 2/2): ❶ Cognitive-Load, ❷ Engagement, ❹ Attraction, and ❺ Memorization, and for each ❸ Menu-Type in decreasing order of their averaged DTW distance, in comparison to the *Static menu* (dark gray) used as a baseline. Error bars show 95% confidence intervals. All five variables are compared to the preference percentage of Menu-Type ❸. Horizontal bars show a statistically significant difference among conditions, with levels omitted in order to preserve legibility.

test for both their mean and the DTW distance (Table 3). This revealed a significant effect of the condition Menu-Type only on the Memorization mean (p=.16*), on the Completion-Time mean (p<.0001****), and on all DTW distances (p<.0001****, all p-values in this paper were Bonferroni corrected for multiple comparisons). The absence of a significant effect of the condition Menu-Type on the mean of the three EEG variables Cognitive-Load, Engagement, and Attraction can be explained by the smoothing of the curves located below the time series. Specifically, when time series data are smoothed, small variations, including peaks and troughs (minima and maxima), are averaged out, which can diminish the ability to detect significant differences in these variables. On the contrary, the presence of a significant effect of the Menu-Type on the DTW distance among the time series shows that the participants converge in their good, average or poor performance in terms of item selection, independently of their assessment. The Menu-Type was found to have an effect on Memorization for only one condition: the *Leaf menu*, probably the strangest condition (the leftmost condition in Fig. 7-❸), required significantly more effort to memorize than the *Static menu* (the rightmost condition in the same figure).

In conclusion, the fist three null hypotheses related to RQ_1, *i.e.*, H_{01}, H_{02}, and H_{03}, are considered supported, while all others are considered rejected, thus confirming their corresponding alternative hypotheses, *i.e.*, H_{1-14}, H_{1-15}, and H_{1-11} to H_{1-14}. In order to better understand the source of the aforementioned effect, we performed a Dunn's pair-wise comparison test with a multiplicity adjusted p-value to determine the differences between groups in the case that an alternative hypothesis was confirmed in terms of: Cognitive-Load DTW distance (Table A.9), Engagement DTW distance (Table A.12), Attraction DTW distance (Table A.16), Memorization DTW distance (Table A.21), and Completion-Time (Table A.24). In each case of a significant difference being found between two groups, we computed Cohen's d and Glass's Δ size measures in order to assess the magnitude of the differences between groups. Any such difference is graphically depicted with a horizontal bar between groups in Fig. 7 and Fig. 8, without reporting the level of significance so as to preserve the legibility.

After computing and compiling Cohen's d and Glass's Δ effect sizes, very similar results were obtained. All dependent variables for which a significant difference was identified in Dunn's test have a predominantly "large" effect size, which reinforces the conclusion. The distribution of effect magnitudes for each dependent variable is summarized in Table 4.

Dependent Variable	Large	Medium	Small
Cognitive-Load DTW	68% ($\frac{34}{50}$)	28% ($\frac{14}{50}$)	4% ($\frac{2}{50}$)
Engagement DTW	75% ($\frac{61}{81}$)	22% ($\frac{18}{81}$)	3% ($\frac{2}{81}$)
Attraction DTW	73% ($\frac{41}{56}$)	27% ($\frac{15}{56}$)	0% ($\frac{0}{56}$)
Memorization DTW	52% ($\frac{31}{60}$)	46% ($\frac{28}{60}$)	2% ($\frac{1}{60}$)
Completion-Time	47% ($\frac{24}{51}$)	33% ($\frac{17}{51}$)	20% ($\frac{10}{51}$)

Table 4: Distribution of Effect Sizes for Dependent Variables

Given the wide spectrum of possibilities, with the multiple properties of the menus (*e.g.*, differences in position, motion, and shape), it was expected that there would be significant differences among the menu types. Overall, these results suggest that using different types of graphical adaptive menus can produce statistically and practical significant differences in the users' UX and performance as regards their cognitive load, engagement, memorization, attraction and completion time.

In order to confirm or refute these results, we conducted an operational internal replication study in October-November 2022 [8]. The same experimental protocol was applied to investigate RQ1 with the same operationalization and experimenters, but to a different population (a more homogeneous group of users). They were 40 students enrolled on a Master's degree in Computer Science at the Universitat Politècnica de València. The results obtained in the replication confirmed the ones obtained in this experiment. This allowed us to find initial evidence that human factors (e.g., gender, age, background) do not affect user experience when employing different adaptive menu types. Nevertheless, this finding needs to be further investigated.

On the one hand, these findings suggest that, if more differences among users were obtained for a menu, this could indicate that the menu is more subjective and that users have varying preferences or cognitive styles that affect how they interact with the graphical adaptive menu. This finding could have implications for the design of graphical adaptive menus and suggests that more adaptive interfaces could be beneficial for users with different cognitive styles. On the other hand, if more similarities among the participants were obtained for a menu, this could suggest that the menu design is more objective and easier to learn and use. This could be a desirable quality for certain types of menus, such as those that are frequently used or critical for task completion time. In order to further investigate these aspects, the UX and performance *vs.* preference are analyzed in the following section.

5.2. RQ_2: Analysis of User Experience and Performance vs. Preference

It is now important to better understand the relationship between preference and the UX/performance analyzed in the previous section, which is related to our second research question.

We first analyzed the correlation between the five UX and performance measures in terms of both their mean values (Fig. 7) and their DTW distances (Fig. 8) with the preference for the 20 graphical adaptive menus (Fig. 3). We calculated Spearman's ρ correlation coefficient, a statistical measure of the strength of a monotonic relationship between two data series constrained by $-1 \leq \rho \leq +1$, here the 5×2 measures and the preference. Since this correlation reflects a potential effect size between the two data series, we can interpret its value as follows [90]: $|\rho| \in [.00, .20[$="very weak", $|\rho| \in [.20, .40[$="weak", $|\rho| \in [.40, .60[$="moderate", $|\rho| \in [.60, .80[$="strong", $|\rho| \in [.80, 1]$ ="very strong". With regard to mean values, all coefficients are "very weak" ($|\rho_{min}|$=.01 to $|\rho_{max}|$=.08) for all measures. Kendall's τ values follow the same trend. With regard to DTW distance, all coefficients are again "very weak" ($|\rho_{min}|$=.007 to $|\rho_{max}|$=.07). The same occurs with Kendall's τ, which suggests that there is no real monotonic correlation between UX/performance and preference. Whereas the values of these coefficients are in the lowest interval, the cosine similarity shows higher values, thus suggesting that even if there is no correlation between pairs of variables, these same pairs could evolve in a similar manner.

Table 5 shows the correlation coefficients calculated between Completion-Time and the mean and the DTW of the four UX measures, respectively. While the Completion-Time is very weakly correlated with the mean of the four variables (all $|\rho| \in [.00, .20]$), their corresponding DTW is moderately correlated (all $|\rho| \in [.40, .60]$) but always very significantly (all corresponding to $p \leq .001$). The correlation between each variable and its corresponding DTW is almost always very weak, but is weak for Memorization: $\rho = .26^{***} \in [.20, .40]$. Indeed, the higher the Completion-Time needed to select items from a graphical adaptive menu, the more demanding the Memorization is for the end user.

As the correlations calculated are mostly (very) weak, we decided to calculate the cosine similarity. Fig. 9 shows that the individual values in the user-independent scenario (UI), rang-

	Cognitive-Load			Engagement			Attraction			Memorization		
	Mean	$\longleftrightarrow$	DTW	Mean	$\longleftrightarrow$	DTW	Mean	$\longleftrightarrow$	DTW	Mean	$\longleftrightarrow$	DTW
Spearman's ρ	0.023	0.17^{***}	0.44^{***}	0.044	0.17^{***}	0.54^{***}	-0.098^{**}	-0.035	0.47^{***}	0.11^{**}	0.26^{***}	0.45^{***}
Kendall's τ	0.013	0.11^{***}	0.31^{***}	0.031	0.10^{***}	0.39^{***}	-0.059	-0.020	0.34^{***}	0.079^{***}	0.16^{***}	0.32^{***}

Table 5: Correlation coefficients between Completion-Time and performance measures (Mean and DTW distance). Double-ended arrows show correlations between each measure and their corresponding DTW.

ing from $S_C(UI)$=0.88 to $S_C(UI)$=0.93 (last column), are all very high, with the exception of Completion-Time ($S_C(UI)$=0.84), thus suggesting a significant goodness of fit between the participants' UX/performance and preference, expressing their perception of the adequacy of the same preferences for the corresponding graphical adaptive menus. For example, Cognitive-Load, Engagement, Attraction, and Memorization all received the highest score $S_C(UI)$=0.93 with respect to their mean values, slightly less with respect to their DTW distance, meaning that the angles between the two vectors could be said to be 93% similar. The cosine value averaged across all conditions is equal to 0.7, which can still be interpreted as a high value given the variety of the 20 graphical adaptive menus. Since similarities were computed on the basis of averaged users' rankings, thus making them user-independent, we also assessed the goodness of fit in a user-dependent scenario by computing the cosine similarity for each measure on the basis of the participants' measures without aggregation and the corresponding preferences. The similarities obtained (M=0.87, SD=0.04) continue to show a favorable adequacy between the data series. These similarities are also indicated in Fig. 7 for the mean values and in Fig. 8 for the DTW distances, respectively.

Fig. 10 shows that the individual values in the user-dependent (UD) scenario, ranging from $S_C(UD)$=0.73 to $S_C(UD)$=0.88 (last column of Table 10), are all very high, with the exception of Completion-Time ($S_C(UD)$=0.55), thus suggesting a significant goodness of fit between the participants' UX/performance and preference, and showing their perception of the adequacy of these same preferences for the corresponding graphical adaptive menus. For example, Attraction received the highest score $S_C(UD)$=0.88, meaning that the angles between the two vectors could be said to be 88% similar. In conclusion, despite the similarity between our variables, the (very) weak correlations between the mean values of our variables led us to reject hypotheses H_{06} to H_{0-10}. However, the moderate correlations between the DTW distances and the preference, along with their high levels of significance, support hypotheses H_{0-11} to H_{0-14}.

In order to further investigate why there is no real correlation despite the high similarity, we now show the results of the IPA. The recommendations for action are derived from the arrangement of the menus in the plot, which consists of four quadrants ([93, 95]) adapted as follows:

1. *Quadrant 1* (Q1, in green) corresponds to high-UX/performance values and high-preference values for menu variables. It is labeled "Keep up the good work", because the variables of the menus contained in this quadrant are positively rated by the participants, who acknowl-

User independent	Engagement	Memorization	Attraction	Cognitive load	Duration	Engagement distance	Memorization distance	Attraction distance	Cognitive load distance	Preference percentage
Engagement	1.00	0.99	1.00	1.00	0.92	0.94	0.96	0.96	0.98	0.93
Memorization		1.00	1.00	1.00	0.92	0.94	0.97	0.96	0.97	0.93
Attraction			1.00	1.00	0.91	0.94	0.96	0.96	0.97	0.93
Cognitive load				1.00	0.92	0.94	0.96	0.96	0.97	0.93
Duration					1.00	0.99	0.98	0.99	0.97	0.84
Engagement distance						1.00	0.98	0.99	0.99	0.89
Memorization distance							1.00	0.99	0.99	0.88
Attraction distance								1.00	0.99	0.88
Cognitive load distance									1.00	0.91
Preference percentage										1.00

Figure 9: Cosine similarity in the user-independent scenario.

User Dependent	Engagement	Memorization	Attraction	Cognitive load	Duration	Engagement distance	Memorization distance	Attraction distance	Cognitive load distance	Preference percentage
Engagement	1.00	0.75	0.89	0.79	0.56	0.77	0.73	0.77	0.78	0.85
Memorization		1.00	0.83	0.92	0.56	0.71	0.78	0.74	0.77	0.83
Attraction			1.00	0.84	0.56	0.74	0.75	0.76	0.79	0.88
Cognitive load				1.00	0.53	0.71	0.73	0.74	0.79	0.84
Duration					1.00	0.87	0.85	0.83	0.84	0.55
Engagement distance						1.00	0.89	0.88	0.90	0.75
Memorization distance							1.00	0.87	0.93	0.73
Attraction distance								1.00	0.88	0.75
Cognitive load distance									1.00	0.79
Preference percentage										1.00

Figure 10: Cosine similarity in the user-dependent scenario.

edge both their preference and their performance. Designers are encouraged to maintain current strategies for these variables and retain these menus as ideal candidates.

2. *Quadrant 2* (Q2, in yellow) corresponds to high-UX/performance values and low-preference values for menu variables. It is labeled "Concentrate here", because the menus that fall into this quadrant should receive the highest priority. Action should be taken to address the challenges raised by these menu attributes, because the participants performed well with these menus but would not prefer to use them for their own purposes. For example, the participants do not like to be forced to operate efficiently in a continuous manner.
3. *Quadrant 3* (Q3, in orange) corresponds to low-UX/performance values and low-preference values for menu variables. It is labeled "Low priority", because the menus located in this quadrant are not essential and do not perform well for the participants, and could be discontinued in the design process without any detriment.
4. *Quadrant 4* (Q4, in ivory) corresponds to low-UX/performance values and high-preference values for menu variables. It is labeled "Possible overkill", because the participants do not perform well with the menus located in this quadrant, yet perceive them to be highly preferred.

Our IPA assigns each menu variable to four different quadrants according to a differentiation provided by the coordinate origin as regards the average value of each individual variable, which is represented by plain orange lines in the subsequent figures. The *Static menu* is represented in bold type in these figures.

Fig. 11 shows the results of the IPA for the COGNITIVE-LOAD variable (left) and its corresponding DTW distance (right). Those that attained the best score are the *Underlying menu*, the *Boxing menu*, the *Morphing menu*, and the *Split menu without any replication*, respectively, which are again all distant from the *Static menu*. The superiority of all the adaptive menus when compared to the *Static menu* in terms of COGNITIVE-LOAD again seems to demonstrate that adaptivity increases the participants' cognitive load with respect to the non-adaptive menu, a suggestion already made in the case of MEMORIZATION. Candidates with a high preference but which continue to have a high COGNITIVE-LOAD are: the *Grayscaling menu*, the *Highlighting menu*, the *Activation area menu*, and the *Bolding menu*. However, their behavior is intrinsically defined, and it is, therefore, challenging to make them more flexible. With regard to the DTW distance, the *Static menu* is, in this case, close to the average point of the quadrants, thus suggesting that the adaptive menus on its left are more representative and successful at computing their DTW distance. The best solutions include the *Highlighting menu*, the *Boxing menu*, and the *Split menu without replication*, respectively. The small DTW distance of these adaptive menus means that the participants converged as regards having a similar reaction to these menus.

Fig. 12 shows the results of the IPA for the ENGAGEMENT variable (left) and its corresponding DTW distance (right). The *Highlighting menu*, the *Boxing menu*, and the *Bolding menu* are

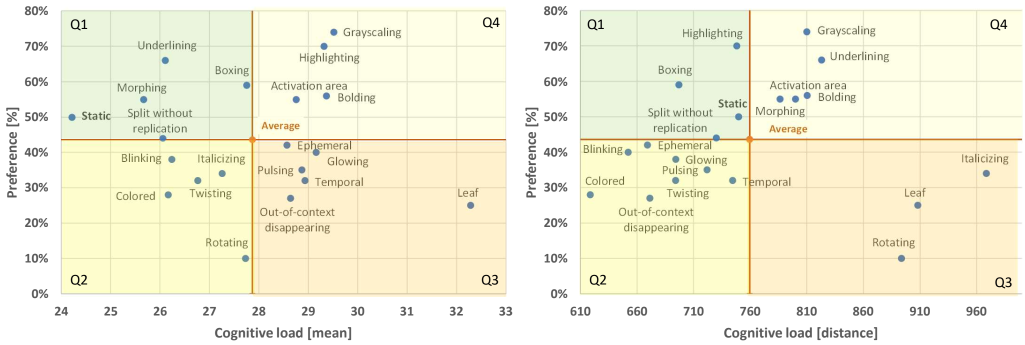

Figure 11: Cognitive-Load *vs.* Preference Analysis: mean value (left), DTW distance (right).

the most engaging and most preferred menus, probably because their common visual variable is positively affected, which is not the case of the *Grayscaling menu* and the *Underlying menu*, which were not perceived to be engaging. Unusually, the *Static menu* also belongs to Q1, but is the least preferred in this category. Q3 covers almost all visually demanding menus, such as the *Glowing menu*, the *Pulsing menu*, the *Out-of-context disappearing menu*, and the *Temporal menu*, all four of which employ animation over time. The participants did not view these types of animation as engaging, thus confirming that animation should be used carefully only in extreme situations, and not in usual tasks. With regard to the DTW distance, the *Highlighting menu* and the *Bolding menu* confirm their advantageous position in Q1, while the *Morphing menu* is tempting: it is among the most engaging and is fairly well preferred. The *Boxing menu* and the *Activation area menu* also belong to Q1. As occurred with the mean value, almost all the menus with employing animation are in Q4, and not are considered to be good.

Fig. 13 shows the results of the IPA for the Attraction variable (left) and its corresponding DTW distance (right). The *Grayscaling menu*, the *Highlighting menu*, and the *Bolding menu* belong to Q1, which makes them simultaneously the most attractive and the most preferred menus. Q2 contains a wide range of menus whose preference should be improved. A representative example is the *Ephemeral menu*, which had a compelling Attraction, with a preference that is fairly close to the average. The *Pulsing menu* seems to be the most attractive menu, but with a smaller preference. Note that the Static menu belongs to Q3 and is, therefore, among the least attractive menus, which suggests that the participants perceived all the other adaptive menus, with the exception of the *Blinking menu*, to be more attractive than the *Static menu*. Adaptive menus are, therefore, usually perceived to be more attractive than rigid static menus. Q4 includes the *Split menu without replication*, which was reasonably preferred but eventually attained a moderate performance, probably because the absence of the replicated items forced the participants to seek the item in the adaptive part. Q4 contains the *Underlying menu* and the *Activation area menu*. The use of menus with movement, such as the *Blinking menu* and the *Rotating menu*, should be reconsidered. With regard to the DTW distance, all of the aforementioned Q1 menus are replaced with three new ones: the *Activation area menu*, the *Morphing menu*, and the *Underlying menu*.

Fig. 14 shows the results of the IPA for the Memorization variable (left) and its corresponding DTW distance (right). Since the goal of memorization is to minimize the amount of data and functions to be memorized, memorization should be minimized while preference should be maximized. The quadrants to be considered are, therefore, inverted and their color scheme is

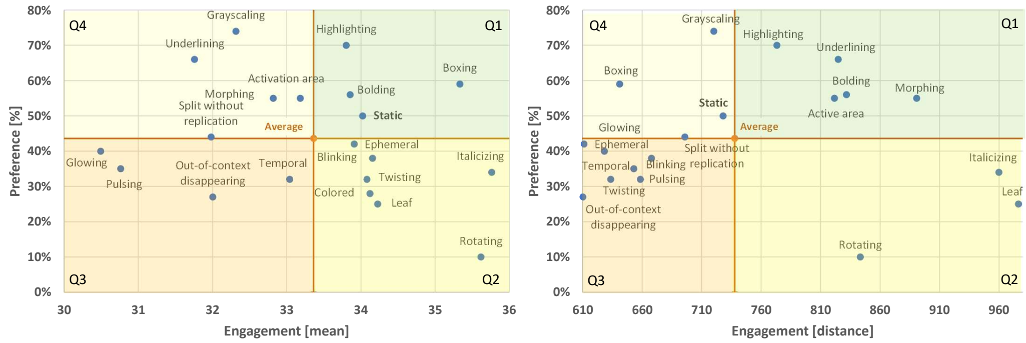

Figure 12: Engagement *vs.* Preference Analysis: mean value (left), DTW distance (right).

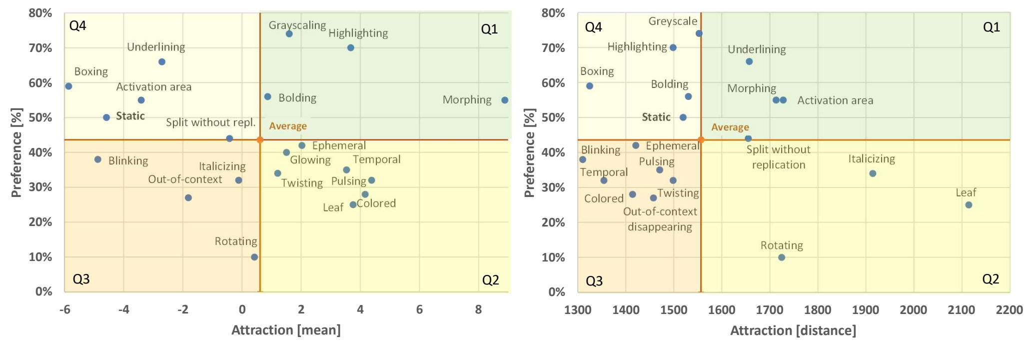

Figure 13: Attraction *vs.* Preference Analysis: mean value (left), DTW distance (right).

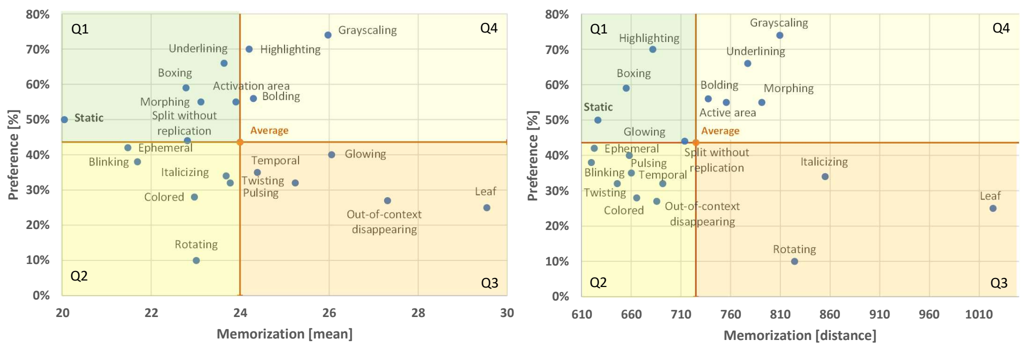

Figure 14: Memorization *vs.* Preference Analysis: mean value (left), DTW distance (right).

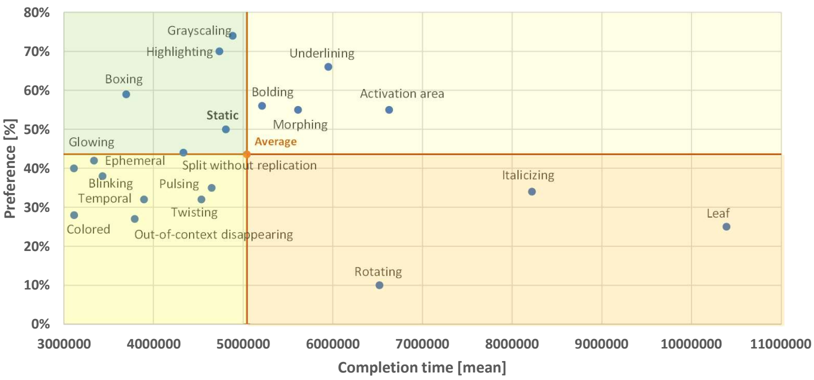

Figure 15: Duration *vs.* Preference Analysis.

updated accordingly. The same occurs with the COGNITIVE-LOAD in Fig. 11. The *Underlying menu* and the *Boxing menu* are the first two, followed equally by the *Morphing menu*, the *Activation area menu*, and the *Split menu without any replication*. Fig. 14 locates the *Static menu*, which is the baseline condition, at the leftmost point in the plot, thus suggesting that the *Static menu* induces the lowest memorization, which is natural since there is no mechanism for adaptation, but it is not as preferred as the aforementioned menus. The potential benefits of adaptive menus, such as efficiency, therefore inevitably come at the cost of certain amount of increase in memorization. Candidates for improvement as regards memorization are the *Grayscaling menu*, the *Highlighting menu*, and the *Bolding menu*, all of which all belong to Q4. With regard to the DTW distance, the *Highlighting menu*, the *Boxing menu*, and the *Glowing menu* maintain their positions. Candidates for improvement include the *Grayscaling menu* and the *Underlining menu*, along with the *Bolding menu*, the *Morphing menu* and the *Activation area menu*.

Fig. 15 shows the results of the IPA for the COMPLETION-TIME: the *Static menu* is located close to the average point of the plot, which suggests that menus located to its left are not only faster than the *Static menu*, but also more preferred if located in the top left-hand quadrant. The *Highlighting menu* and the *Grayscaling menu* are about the same in terms of *Completion-Time*, but are definitely more preferred than the *Static menu*. The *Boxing menu* is even faster, but less preferred than the *Grayscaling* and *Highlighting menus*. The bottom left quadrant (depicted in yellow in Fig. 15) contains 9 menu conditions that have all been shown to be as faster than the *Static menu*, sometimes only slightly (as is the case of the *Pulsing* and *Twisting menus*) and sometimes more significantly (as is the case of the *Blinking* and the *Colored menus*), but are less preferred than the *Static menu*. Three menu conditions are slower and less preferred than the *Static menu*: the *Rotating menu*, the *Italicizing menu*, and the *Leaf menu*, which is the slowest. Four candidates are slower than the *Static menu*, but are more preferred: the *Bolding menu*, the *Morphing menu*, the *Underlying menu*, and the *Activation area menu*.

6. Implications for Research and Practice

The findings of this study provide significant information that can serve as a basis for both future research and practical applications in the field of user interface design. As we explore the implications of our findings, we distinguish between recommendations for researchers, who can build on these findings to advance understanding of how different elements of the user interface influence the user experience - and practical guidance for practitioners - who can leverage these insights to optimise the user experience in software design . This section explores these implications further and presents an analysis that aims to bridge the gap between empirical research and real-world application, ultimately improving both the design process and the user experience.

6.1. Implications for Researchers

This experiment provides a starting point for researchers interested in exploring how user interface elements affect user experience, along with how to quantify user experience using biometric data. While our study utilized EEG to measure cognitive load, engagement, attraction, and memorization, there are several other neurological and physiological measures that can complement or expand upon these findings. Researchers can employ methods such as eye tracking, which can provide data on visual attention and scanning patterns, or facial expression analysis to assess emotional responses. Galvanic skin response and heart rate variability (HRV) are also valuable for assessing stress and arousal levels.

Moreover, while we have focused on graphical adaptive menus, there are many other user interface elements that could be studied, such as forms, text fields, and action buttons. Expanding the range of interface elements studied will allow researchers to develop a more comprehensive understanding of how different UI elements impact on user experience.

Furthermore, future studies could also explore how different user groups, such as those with different ages, cultural backgrounds, or cognitive abilities, respond to different UI elements. These studies could lead to the development of more inclusive and accessible software systems. In fact, we have conducted recently a replication of part of this experiment to understand how graphical adaptive menus have a different influence on UX (i.e., RQ1), by employing a different (more homogeneous) user group. As previously reported, the findings of the replication study confirmed the findings of this experiment.

Finally, as more studies are conducted and more knowledge is gained, it would be valuable to develop guidelines that UI designers or developers could use when creating effective graphical adaptive user interfaces. These guidelines could be based on empirical evidence of how different user interface elements impact on user experience, along with expert opinions, best practices from industry and industry standards. This can help to make sure that designs are not only effective but also adhere to widely accepted conventions. This includes following established design principles like the principles of Fitts' law for predicting user interaction time, and the use of familiar UI patterns that users recognize and understand.

Moreover, the development of guidelines should be an ongoing process, incorporating new research findings and technological advancements. Establishing a feedback loop where designers and developers can share insights and experiences will help continuously refine and improve the guidelines. Finally, it is important to increase awareness about the importance of these guidelines and how to implement them effectively through training programs, workshops, and resources.

Such guidelines would help ensure that UI designers or software developers create interfaces that are not only aesthetically pleasing, but also functional, intuitive and adapted to the specific needs of their target users.

6.2. Implications for Practitioners

Using our results as a basis, this section suggests some implications for practitioners, particularly in the design of graphical adaptive menus, based on the debate regarding performance *vs.* preference. Since the UX and performance measures are not correlated with preference, we shall distinguish two families of design implications according to priority: if preference takes precedence over UX/performance, we shall synthesize the results obtained in the IPA, while if a UX/performance measure takes precedence over preference, we shall compute the score for each condition using the *de Borda method* [96]. This method ranks graphical adaptive menus for a selection based on the results obtained for each measure of UX and performance while taking the preference into account.

We rank the adaptive menus in ascending order of their standing in the UX/performance measures combined with their preference rankings. The lowest score is commonly assigned to the most preferred candidate and the score increases with the ranking. For example, the scores assigned to preference have the following distribution (Fig. 3): 1 for the Grayscaling menu, the most preferred menu, 2 for the *Highlighting menu*, 3 for the *Underlining menu*, and so on until the least preferred menu, *i.e.*, 20 for the *Rotating menu*. The scores assigned to the Cognitive-Load follow the order of the mean attained (Fig. 7): 1 for the menu with the lowest mean, *i.e.*, the *Static menu*, 2 for the *Morphing menu*, 3 for the *Split without replication menu*, and so on to the menu with the highest mean, i.e., 20 for the Leaf menu. The same applies in the case of the DTW distance. The final de Borda score is simply the sum of these rankings, and the best menus are those for which this score is minimized.

Fig. 16 graphically depicts the final score by UX and performance measure in order to eventually obtain the overall score. For example, the *Boxing menu* ranks best with respect to the Cognitive-Load, drops to ninth place for Engagement and eighth for Attraction, but rises to second place for Memorization and first place for Completion-Time, to finish in fifth place overall. One edifying case is that of the *Ephemeral menu*, which obtains the best overall score, as it ranks very well in most measures, but is not always in first place: seventh for Cognitive-Load and Engagement, fifth for Attraction, and third for Completion-Time, despite a preference of 42%. The bottom part contains three menus that always receive low rankings and which should, therefore, be avoided: *Italicizing menu*, *Rotating menu*, and *Leaf menu*. The design implications are summarized in Table 6.

Moreover, the information obtained in this experiment may be valuable for software designers who wish to create good user experiences. Our experiment provides a way in which to objectively quantify the user experience of different user interface variants and identify which ones have a greater impact on users. This knowledge could be used to make informed decisions when designing user interfaces of software systems, ultimately leading to higher quality software products.

By using the metrics obtained in this experiment, designers can evaluate the effectiveness of their designs and make changes that will improve user experience. For example, if the design of a menu is causing high cognitive load, then the UI designer or developer could simplify the structure of the menu or change the font style in order to make it easier to read. Alternatively, if the engagement levels are low, UI designers could consider adding movement to the menu in order to make it more interesting.

The data obtained could also be used to adapt the user interface design to particular groups of users. By analyzing the neurological responses of different groups of users, UI designers could create interfaces that meet the needs of specific user groups, such as those with different levels of experience. For example, if the cognitive load is too high for novice users, UI designers could simplify the user interface element (*e.g.*, a graphical menu) in order to make navigation easier.

Furthermore, UI designers could use the information obtained in this study to identify trends as regards which the most popular and effective menus are among different users. This knowledge could then be incorporated into future UI designs to in order to create interfaces that are more

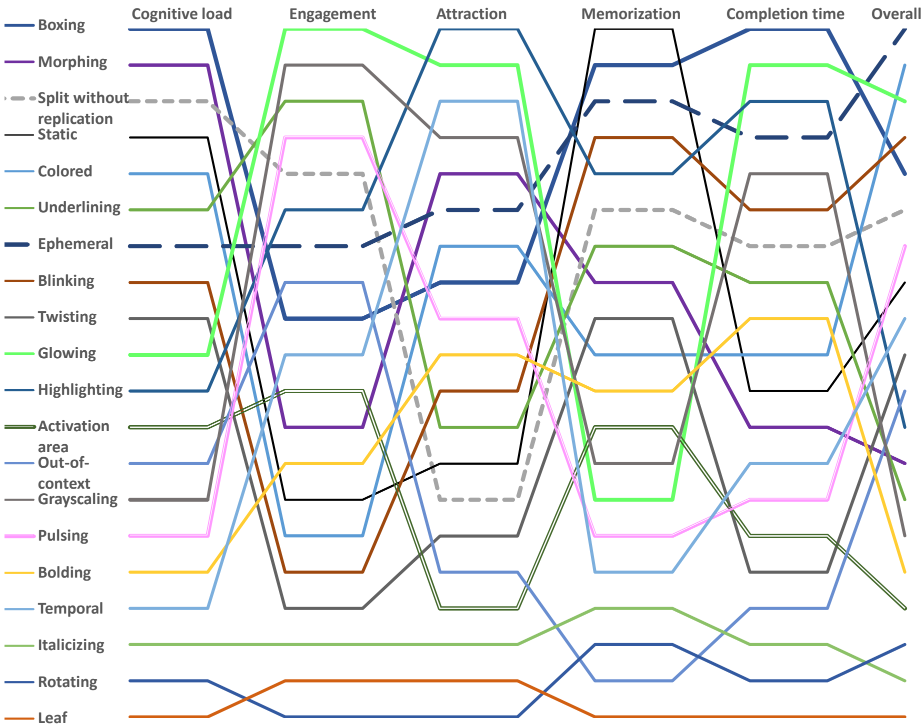

Figure 16: Ranking of the twenty graphical adaptive menus according to their de Borda ranking.

user-friendly and engaging. Designers could use these data in order to ensure that their products are adjusted to their users' needs and preferences, leading to higher satisfaction and increased user retention.

7. Threats to Validity

The threats to the validity of our empirical study are explained in terms of four categories: threats to internal, external, construct, and conclusion validity [97].

7.1. Internal validity

The learning effect, the understandability of the slides used to represent the menus, and instrumentation validity must be considered. The learning effect was mitigated by randomizing 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 4.8. Finally, in order to avoid any possible sources 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 that researches and develops high-quality EEG devices. The electrodes of this device are located over specific areas of the brain, according to the 10-10 location system, which is optimal, as regards computing emotional states, memorization, and cognitive load [98].

To optimize	Top preferred menus are	Top performing menus are
COGNITIVE-LOAD	*Underlining menu, Boxing menu, Morphing menu, Split without replication menu*	*Boxing menu, Morphing menu, Split without replication menu, Colored menu, Underlining menu, Ephemeral menu*
ENGAGEMENT	*Highlighting menu, Boxing menu, Bolding menu*	*Glowing menu, Grayscaling menu, Underling menu, Pulsing menu, Split without replication menu, Highlighting menu, Ephemeral menu*
ATTRACTION	*Grayscaling menu, Highlighting menu, Bolding menu, Morphing menu*	*Highlighting menu, Glowing menu, Temporal menu, Grayscaling menu, Morphing menu, Ephemeral menu*
MEMORIZATION	*Underlining menu, Boxing menu, Morphing menu, Activation area menu, Split without replication menu*	*Boxing menu, Ephemeral menu, Blinking menu, Highlighting menu, Split without replication menu, Underlining menu, Morphing menu*
COMPLETION-TIME	*Grayscaling menu, Highlighting menu, Boxing menu, Glowing menu*	*Boxing menu, Glowing menu, Highlighting menu, Ephemeral menu, Grayscaling menu, Blinking menu*
OVERALL	*Ephemeral menu, Colored menu, Glowing menu, Blinking menu, Boxing menu, Split without replication menu, Pulsing menu*	

Table 6: Selection of graphical adaptive menus based on priority.

7.2. External Validity

The representativeness of the results and the size and complexity of tasks that might affect the generalization of the results must be considered. The representatives of the results could have been affected by the number of menus compared. We believe that the menus selected for this study can be considered as a baseline from which to obtain some indications of which properties attain better results as regards UX and performance. However, they cannot represent the whole design space of graphical adaptive menus. We plan to conduct more studies with more and different graphical adaptive menus in order to increase the reliability of our results. With regard to the size and complexity of the task, we decided to use relatively small tasks in order to enable the subjects to complete the whole experimental session within 30 minutes. Replications with different and more complex graphical menus are, nevertheless, necessary in order to study the effect of the graphical adaptive menus variables on the observed results.

Moreover, regarding the generalizability of the EEG-based findings, we acknowledge that the reliability of EEG results can be affected by various factors, including individual differences in brain activity and task familiarity. In this study, we focused on EEG because of its ability to provide real-time insights into cognitive processes. However, the results may not be fully generalizable to other populations or tasks, which is a common limitation in EEG studies. To validate the insights gained from EEG, we conducted a replication study [8] with a different participant pool. In this replication, we correlated EEG metrics with subjective measures obtained through questionnaires, and we found significant correlations between the two, strengthening the validity of our findings. This replication serves as initial evidence that our findings are reliable across different user groups, but further replication studies are needed.

Although this study focused specifically on EEG as a biomarker for user experience, we acknowledge that the generalizability of EEG to other biometric data may be limited. Methods such as eye tracking, heart rate monitoring, and facial recognition could offer complementary insights into user engagement and cognitive load, providing a more comprehensive view of the user experience. EEG, while valuable for real-time insights into cognitive processes, may not fully capture all aspects of user experience as other biometric methods might. Future studies should incorporate these additional methods to enhance the reproducibility and comparability of results.

7.3. Construct Validity

The measures used to obtain a quantitative evaluation of the subjects' UX (*i.e.*, the metrics obtained from the EEG data) and the participants' apprehension about being evaluated must be considered. We mitigated the threat regarding the measures used by employing the Bitbrain processing service, in which calculations are based on relevant previous research in EEG data analysis ([19, 80, 21, 17, 18, 81, 82]).

With regard to the participants' possible apprehension about being evaluated, this was largely avoided, since the participants were not evaluated as regards their performance using different menus. The participants were told that the error rate or speed was not being evaluated. In addition, the bias introduced into the study by expectations on the part of the experimenter was mitigated by the example menu that the participants used and also by interacting with the participants before each session.

7.4. Conclusion Validity

Finally, with regard to conclusion validity, the data collection and the validity of the statistical tests applied must be considered. In order to decrease the threat related to data collection, we systematically applied the same data-extraction procedure in each session with each participant. That is, before each session, and during the EEG device placement, we ensured that each electrode was in contact with the scalp and was placed correctly according to the guidelines [31]. With regard to the statistical tests, the most appropriate tests with which to test the hypothesis were performed [99]. The statistical tests were selected by considering the design of the experiment and the nature of the variables, and their assumptions. The data collection and the validity of the statistical tests applied must be considered. In the case of the statistical tests, rigorous tests were performed in order to statistically reject the null hypotheses. If differences were detected but were not significant, this was explicitly mentioned.

8. Conclusions

In this paper, we present an experiment carried out in order to obtain biometric data concerning the selection of menu items in a pool of twenty graphical adaptive menus so as to address two main research questions: RQ_1=Do graphical adaptive menus have a different influence on user experience? and RQ_2 =Do users' preferences for graphical adaptive menus relate to performance and user experience? The monitoring of brainwaves through the use of electroencephalography made it possible to obtain quantitative measures, such as cognitive load, engagement, attraction, and memorization, along with completion time, in order to evaluate the performance of the menus.

Our findings suggest that menus with atypical structures, different font styles, movement, and an increased area of menu items may have a more subjective and user-dependent effect on the user experience, while menus with temporality or color tend to be more objective and easier to use and learn. In order to contribute clarifying the debate regarding the many graphical adaptive menus for which no consensus has been reached and there is no common basis for comparison (*e.g.*, menu X says it is faster than menu Y, but menu Z does not say anything about X or Y but specifies that it is preferred to menu T), this experiment combines UX and performance analysis, embodied in four biometric measures and completion time, with the preferences expressed by the participants. The result is the ability to deduce which menu can optimize which performance measure by taking or not taking preference into account, which is crucial, since no significant correlation was found between UX/performance and preference, although there is some similarity between their parallel evolution.

In conclusion, only memorization and completion time were found to be significantly different as regards item selection among the twenty menu types, whereas none of the other measures did. However, the DTW distance was found to be significantly different as regards item selection for the four EEG measures, thus indicating that the participants reached some sort of consensus in the case of appreciating or not appreciating the various menu conditions. A weak or very weak correlation was found between the preference for and the UX/performance of the twenty menu conditions, thus showing the need to perform a performance-preference analysis in order to determine which conditions could optimize a UX or performance measure without deteriorating the corresponding preference.

These results highlight key insights for adaptive menu design: Boxing menus were the most effective in reducing cognitive load, making them suitable for minimizing mental effort. Glowing menus increased user engagement, though this did not necessarily improve performance. Highlighting and Ephemeral menus were the most visually attractive and improved memorization, suggesting their potential for interfaces that require strong recall. Additionally, Ephemeral menus yielded the fastest completion times, optimizing efficiency while maintaining UX quality. These findings reinforce the trade-offs between usability, cognitive effort, and user preference, offering practical insights for the development of effective adaptive interfaces.

Acknowledgement

This research has been funded by the State Research Agency under grant no. PID2022-140106NB-I00 (UCI-Adapt project) and Generalitat Valenciana (GVA) under grant CIAICO-/2021/303 (AKILA project). Daniel Gaspar-Figueiredo is funded by the GVA through a research fellowship with reference ACIF/2021/172. Jean Vanderdonckt is supported by the EU EIC Pathfinder-Awareness Inside challenge "Symbiotik" project (1 Oct. 2022-30 Sep. 2026) under Grant no. 101071147. These actions are also co-funded by the European Union through the Operational Program of the European Social Fund (ESF). We would also like to thank Quentin Sellier for helping us in the experiment.

References

[1] S. Abrahão, E. Insfran, A. Sluÿters, J. Vanderdonckt, Model-based intelligent user interface adaptation: challenges and future directions, Software and Systems Modeling 20 (2021) 1335–1349. doi:10.1007/s10270-021-00909-7.

[2] L. Findlater, K. Z. Gajos, Design space and evaluation challenges of adaptive graphical user interfaces, AI Magazine 30 (2009) 68–73. URL: https://doi.org/10.1609/aimag.v30i4.2268. doi:10.1609/aimag.v30i4.2268.

[3] K. Z. Gajos, K. Chauncey, The influence of personality traits and cognitive load on the use of adaptive user interfaces, in: G. A. Papadopoulos, T. Kuflik, F. Chen, C. Duarte, W. Fu (Eds.), Proceedings of the 22nd International Conference on Intelligent User Interfaces, IUI 2017, Limassol, Cyprus, March 13-16, 2017, Association for Computing Machinery, 2017, pp. 301–306. URL: https://doi.org/10.1145/3025171.3025192. doi:10.1145/3025171.3025192.

[4] T. Lavie, J. Meyer, Benefits and costs of adaptive user interfaces, International Journal of Human-Computer Studies 68 (2010) 508 – 524. URL: http://www.sciencedirect.com/science/article/pii/S1071581910000145. doi:https://doi.org/10.1016/j.ijhcs.2010.01.004.

[5] J. Nielsen, J. Levy, Measuring usability: Preference vs. performance, Commun. ACM 37 (1994) 66–75. URL: https://doi.org/10.1145/175276.175282. doi:10.1145/175276.175282.

[6] J. Vanderdonckt, S. Bouzit, G. Calvary, D. Chêne, Exploring a design space of graphical adaptive menus: Normal vs. small screens, ACM Trans. Interact. Intell. Syst. 10 (2020) 2:1–2:40. URL: https://doi.org/10.1145/3237190. doi:10.1145/3237190.

[7] L. Findlater, J. McGrenere, A comparison of static, adaptive, and adaptable menus, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '04, Association for Computing Machinery, New York, NY, USA, 2004, pp. 89–96. URL: http://doi.acm.org/10.1145/985692.985704. doi:10.1145/985692.985704.

[8] D. Gaspar-Figueiredo, S. Abrahao, E. Insfran, J. Vanderdonckt, Measuring user experience of adaptive user interfaces using eeg: A replication study, in: Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering, EASE '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 52–61. URL: https://doi.org/10.1145/3593434.3593452. doi:10.1145/3593434.3593452.

[9] J. Vanderdonckt, S. Bouzit, G. Calvary, D. Chêne, Cloud menus: A circular adaptive menu for small screens, in: 23rd International Conference on Intelligent User Interfaces, IUI '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 317–328. URL: http://doi.acm.org/10.1145/3172944.3172975. doi:10.1145/3172944.3172975.

[10] J. Bertin, Sémiologie graphique, paris, mouton/gauthier-villard, Réédition (2005) EHESS (1967).

[11] ISO, ISO/IEC 9241, Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts, standard, International Standard Organization, Geneva, 2018. URL: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:en.

[12] T. Zaki, M. N. Islam, 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. URL: https://www.sciencedirect.com/science/article/pii/S1574013721000150. doi:https://doi.org/10.1016/j.cosrev.2021.100375.

[13] S. Kujala, V. Roto, K. Väänänen-Vainio-Mattila, E. Karapanos, A. Sinnelä, Ux curve: A method for evaluating long-term user experience, Interacting with computers 23 (2011) 473–483.

[14] R. McCarney, J. Warner, S. Iliffe, R. van Haselen, M. Griffin, P. Fisher, The hawthorne effect: a randomised, controlled trial, BMC Med Res Methodol 7 (2007) 119–142. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-7-30.

[15] J. R. Bergstrom, S. Duda, D. Hawkins, M. McGill, Physiological response measurements, in: Eye tracking in user experience design, Elsevier, 2014, pp. 81–108.

[16] S. D. Kreibig, Autonomic nervous system activity in emotion: A review, Biological Psychology 84 (2010) 394–421. doi:https://doi.org/10.1016/j.biopsycho.2010.03.010.

[17] J. J. Allen, J. A. Coan, M. Nazarian, Issues and assumptions on the road from raw signals to metrics of frontal eeg asymmetry in emotion, Biological Psychology 67 (2004) 183–218. URL: https://www.sciencedirect.com/science/article/pii/S0301051104000377. doi:https://doi.org/10.1016/j.biopsycho.2004.03.007, frontal EEG Asymmetry, Emotion, and Psychopathology.

[18] E. Harmon-Jones, P. A. Gable, C. K. Peterson, The role of asymmetric frontal cortical activity in emotion-related phenomena: A review and update, Biological Psychology 84 (2010) 451–462. URL: https://www.sciencedirect.com/science/article/pii/S0301051109001823. doi:https://doi.org/10.1016/j.biopsycho.2009.08.010.

[19] W. Klimesch, Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis, Brain Research Reviews 29 (1999) 169–195. URL: https://www.sciencedirect.com/science/article/pii/S0165017398000563. doi:https://doi.org/10.1016/S0165-0173(98)00056-3.

[20] A. T. Pope, E. H. Bogart, D. S. Bartolome, Biocybernetic system evaluates indices of operator engagement in automated task, Biological Psychology 40 (1995) 187–195. URL: https://www.sciencedirect.com/science/article/pii/0301051195051163. doi:https://doi.org/10.1016/0301-0511(95)05116-3.

[21] P. J. Mikulka, M. W. Scerbo, F. G. Freeman, Effects of a biocybernetic system on vigilance performance, Human Factors 44 (2002) 654–664. URL: https://doi.org/10.1518/0018720024496944. doi:10.1518/0018720024496944. arXiv:https://doi.org/10.1518/0018720024496944, pMID: 12691372.
[22] N.-H. Liu, C.-Y. Chiang, H.-C. Chu, Recognizing the degree of human attention using EEG signals from mobile sensors, Sensors 13 (2013) 10273–10286. URL: https://doi.org/10.3390/s130810273. doi:10.3390/s130810273.
[23] M. Beauregard, J. Courtemanche, V. Paquette, Brain activity in near-death experiencers during a meditative state, Resuscitation 80 (2009) 1006–1010. URL: https://doi.org/10.1016/j.resuscitation.2009.05.006. doi:10.1016/j.resuscitation.2009.05.006.
[24] L. J. Nealis, Z. M. van Allen, J. M. Zelenski, Positive affect and cognitive restoration: Investigating the role of valence and arousal, PloS one 11 (2016) e0147275.
[25] Q. Q. Yi Ding, Yaqin Cao, V. G. Duffy, An exploratory study using electroencephalography (eeg) to measure the smartphone user experience in the short term, International Journal of Human–Computer Interaction 36 (2020) 1008–1021. URL: https://doi.org/10.1080/10447318.2019.1709330. doi:10.1080/10447318.2019.1709330. arXiv:https://doi.org/10.1080/10447318.2019.1709330.
[26] S. Haddad, K. Latifzadeh, S. Duraisamy, J. Vanderdonckt, O. Daassi, S. Belghith, L. A. Leiva, Good guis, bad guis: Affective evaluation of graphical user interfaces, in: Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 232–243. URL: https://doi.org/10.1145/3627043.3659549. doi:10.1145/3627043.3659549.
[27] K. Latifzadeh, N. Gozalpour, V. J. Traver, T. Ruotsalo, A. Kawala-Sterniuk, L. A. Leiva, Efficient decoding of affective states from video-elicited eeg signals: An empirical investigation, ACM Trans. Multimedia Comput. Commun. Appl. 20 (2024). URL: https://doi.org/10.1145/3663669. doi:10.1145/3663669.
[28] C. S. Nayak, A. C. Anilkumar, EEG Normal Waveforms, StatPearls Publishing, Treasure Island (FL), 2021.
[29] E. Başar, Brain function and oscillations. integrative brain function, Neurophysiology and cognitive processes 2 (1999). URL: https://link.springer.com/book/10.1007/978-3-642-59893-7.
[30] W. Klimesch, B. Schack, P. Sauseng, The functional significance of theta and upper alpha oscillations, Experimental psychology 52 (2005) 99–108.
[31] E. P. N. Committee, et al., Guideline thirteen: guidelines for standard electrode position nomenclature, J. Clin. Neurophysiol. 11 (1994) 111–113. URL: https://pubmed.ncbi.nlm.nih.gov/8195414/.
[32] N. Jamil, A. N. Belkacem, S. Ouhbi, A. Lakas, Noninvasive electroencephalography equipment for assistive, adaptive, and rehabilitative brain–computer interfaces: A systematic literature review, Sensors 21 (2021) 4754. URL: https://doi.org/10.3390/s21144754. doi:10.3390/s21144754.
[33] A. Colomer Granero, F. Fuentes-Hurtado, V. Naranjo Ornedo, J. Guixeres Provinciale, J. M. Ausín, M. Alcañiz Raya, A comparison of physiological signal analysis techniques and classifiers for automatic emotional evaluation of audiovisual contents, Frontiers in Computational Neuroscience 10 (2016). URL: https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2016.00074. doi:10.3389/fncom.2016.00074.
[34] G. R. Müller-Putz, P. Ofner, J. Pereira, A. Pinegger, A. Schwarz, M. Zube, U. Eck, B. Hessing, M. Schneiders, R. Rupp, Applying intuitive eeg-controlled grasp neuroprostheses in individuals with spinal cord injury: Preliminary results from the moregrasp clinical feasibility study, in: Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC '19, 2019, pp. 5949–5955. doi:10.1109/EMBC.2019.8856491.
[35] X. Wang, M. Hersche, B. Tömekce, B. Kaya, M. Magno, L. Benini, An accurate eegnet-based motor-imagery brain–computer interface for low-power edge computing, in: Proceedings of the IEEE International Symposium on Medical Measurements and Applications (MeMeA), MeMea '20, IEEE Press, 2020, pp. 1–6. doi:10.1109/MeMeA49120.2020.9137134.
[36] X. Wang, M. Hersche, M. Magno, L. Benini, Mi-bminet: An efficient convolutional neural network for motor imagery brain–machine interfaces with eeg channel selection, 2023. URL: https://arxiv.org/abs/2203.14592. arXiv:2203.14592.
[37] N. Yamamoto, J. Enjoji, I. Chanpornpakdi, R. Ozasa, F. Aki, T. Kimura, H. Ohshima, K. Yamazaki, Investigation on cnn-based state classification towards BCI application using amplitude probability density distribution, in: C. Stephanidis, M. Antona (Eds.), HCI International 2020 - Posters - 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings, Part I, volume 1224 of *Communications in Computer and Information Science*, Springer, 2020, pp. 683–689. URL: https://doi.org/10.1007/978-3-030-50726-8_89. doi:10.1007/978-3-030-50726-8_89.
[38] L. Cabañero Gómez, R. Hervás, I. González, V. Villarreal, Studying the generalisability of cognitive load measured with eeg, Biomedical Signal Processing and Control 70 (2021) 103032. URL: https://www.sciencedirect.com/science/article/pii/S1746809421006297. doi:https://doi.org/10.1016/j.bspc.2021.103032.

[39] L. Cabañero, R. Hervás, I. González, J. Fontecha, T. Mondéjar, J. Bravo, Characterisation of mobile-device tasks by their associated cognitive load through eeg data processing, Future Generation Computer Systems 113 (2020) 380–390. URL: https://www.sciencedirect.com/science/article/pii/S0167739X20305112. doi:https://doi.org/10.1016/j.future.2020.07.013.

[40] S. Han, Y. Kong, Z. Lei, Q. Zhao, Evaluation method of human-vehicle tactile interaction experience based on eeg, in: Proceedings of 12th International Conference on Intelligent Human-Machine Systems and Cybernetics, volume 2 of *IHMSC '20*, 2020, pp. 75–78. doi:10.1109/IHMSC49165.2020.10095.

[41] E. T. Martínez Beltrán, M. Quiles Pérez, S. López Bernal, G. Martínez Pérez, A. Huertas Celdrán, Safecar: A brain–computer interface and intelligent framework to detect drivers' distractions, Expert Syst. Appl. 203 (2022). URL: https://doi.org/10.1016/j.eswa.2022.117402. doi:10.1016/j.eswa.2022.117402.

[42] A. Sluÿters, Q. Sellier, J. Vanderdonckt, V. Parthiban, P. Maes, Consistent, continuous, and customizable mid-air gesture interaction for browsing multimedia objects on large displays, International Journal of Human–Computer Interaction (2022) 1–32. URL: https://doi.org/10.1080/10447318.2022.2078464. doi:10.1080/10447318.2022.2078464. arXiv:https://doi.org/10.1080/10447318.2022.2078464.

[43] K. Z. Gajos, M. Czerwinski, D. S. Tan, D. S. Weld, Exploring the design space for adaptive graphical user interfaces, in: Proceedings of the Working Conference on Advanced Visual Interfaces, AVI '06, Association for Computing Machinery, New York, NY, USA, 2006, pp. 201–208. URL: http://doi.acm.org/10.1145/1133265.1133306. doi:10.1145/1133265.1133306.

[44] G. Bailly, E. Lecolinet, L. Nigay, Visual menu techniques, ACM Comput. Surv. 49 (2017) 60:1–60:41. URL: https://doi.org/10.1145/3002171. doi:10.1145/3002171.

[45] S. Bouzit, G. Calvary, D. Chêne, J. Vanderdonckt, A comparison of shortcut and step-by-step adaptive menus for smartphones, in: , 2016, p. . doi:10.14236/ewic/HCI2016.26.

[46] J. Park, S. H. Han, Y. S. Park, Y. Cho, Usability of adaptable and adaptive menus, in: N. Aykin (Ed.), Usability and Internationalization. HCI and Culture, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 405–411. URL: https://doi.org/10.1007/978-3-540-73287-7_49.

[47] L. Findlater, K. Moffatt, J. McGrenere, J. Dawson, Ephemeral adaptation: The use of gradual onset to improve menu selection performance, in: Proceedings of the ACM Conference on Human Factors in Computing Systems, CHI '09, Association for Computing Machinery, New York, NY, USA, 2009, pp. 1655–1664. URL: http://doi.acm.org/10.1145/1518701.1518956. doi:10.1145/1518701.1518956.

[48] L. Jiang, Y.-H. Chen, Menu design on small display user interfaces: Measuring the influence of menu type, number of preview items, and menu breadth on navigation efficiency, International Journal of Human–Computer Interaction 38 (2022) 631–645. URL: https://doi.org/10.1080/10447318.2021.1954781. doi:10.1080/10447318.2021.1954781.

[49] L. Findlater, K. Z. Gajos, Design space and evaluation challenges of adaptive graphical user interfaces, AI Magazine 30 (2009) 68–73. URL: http://www.aaai.org/ojs/index.php/aimagazine/article/view/2268.

[50] J. Siegmund, C. Kästner, S. Apel, C. Parnin, A. Bethmann, T. Leich, G. Saake, A. Brechmann, Understanding understanding source code with functional magnetic resonance imaging, in: Proceedings of the 36th international conference on software engineering, 2014, pp. 378–389.

[51] S. Fakhoury, Y. Ma, V. Arnaoudova, O. Adesope, The effect of poor source code lexicon and readability on developers' cognitive load, in: Proc. Int. Conf. Program Comprehension (ICPC), IEEE, 2018, pp. 286–28610.

[52] B. Weber, T. Fischer, R. Riedl, Brain and autonomic nervous system activity measurement in software engineering: A systematic literature review, Journal of Systems and Software 178 (2021) 110946. URL: https://www.sciencedirect.com/science/article/pii/S0164121221000431. doi:https://doi.org/10.1016/j.jss.2021.110946.

[53] R. Feldt, R. Torkar, L. Angelis, M. Samuelsson, Towards individualized software engineering: empirical studies should collect psychometrics, in: Proceedings of the 2008 international workshop on Cooperative and human aspects of software engineering, 2008, pp. 49–52.

[54] A. P. O. S. Vermeeren, E. L.-C. Law, V. Roto, M. Obrist, J. Hoonhout, K. Väänänen-Vainio-Mattila, User experience evaluation methods: Current state and development needs, in: Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, NordiCHI '10, Association for Computing Machinery, New York, NY, USA, 2010, p. 521–530. URL: https://doi.org/10.1145/1868914.1868973. doi:10.1145/1868914.1868973.

[55] A. Hinderks, A. Meiners, F. J. D. Mayo, J. Thomaschewski, Interpreting the results from the user experience questionnaire (UEQ) using importance-performance analysis (IPA), in: A. Bozzon, F. J. D. Mayo, J. Filipe (Eds.), Proceedings of the 15th International Conference on Web Information Systems and Technologies, WEBIST 2019, ScitePress, 2019, pp. 388–395. URL: https://doi.org/10.5220/0008366503880395. doi:10.5220/0008366503880395.

[56] S. G. Hart, L. E. Staveland, Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research, in: P. A. Hancock, N. Meshkati (Eds.), Human Mental Workload, volume 52 of *Advances in Psychol-*

ogy, North-Holland, 1988, pp. 139 – 183. URL: http://www.sciencedirect.com/science/article/pii/S0166411508623869. doi:https://doi.org/10.1016/S0166-4115(08)62386-9.
[57] P. Schmutz, S. Heinz, Y. Métrailler, K. Opwis, Cognitive load in ecommerce applications—measurement and effects on user satisfaction, Advances in Human-Computer Interaction (2009).
[58] E. L.-C. Law, P. van Schaik, V. Roto, Attitudes towards user experience (ux) measurement, International Journal of Human-Computer Studies 72 (2014) 526–541. URL: https://www.sciencedirect.com/science/article/pii/S1071581913001304.
[59] D. Graziotin, P. Lenberg, R. Feldt, S. Wagner, Psychometrics in behavioral software engineering: A methodological introduction with guidelines, ACM Transactions on Software Engineering and Methodology 31 (2022) 7. doi:10.1145/3469888.
[60] H. Lee, J. Lee, S. Seo, Brain response to good and bad design, in: J. A. Jacko (Ed.), Human-Computer Interaction. New Trends, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 111–120.
[61] X. Hou, F. Trapsilawati, Y. Liu, O. Sourina, C.-H. Chen, W. Mueller-Wittig, W. T. Ang, 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, 2017, pp. 885–897.
[62] Y. Ding, Y. Cao, Q. Qu, V. G. Duffy, An exploratory study using electroencephalography (eeg) to measure the smartphone user experience in the short term, International Journal of Human–Computer Interaction 36 (2020) 1008–1021.
[63] A. J. Ko, T. D. LaToza, M. M. Burnett, A practical guide to controlled experiments of software engineering tools with human participants, Empirical Software Engineering 20 (2015) 110–141. URL: https://doi.org/10.1007/s10664-013-9279-3. doi:10.1007/s10664-013-9279-3.
[64] Y. Mashiko, V. R. Basili, Using the GQM paradigm to investigate influential factors for software process improvement, J. Syst. Softw. 36 (1997) 17–32. URL: https://doi.org/10.1016/0164-1212(95)00194-8. doi:10.1016/0164-1212(95)00194-8.
[65] A. Apraiz Iriarte, G. Lasa, M. Mazmela, Evaluating user experience with physiological monitoring: A systematic literature review, Dyna New Technologies 8 (2021) 21. doi:https://doi.org/10.6036/NT10072.
[66] A. Ponsard, K. Ardekani, K. Zhang, F. Ren, M. Negulescu, J. McGrenere, Twist and pulse: Ephemeral adaptation to improve icon selection on smartphones, in: Proceedings of the 41st Graphics Interface Conference, GI '15, Canadian Information Processing Society, Toronto, Ont., Canada, Canada, 2015, pp. 219–222. URL: http://dl.acm.org/citation.cfm?id=2788890.2788929.
[67] D.-S. Lee, W. C. Yoon, Quantitative results assessing design issues of selection-supportive menus, International Journal of Industrial Ergonomics 33 (2004) 41 – 52. URL: http://www.sciencedirect.com/science/article/pii/S0169814103001331. doi:https://doi.org/10.1016/j.ergon.2003.07.004.
[68] A. Cockburn, C. Gutwin, S. Greenberg, A predictive model of menu performance, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '07, Association for Computing Machinery, New York, NY, USA, 2007, pp. 627–636. URL: http://doi.acm.org/10.1145/1240624.1240723. doi:10.1145/1240624.1240723.
[69] S. Bouzit, D. Chêne, G. Calvary, From appearing to disappearing ephemeral adaptation for small screens, in: Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design, OzCHI '14, Association for Computing Machinery, New York, NY, USA, 2014, pp. 41–48. URL: http://doi.acm.org/10.1145/2686612.2686619. doi:10.1145/2686612.2686619.
[70] A. Sears, B. Shneiderman, Split menus: Effectively using selection frequency to organize menus, ACM Trans. Comput.-Hum. Interact. 1 (1994) 27–51. URL: http://doi.acm.org/10.1145/174630.174632. doi:10.1145/174630.174632.
[71] A. Roudaut, G. Bailly, E. Lecolinet, L. Nigay, Leaf menus: Linear menus with stroke shortcuts for small handheld devices, in: T. Gross, J. Gulliksen, P. Kotzé, L. Oestreicher, P. Palanque, R. O. Prates, M. Winckler (Eds.), Human-Computer Interaction – INTERACT 2009, Springer, Berlin, Heidelberg, 2009, pp. 616–619. doi:https://doi.org/10.1007/978-3-642-03655-2_69.
[72] P. Baudisch, D. Tan, M. Collomb, D. Robbins, K. Hinckley, K. Hinckley, M. Agrawala, S. Zhao, G. Ramos, Phosphor: Explaining transitions in the user interface using afterglow effects, in: Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology, UIST '06, Association for Computing Machinery, New York, NY, USA, 2006, pp. 169–178. URL: http://doi.acm.org/10.1145/1166253.1166280. doi:10.1145/1166253.1166280.
[73] E. Tanvir, J. Cullen, P. Irani, A. Cockburn, Aamu: Adaptive activation area menus for improving selection in cascading pull-down menus, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08, Association for Computing Machinery, New York, NY, USA, 2008, pp. 1381–1384. URL: http://doi.acm.org/10.1145/1357054.1357270. doi:10.1145/1357054.1357270.
[74] Bitbrain, Sennslab: Definition of analysis metrics in sennsmetrics, https://help.bitbrain.com/sennsmetrics-definition-of-analysis-metrics, 2020.

[75] P. Antonenko, F. Paas, R. Grabner, T. van Gog, Using electroencephalography to measure cognitive load, Educational Psychology Review 22 (2010) 425–438. URL: https://link.springer.com/article/10.1007/s10648-010-9130-y. doi:10.1007/s10648-010-9130-y.
[76] J. Lehmann, M. Lalmas, E. Yom-Tov, G. Dupret, Models of user engagement, in: J. Masthoff, B. Mobasher, M. C. Desmarais, R. Nkambou (Eds.), Proceedings of ACM Int. Conf. on User Modeling, Adaptation, and Personalization, UMAP '20'12, volume 7379 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 164–175. URL: https://link.springer.com/chapter/10.1007/978-3-642-31454-4_14. doi:10.1007/978-3-642-31454-4_14.
[77] C. Laming, K. Mason, A robust implicit measure of facial attractiveness discrimination, Social cognitive and affective neuroscience 14 (2019) 737–746. URL: https://academic.oup.com/scan/article/14/7/737/5520404. doi:10.1093/scan/nsz043.
[78] E. Noh, K. Liao, M. V. Mollison, T. Curran, V. R. d. Sa, Single-trial eeg analysis predicts memory retrieval and reveals source-dependent differences, Frontiers in Human Neuroscience 12 (2018). URL: https://www.frontiersin.org/articles/10.3389/fnhum.2018.00258. doi:10.3389/fnhum.2018.00258.
[79] Bitbrain, Sennsmetrics: Analysis software of biometrics, https://www.bitbrain.com/neurotechnology-products/software/sennsmetrics, 2020.
[80] A.-M. Brouwer, M. A. Hogervorst, J. B. F. van Erp, T. Heffelaar, P. H. Zimmerman, R. Oostenveld, Estimating workload using EEG spectral power and ERPs in the n-back task, Journal of Neural Engineering 9 (2012) 045008. URL: https://doi.org/10.1088/1741-2560/9/4/045008. doi:10.1088/1741-2560/9/4/045008.
[81] W. Klimesch, M. Doppelmayr, H. Schimke, B. Ripper, Theta synchronization and alpha desynchronization in a memory task, Psychophysiology 34 (1997) 169–176. doi:https://doi.org/10.1111/j.1469-8986.1997.tb02128.x.
[82] N. M. Long, J. F. Burke, M. J. Kahana, Subsequent memory effect in intracranial and scalp eeg, NeuroImage 84 (2014) 488–494. doi:https://doi.org/10.1016/j.neuroimage.2013.08.052.
[83] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing 26 (1978) 43–49. doi:10.1109/TASSP.1978.1163055.
[84] M. Müller, Dynamic time warping, in: Information Retrieval for Music and Motion, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 69–84. URL: https://doi.org/10.1007/978-3-540-74048-3_4.
[85] S. Bouzit, G. Calvary, D. Chêne, J. Vanderdonckt, A design space for engineering graphical adaptive menus, in: Proceedings of the 8th ACM SIGCHI Symposium on Engineering Interactive Computing Systems, EICS '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 239–244. URL: http://doi.acm.org/10.1145/2933242.2935874. doi:10.1145/2933242.2935874.
[86] S. S. Shapiro, M. B. Wilk, An analysis of variance test for normality (complete samples), Biometrika 52 (1965) 591–611. doi:10.1093/biomet/52.3-4.591.
[87] S. S. Shapiro, M. B. Wilk, H. J. Chen, A comparative study of various tests for normality, Journal of the American Statistical Association 63 (1968) 1343–1372. doi:10.1080/01621459.1968.10480932.
[88] H. Levene, Robust tests for equality of variances, in: I. Olkin, H. H. et al. (Eds.), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, Stanford University Press, Palo Alto, CA, USA, 1960, pp. 278–292.
[89] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of the American Statistical Association 32 (1937) 675–701. doi:10.1080/01621459.1937.10503522.
[90] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Routledge, New York, NY, USA, 2013. URL: https://www.taylorfrancis.com/books/mono/10.4324/9780203771587/statistical-power-analysis-behavioral-sciences-jacob-cohen. doi:https://doi.org/10.4324/9780203771587.
[91] A. R. Lahitani, A. E. Permanasari, N. A. Setiawan, Cosine similarity to determine similarity measure: Study case in online essay assessment, in: 2016 4th International Conference on Cyber and IT Service Management, IEEE, 2016, pp. 1–6.
[92] N. I. o. S. U.S. Commerce Department, S. E. D. Technology, Cosine distance, cosine similarity, angular cosine distance, angular cosine similarity, 2017. URL: https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/cosdist.htmm.
[93] J. A. Martilla, J. C. James, Importance-performance analysis, Journal of Marketing 41 (1977) 77–79. URL: https://doi.org/10.1177/002224297704100112. doi:10.1177/002224297704100112. arXiv:https://doi.org/10.1177/002224297704100112.
[94] A. Hinderks, F. J. D. Mayo, A. Meiners, J. Thomaschewski, Applying importance-performance analysis (IPA) to interpret the results of the user experience questionnaire (UEQ), J. Web Eng. 19 (2020) 243–266. URL: https://doi.org/10.13052/jwe1540-9589.1926. doi:10.13052/jwe1540-9589.1926.
[95] S. R. Magal, N. M. Levenburg, Using importance-performance analysis to evaluate e-business strategies among small businesses, in: Proceedings of 38th Hawaii International Conference on System Sciences, HICSS-38, IEEE

Computer Society, 2005, p. . URL: https://doi.org/10.1109/HICSS.2005.661. doi:10.1109/HICSS.2005.661.

[96] P. Emerson, The original Borda count and partial voting, Social Choice and Welfare 40 (2013) 353–358. URL: https://doi.org/10.1007/s00355-011-0603-9.

[97] C. Wohlin, M. Höst, K. Henningsson, Empirical research methods in software engineering, in: Empirical Methods and Studies in Software Engineering, Springer Berlin Heidelberg, 2003, pp. 7–23. URL: https://doi.org/10.1007/978-3-540-45143-3_2. doi:10.1007/978-3-540-45143-3_2.

[98] A. Alda, A. Ortiz, N. Torreblanca, L. Montesano, J. Minguez, What is minimal eeg? user centered and reliable eeg headsets for real-world applications, 2019.

[99] M. K. D., Applied statistics for software managers, Applied Statistics for Software Managers (2002). URL: https://cir.nii.ac.jp/crid/1573668924056253312.

Appendix A. Descriptive Analyses

Appendix A.1. COGNITIVE-LOAD

Menu	Minimum	Maximum	Mean	Median	Std. dev.	Coeff. var.	Skewness	Kurtosis
Activation area	3.49	54.52	28.76	27.92	11.32	39%	0.20	0.04
Blinking	6.82	59.97	26.24	25.26	13.52	51%	0.72	−0.13
Bolding	3.21	64.33	29.37	28.56	14.41	49%	0.81	0.49
Boxing	9.67	66.14	27.76	24.07	15.09	54%	0.76	−0.33
Colored	6.58	68.09	26.17	23.13	14.45	55%	1.13	0.94
Ephemeral	5.09	72.46	28.57	24.69	15.32	53%	0.90	0.40
Glowing	7.73	69.22	29.16	25.83	14.01	48%	0.88	0.51
Grayscaling	8.13	69.41	29.52	27.77	13.97	47%	0.54	0.23
Highlighting	3.90	53.51	29.32	28.25	11.13	37%	0.14	−0.19
Italicizing	7.35	59.01	27.26	26.05	10.46	38%	0.86	1.48
Leaf	10.63	74.62	32.28	30.85	12.23	37%	0.98	2.17
Morphing	6.15	54.69	25.67	26.14	12.12	47%	0.15	−0.47
Out-of-context	8.44	58.31	28.64	26.88	12.89	45%	0.63	0.03
Pulsing	6.85	65.62	28.87	27.81	13.55	46%	0.67	0.19
Rotating	10.00	54.65	27.73	25.78	11.63	41%	0.56	−0.26
Split without repl.	1.46	59.46	26.06	23.44	11.65	44%	0.51	0.46
Static	4.73	51.62	24.22	21.96	12.08	49%	0.47	−0.73
Temporal	3.99	77.2	28.93	26.52	14.17	48%	1.08	2.24
Twisting	6.90	49.06	26.76	26.83	10.67	39%	0.18	−0.16
Underlining	6.68	53.47	26.11	26.98	10.95	41%	0.33	0.12

Table A.7: Descriptive analysis of the COGNITIVE-LOAD value for all MENU-TYPE.

Menu	Minimum	Maximum	Mean	Median	Std. dev.	Coeff. var.	Skewness	Kurtosis
Activation area	596.66	1522.59	799.98	696.22	226.82	28%	1.7	2.5
Blinking	499.95	1277.42	694.13	649.97	162.35	23%	1.7	3.2
Bolding	576.91	1681.81	810.30	766.56	234.01	29%	2.3	5.6
Boxing	485.30	1232.15	696.91	663.38	158.34	23%	2.5	1.5
Colored	392.59	1020.74	618.60	564.16	166.81	27%	1.2	0.82
Ephemeral	448.04	1416.50	669.15	602.04	189.79	28%	2.1	5.6
Glowing	451.48	1150.83	652.20	581.93	153.41	24%	1.2	1.2
Grayscaling	605.78	1477.32	810.01	773.22	177.44	22%	2.0	5.0
Highlighting	550.97	1883.38	748.00	661.58	241.66	32%	3.1	12.0
Italicizing	712.42	1940.24	968.47	884.56	262.60	27%	2.0	4.8
Leaf	676.55	1389.39	907.96	857.81	159.32	18%	1.1	0.87
Morphing	566.84	2368.10	786.24	661.02	328.90	42%	3.3	14.0
Out-of-context	474.20	1126.61	671.22	618.11	160.26	24%	1.1	0.97
Pulsing	544.78	1273.81	721.81	685.69	161.49	22%	1.9	3.9
Rotating	640.08	2692.02	893.53	778.55	346.83	39%	3.9	19.0
Split without repl.	521.58	1251.92	730.05	661.05	190.00	26%	1.5	1.6
Static	491.89	1460.30	749.76	671.76	210.21	28%	1.8	3.3
Temporal	490.83	1653.98	744.48	696.62	220.25	30%	2.5	7.8
Twisting	536.65	2003.14	693.98	641.83	231.96	33%	4.8	27.0
Underlining	566.98	2289.01	823.05	759.38	273.56	33%	4.1	22.0

Table A.8: Descriptive analysis of the COGNITIVE-LOAD DTW distance for all MENU-TYPE.

Group 1	Group 2	Rank sum diff.	p-value	Cohen's d	Glass's Δ	Interp.
Activation area	Colored	244.0	.0008***	0.90	0.79	Large
”	Italicizing	−205.0	.0203**	0.68	0.74	Medium
Blinking	Italicizing	−322.0	<.0001****	1.25	1.69	Large
”	Leaf	−298.0	<.0001****	1.33	1.32	Large
”	Rotating	−228.0	.0031**	0.73	1.23	Medium
Bolding	Colored	292.0	<.0001****	0.93	0.81	Large
”	Ephemeral	228.0	.0031**	0.66	0.60	Medium
”	Glowing	219.0	.0066**	0.80	0.69	Large
”	Out-of-context	209.0	.0149*	0.69	0.59	Medium
”	Twisting	195.0	.0434*	0.49	0.49	Small
Boxing	Italicizing	−307.0	<.0001****	1.24	1.75	Large
”	Leaf	−283.0	<.0001****	1.33	1.33	Large
”	Rotating	−213.0	.0108*	0.73	1.24	Medium
Colored	Grayscaling	−316.0	<.0001****	1.10	1.14	Large
”	Italicizing	−449.0	<.0001****	1.58	2.08	Large
”	Leaf	−425.0	<.0001****	1.77	1.73	Large
”	Rotating	−355.0	<.0001****	1.01	1.64	Large
”	Underlining	−355.0	<.0001****	0.90	1.22	Large
Ephemeral	Grayscaling	−252.0	.0004***	0.76	0.74	Medium
”	Italicizing	−385.0	<.0001****	1.30	1.57	Large
”	Leaf	−361.0	<.0001****	1.36	1.25	Large
”	Rotating	−291.0	<.0001****	0.80	1.18	large
”	Underlining	−232.0	.0022**	0.65	0.81	Medium
Glowing	Grayscaling	−243.0	.0008***	0.95	1.03	Large
”	Italicizing	−376.0	<.0001****	1.46	2.06	Large
”	Leaf	−352.0	<.0001****	1.64	1.67	Large
”	Rotating	−282.0	<.0001****	0.90	1.58	Large
”	Underlining	−223.0	.0048**	0.77	1.11	Medium
Grayscaling	Out-of-context	233.0	.0020**	0.86	0.82	Large
”	Twisting	219.0	.0066**	0.59	0.69	Medium
Highlighting	Italicizing	−271.0	<.0001****	0.87	0.90	Large
”	Leaf	−247.0	<.0001****	0.78	0.66	Medium
Italicizing	Morphing	260.0	.0002***	0.61	0.69	Medium
”	Out-of-context	366.0	<.0001****	1.36	1.12	Large
”	Pulsing	277.0	<.0001****	1.12	0.93	Large
”	Split without repl.	273.0	<.0001****	1.03	0.90	Large
”	Static	277.0	<.0001****	0.91	0.82	Large
”	Temporal	259.0	.0002***	0.92	0.85	Large
”	Twisting	352.0	<.0001****	1.14	1.04	Large
Leaf	Morphing	236.0	.0016**	0.47	0.76	Small
”	Out-of-context	342.0	<.0001****	1.48	1.49	Large
”	Pulsing	253.0	.0003***	1.16	1.17	Large
”	Split without repl.	249.0	.0005***	1.01	1.12	Large
”	Static	253.0	.0003***	0.84	0.99	Large
”	Temporal	235.0	.0017**	0.85	1.03	Large
”	Twisting	328.0	<.0001****	1.07	1.34	Large
Out-of-context	Rotating	−272.0	<.0001****	1.10	1.39	Large
”	Underlining	−213.0	.0108*	0.75	0.95	Medium
Rotating	Twisting	258.0	.0002***	0.67	0.57	Medium
Twisting	Underlining	−199.0	.0322*	0.50	0.55	Medium

Table A.9: Dunn's pair-wise comparison test with multiplicity adjusted p-value: significant differences between COGNITIVE-LOAD DTW distances of the various MENU-TYPE. Interpretation of Cohen's d [90]: $d \geq 0.8$=Large, $0.5 \leq d < 0.8$=Medium, $0.2 \leq d < 0.5$=Small.

Appendix A.2. Engagement

Menu	Minimum	Maximum	Mean	Median	Std. dev.	Coeff. var.	Skewness	Kurtosis
Activation area	4.38	55.19	33.18	34.26	10.9	32%	−0.33	0.098
Blinking	11.01	70.5	34.16	31.7	13.76	40%	0.98	0.92
Bolding	5.29	59.43	33.85	33.39	14.12	41%	0.053	−0.48
Boxing	9.18	72.44	35.33	35.09	13.12	37%	0.43	0.63
Colored	8.12	60.97	34.12	32.07	14.64	42%	0.16	−1.04
Ephemeral	11.32	58.86	33.91	36.16	10.97	32%	−0.074	−0.47
Glowing	4.42	61.52	30.49	30.51	14.18	46%	0	−1
Grayscaling	4.92	60.45	32.31	30.04	12.41	38%	0.48	0.03
Highlighting	11.09	60.31	33.80	33.5	12.76	37%	0.16	−0.59
Italicizing	15.72	76.04	35.76	33.64	12.34	34%	0.94	1.77
Leaf	7.62	60.91	34.23	33.94	11.67	34%	0.25	0.0069
Morphing	9.50	96.8	32.82	30.28	16.24	49%	2	5
Out-of-context	5.25	54.65	32.00	32.22	11.35	35%	−0.33	−0.036
Pulsing	10.37	50.73	30.76	30.67	9.99	32%	0.14	−0.56
Rotating	12.52	72.73	35.62	34.7	13.45	37%	0.62	0.28
Split without repl.	13.93	70.55	31.98	30.52	12.61	39%	0.75	0.88
Static	8.55	68.32	34.02	34.67	12.4	36%	0.0099	0.73
Temporal	7.17	54.11	33.04	32.44	11.66	35%	−0.0005	−0.36
Twisting	9.74	75.85	34.08	33.32	12.38	36%	0.95	2.49
Underlining	8.48	57.6	31.76	30.72	12.32	38%	0.05	−0.52

Table A.10: Descriptive analysis of the Engagement value for all Menu-Type.

Menu	Minimum	Maximum	Mean	Median	Std. dev.	Coeff. var.	Skewness	Kurtosis
Activation area	567.66	1727.34	821.63	737.20	240.36	29%	2.1	4.6
Blinking	463.42	1135.45	667.80	626.38	135.14	20%	1.2	2.2
Bolding	569.45	1638.70	831.70	763.37	250.81	30%	1.8	3.0
Boxing	461.44	1066.13	641.07	600.67	136.96	21%	1.3	1.8
Colored	417.01	986.77	586.96	558.90	137.77	23%	1.2	1.3
Ephemeral	426.55	856.40	611.11	589.07	108.97	18%	0.27	−0.73
Glowing	446.35	1075.10	628.19	568.75	154.81	25%	1.3	1.3
Grayscaling	495.00	1608.62	720.22	635.77	230.18	32%	2.1	5.0
Highlighting	555.85	1723.18	773.26	727.94	212.01	27%	2.7	9.8
Italicizing	661.74	2083.23	959.77	862.65	286.70	30%	2.6	7.3
Leaf	750.24	1488.74	976.27	899.46	203.18	21%	1.0	−0.035
Morphing	584.86	2529.75	890.78	748.51	382.35	43%	2.7	8.5
Out-of-context	422.01	1167.27	610.03	554.00	154.56	25%	1.9	4.0
Pulsing	475.50	1046.17	652.93	610.72	150.78	23%	1.1	0.56
Rotating	609.57	1963.44	843.28	760.01	269.95	32%	2.9	9.0
Split without repl.	479.92	1609.15	695.82	643.30	246.82	35%	2.3	5.8
Static	508.11	1728.06	727.98	654.56	254.99	35%	2.8	8.7
Temporal	448.84	1910.51	633.49	580.96	235.54	37%	4.3	23
Twisting	491.66	1785.27	658.39	603.70	209.21	32%	4.3	22
Underlining	568.15	2244.80	824.80	736.13	288.11	35%	3.5	15

Table A.11: Descriptive analysis of the Engagement DTW distance for all Menu-Type.

Group 1	Group 2	Rank sum diff.	*p*-value	Cohen's *d*	Glass's Δ	Interp.
Activation area	Blinking	195.0	.0434*	0.79	0.64	Medium
”	Boxing	245.0	.0007****	0.92	0.75	Large
”	Colored	322.0	<.0001****	1.20	0.97	Large
”	Ephemeral	271.0	<.0001****	1.13	0.87	Large
”	Glowing	258.0	.0002***	0.96	0.80	Large
”	Out-of-context	289.0	<.0001****	1.04	0.88	Large
”	Split without repl.	201.0	.0277*	0.51	0.52	Medium
”	Temporal	277.0	<.0001****	0.79	0.78	Medium
”	Twisting	227.0	.0034**	0.72	0.68	Medium
Blinking	Bolding	−207.0	.0174*	0.81	1.21	Large
”	Italicizing	−333.0	<.0001****	1.30	2.16	Large
”	Leaf	−363.0	<.0001****	1.78	2.28	Large
”	Morphing	−227.0	.0034**	0.77	1.65	Medium
”	Rotating	−249.0	.0005***	0.81	1.29	Large
Bolding	Boxing	257.0	.0002***	0.94	0.76	Large
”	Colored	334.0	<.0001****	1.20	0.97	Large
”	Ephemeral	283.0	<.0001****	1.14	0.88	Large
”	Glowing	270.0	<.0001****	0.97	0.81	Large
”	Out-of-context	301.0	<.0001****	1.06	0.88	Large
”	Pulsing	249.0	.0005***	0.86	0.71	Large
”	Split without repl.	213.0	.0108*	0.54	0.54	Medium
”	Temporal	289.0	<.0001****	0.81	0.79	Large
”	Twisting	239.0	.0012**	0.75	0.69	Medium
Boxing	Highlighting	−210.0	<.0001****	0.73	0.96	Medium
”	Italicizing	−383.0	<.0001****	1.41	2.32	Large
”	Leaf	−413.0	<.0001****	1.93	2.44	Large
”	Morphing	−277.0	<.0001****	0.87	1.82	Large
”	Rotating	−299.0	<.0001****	0.94	1.47	Large
”	Underlining	−239.0	.0012**	0.81	1.34	Large
Colored	Highlighting	−287.0	<.0001****	1.03	1.34	Large
”	Italicizing	−460.0	<.0001****	1.65	2.70	Large
”	Leaf	−490.0	<.0001****	2.24	2.81	Large
”	Morphing	−354.0	<.0001****	1.05	2.20	Large
”	Rotating	−376.0	<.0001****	1.19	1.85	Large
”	Underlining	−316.0	<.0001****	1.05	1.72	Large
Ephemeral	Highlighting	−236.0	.0016**	0.96	1.48	Large
”	Italicizing	−409.0	<.0001****	1.60	3.20	Large
”	Leaf	−439.0	<.0001****	2.24	3.34	Large
”	Morphing	−303.0	<.0001****	0.99	2.56	Large
”	Rotating	−325.0	<.0001****	1.26	2.12	Large
”	Underlining	−265.0	.0001***	0.98	1.96	Large
Glowing	Highlighting	−223.0	.0048**	0.78	0.93	Medium
”	Italicizing	−396.0	<.0001****	1.43	2.14	Large
”	Leaf	−426.0	<.0001****	1.92	2.24	Large
”	Morphing	−290.0	<.0001****	0.90	1.69	Large
”	Rotating	−312.0	<.0001****	0.97	1.38	Large
”	Underlining	−252.0	.0004***	0.85	1.27	Large

Table A.12: Dunn's pair-wise comparison test with multiplicity adjusted *p*-value: significant differences between Engagement DTW distances of the various Menu-Type. Interpretation of Cohen's *d* [90]: $d \geq 0.8$=Large, $0.5 \leq d < 0.8$=Medium, $0.2 \leq d < 0.5$=Small. (Part 1/2)

Group 1	Group 2	Rank sum diff.	p-value	Cohen's d	Glass's Δ	Interp.
Grayscaling	Italicizing	−285.0	$<.0001^{****}$	0.92	1.04	Large
”	Leaf	−315.0	$<.0001^{****}$	1.18	1.11	Large
”	Rotating	−201.0	$.0277^{*}$	0.49	0.53	Small
Highlighting	Leaf	−203.0	$.0237^{*}$	0.97	0.95	Large
”	Out-of-context	254.0	$.0003^{***}$	0.87	0.76	Large
”	Pulsing	202.0	$.0256^{*}$	0.65	0.56	Medium
”	Temporal	242.0	$.0009^{***}$	0.62	0.66	Medium
Italicizing	Out-of-context	427.0	$<.0001^{****}$	1.51	1.21	Large
”	Pulsing	375.0	$<.0001^{****}$	1.33	1.06	Large
”	Split without repl.	339.0	$<.0001^{****}$	0.98	0.91	Large
”	Static	272.0	$<.0001^{****}$	0.85	0.80	Large
”	Temporal	415.0	$<.0001^{****}$	1.24	1.13	Large
”	Twisting	365.0	$<.0001^{****}$	1.20	1.05	Large
Leaf	Out-of-context	457.0	$<.0001^{****}$	2.02	1.80	Large
”	Pulsing	405.0	$<.0001^{****}$	1.80	1.59	Large
”	Split without repl.	369.0	$<.0001^{****}$	1.23	1.37	Large
”	Static	302.0	$<.0001^{****}$	1.07	1.22	Large
”	Temporal	445.0	$<.0001^{****}$	1.55	1.68	Large
”	Twisting	395.0	$<.0001^{****}$	1.54	1.56	Large
Morphing	Out-of-context	321.0	$<.0001^{****}$	0.96	0.73	Large
”	Pulsing	269.0	$<.0001^{****}$	0.81	0.62	Large
”	Split without repl.	233.0	$.0020^{**}$	0.60	0.51	Medium
”	Temporal	309.0	$<.0001^{****}$	0.81	0.67	Large
”	Twisting	259.0	$.0002^{***}$	0.75	0.60	Medium
Out-of-context	Rotating	−343.0	$<.0001^{****}$	1.05	1.50	Large
”	Underlining	−283.0	$<.0001^{****}$	0.92	1.38	Large
Pulsing	Rotating	−291.0	$<.0001^{****}$	0.86	1.25	Large
”	Underlining	−231.0	$.0024^{**}$	0.74	1.13	Medium
Rotating	Split without repl.	255.0	$.0003^{***}$	0.56	0.54	Medium
”	Temporal	331.0	$<.0001^{****}$	0.82	0.77	Large
”	Twisting	281.0	$<.0001^{****}$	0.76	0.68	Medium
Split without repl.	Underlining	−195.0	$.0434^{*}$	0.48	0.52	Small
Temporal	Underlining	−271.0	$<.0001^{****}$	0.72	0.81	Medium
Twisting	Underlining	−221.0	$.0056^{**}$	0.66	0.79	Medium

Table A.13: Dunn's pair-wise comparison test with multiplicity adjusted p-value: significant differences between ENGAGEMENT DTW distances of the various MENU-TYPE. Interpretation of Cohen's d [90]: $d \geq 0.8$=Large, $0.5 \leq d < 0.8$=Medium, $0.2 \leq d < 0.5$=Small. (Part 2/2)

Appendix A.3. ATTRACTION

Menu	Minimum	Maximum	Mean	Median	Std. dev.	Coeff. var.	Skewness	Kurtosis
Activation area	−38.81	46	−3.40	−3.46	20.14	591%	0.37	−0.062
Blinking	−48.82	61.34	−4.87	−8.77	24.69	506%	0.57	0.033
Bolding	−51.94	66.4	0.868	0.585	26.88	3095%	0.61	0.38
Boxing	−53.62	59.43	−5.85	−7.92	26.15	446%	0.50	0.024
Colored	−66.71	76.25	4.16	2.23	31.18	749%	0.012	−0.28
Ephemeral	−71.77	61.39	2.02	5.97	29.38	1451%	−0.61	0.32
Glowing	−37.16	58.86	1.50	−0.17	23.54	1564%	0.37	−0.68
Grayscaling	−46.08	65.12	1.59	2.68	26.86	1682%	0.22	−0.26
Highlighting	−40.77	56.15	3.67	0.29	27.13	737%	0.29	−0.63
Italicizing	−44.78	69.28	1.20	−1.80	23.19	1921%	0.62	1.13
Leaf	−41.83	68.94	3.75	−3.29	24.52	653%	0.73	0.44
Morphing	−55.97	56.22	8.88	5.11	26.49	298%	−0.0218	−0.058
Out-of-context	−59.15	73.32	−1.81	−2.66	28.17	1553%	0.31	0.24
Pulsing	−60.72	73.43	3.53	6.59	26.18	741%	−0.209	0.82
Rotating	−61.16	64.15	0.417	−2.68	25.86	6194%	0.399	1.05
Split without repl.	−52.8	84.95	−0.421	1.99	32.45	7703%	0.684	0.66
Static	−52.55	51.87	−4.58	−0.083	25.35	553%	−0.071	−0.46
Temporal	−55.73	59.67	4.38	6.97	27.05	617%	−0.061	−0.28
Twisting	−59.16	83.93	−0.109	−0.21	28.15	25614%	0.44	1.15
Underlining	−47.15	50.41	−2.71	−4.16	23.49	867%	0.27	−0.33

Table A.14: Descriptive analysis of the Attraction value for all Menu-Type.

Menu	Minimum	Maximum	Mean	Median	Std. dev.	Coeff. var.	Skewness	Kurtosis
Activation area	1224.18	3163.83	1727.61	1611.07	442.75	26%	1.8	3.7
Blinking	881.52	2415.34	1310.54	1274.83	275.80	21%	1.8	5.3
Bolding	1034.34	3342.43	1530.28	1357.01	484.41	32%	2.3	5.6
Boxing	938.50	2433.97	1324.78	1248.84	323.64	24%	1.8	4.0
Colored	946.54	2402.07	1414.33	1329.55	328.16	23%	1.2	1.4
Ephemeral	1098.95	2272.67	1420.92	1395.02	255.53	18%	1.6	3.3
Glowing	959.53	2406.51	1276.95	1190.73	322.20	25%	2.1	4.7
Grayscaling	1012.25	3408.58	1552.75	1450.14	443.01	29%	2.5	5.9
Highlighting	1138.34	2676.80	1498.42	1417.66	355.95	24%	2.0	4.4
Italicizing	1386.74	3773.39	1914.22	1794.06	505.89	26%	2.4	6.9
Leaf	1553.64	4651.44	2113.91	1945.14	557.03	26%	2.8	11.0
Morphing	1223.55	5310.85	1712.95	1574.82	652.45	38%	4.5	25.0
Out-of-context	976.55	2605.88	1457.71	1309.54	388.29	27%	1.5	2.1
Pulsing	1123.10	2520.73	1470.62	1398.27	304.74	25%	1.5	2.8
Rotating	1267.68	3965.37	1724.55	1591.24	484.06	28%	2.9	11.0
Split without repl.	1142.74	3318.66	1655.05	1497.84	459.64	28%	1.8	3.6
Static	1143.57	3046.27	1519.40	1424.20	434.48	29%	2.3	5.4
Temporal	991.84	3013.82	1354.62	1271.14	375.12	28%	3.0	11
Twisting	1051.52	3304.44	1498.46	1373.13	451.52	30%	2.5	7.3
Underlining	1173.02	4857.58	1657.14	1549.97	599.33	36%	4.2	21.0

Table A.15: Descriptive analysis of the Attraction DTW distance for all Menu-Type.

Appendix A.4. Memorization

Group 1	Group 2	Rank sum diff.	*p*-value	Cohen's *d*	Glass's Δ	Interp.
Activation area	Blinking	298.0	<.0001****	1.12	0.94	Large
”	Boxing	299.0	<.0001****	1.03	0.91	Large
”	Colored	245.0	.0007***	0.80	0.71	Large
”	Ephemeral	194.0	.046*	0.84	0.69	Large
”	Glowing	339.0	<.0001****	1.16	1.01	Large
”	Out-of-context	210.0	.013*	0.64	0.60	Medium
”	Temporal	291.0	<.0001****	0.90	0.84	Large
Blinking	Italicizing	-401.0	<.0001****	1.47	2.18	Large
”	Leaf	-472.0	<.0001****	1.82	2.90	Large
”	Morphing	-255.0	.0003***	0.80	1.45	Large
”	Rotating	-296.0	<.0001****	1.05	1.50	Large
”	Split without repl.	-242.0	.0009***	0.90	1.24	Large
”	Underlining	-241.0	.0010***	0.74	1.25	Medium
Bolding	Italicizing	-280.0	<.0001****	0.77	0.79	Medium
”	Leaf	-351.0	<.0001****	1.12	1.20	Large
Boxing	Italicizing	-402.0	<.0001****	1.38	1.81	Large
”	Leaf	-473.0	<.0001****	1.73	2.43	Large
”	Morphing	-256.0	.0002***	0.75	1.19	Medium
”	Rotating	-297.0	<.0001****	0.97	1.23	Large
”	Split without repl.	-243.0	.0008***	0.82	1.01	Large
”	Underlining	-242.0	.0009***	0.69	1.02	Medium
Colored	Italicizing	-348.0	<.0001****	1.17	1.52	Large
”	Leaf	-419.0	<.0001****	1.53	2.13	Large
”	Morphing	-202.0	.0256*	0.58	0.91	Medium
”	Rotating	-243.0	.0008***	0.75	0.94	Medium
Ephemeral	Italicizing	-297.0	<.0001****	1.22	1.93	Large
”	Leaf	-368.0	<.0001****	1.59	2.70	Large
Glowing	Grayscaling	-220.0	.0061**	0.71	0.85	Medium
”	Italicizing	-442.0	<.0001****	1.50	1.97	Large
”	Leaf	-513.0	<.0001****	1.83	2.59	Large
”	Morphing	-296.0	<.0001****	0.84	1.35	Large
”	Rotating	-337.0	<.0001****	1.08	1.39	Large
”	Split without repl.	-283.0	<.0001****	0.95	1.17	Large
”	Underlining	-282.0	<.0001****	0.79	1.18	Medium

Table A.16: Dunn's pair-wise comparison test with multiplicity adjusted *p*-value: significant differences between Attraction DTW distances of the various Menu-Type. Interpretation of Cohen's *d* [90]: $d \geq 0.8$=Large, $0.5 \leq d < 0.8$=Medium, $0.2 \leq d < 0.5$=Small, $d < 0.2$=Very small. (Part 1/2)

Appendix A.5. Completion-Time

Group 1	Group 2	Rank sum diff.	p-value	Cohen's d	Glass's Δ	Interp.
Grayscaling	Italicizing	-222.0	.0052**	0.75	0.81	Medium
”	Leaf	-293.0	<.0001****	1.11	1.26	Large
Highlighting	Italicizing	-251.0	.0004***	0.95	1.16	Large
”	Leaf	-322.0	<.0001****	1.31	1.73	Large
Italicizing	Out-of-context	313.0	<.0001****	1.01	0.90	Large
”	Pulsing	273.0	<.0001****	1.06	0.87	Large
”	Static	258.0	.0002***	0.83	0.78	Large
”	Temporal	394.0	<.0001****	1.25	1.10	Large
”	Twisting	277.0	<.0001****	0.86	0.82	Large
Leaf	Morphing	217.0	.0078**	0.66	0.71	Medium
”	Out-of-context	384.0	<.0001****	1.36	1.17	Large
”	Pulsing	344.0	<.0001****	1.43	1.15	Large
”	Split without repl.	230.0	.0026**	0.89	0.82	Large
”	Static	329.0	<.0001****	1.19	1.06	Large
”	Temporal	465.0	<.0001****	1.59	1.36	Large
”	Twisting	348.0	<.0001****	1.21	1.10	Large
”	Underlining	231.0	0.0024**	0.83	0.82	Large
Morphing	Temporal	248.0	.0005***	0.67	0.54	Medium
Out-of-context	Rotating	-208.0	.0161*	0.60	0.68	Large
Rotating	Temporal	289.0	<.0001****	0.85	0.76	Medium
Split without repl.	Temporal	235.0	.0017**	0.71	0.65	Medium
Temporal	Underlining	-234.0	.0019**	0.65	0.80	Medium

Table A.17: Dunn's pair-wise comparison test with multiplicity adjusted p-value: significant differences between ATTRACTION DTW distances of the various MENU-TYPE. Interpretation of Cohen's d [90]: $d \geq 0.8$=Large, $0.5 \leq d < 0.8$=Medium, $0.2 \leq d < 0.5$=Small, $d < 0.2$=Very small.(Part 2/2)

Menu	Minimum	Maximum	Mean	Median	Std. dev.	Coeff. var.	Skewness	Kurtosis
Activation area	2.3	44	24	24	11	46%	0.09	−0.82
Blinking	3.8	66	22	22	12	55%	1.3	3.6
Bolding	2.1	47	24	26	9.6	39%	−0.37	0.23
Boxing	6.1	53	23	18	13	58%	0.84	−0.37
Colored	2.9	83	23	20	16	68%	1.6	4.4
Ephemeral	2.7	62	21	22	11	52%	1	3.1
Glowing	4	66	26	23	14	54%	0.92	0.68
Grayscaling	4.9	61	26	25	11	43%	0.79	1.4
Highlighting	3.2	49	24	22	10	43%	0.26	−0.059
Italicizing	11	54	24	23	9.1	38%	1.1	1.8
Leaf	11	64	30	28	13	44%	0.84	0.016
Morphing	3	62	23	20	13	57%	0.78	0.85
Out-of-context	4.2	77	27	23	14	53%	1.1	2.3
Pulsing	5	56	24	24	11	44%	0.66	0.94
Rotating	5.4	51	23	22	11	46%	0.66	−0.32
Split without repl.	1.3	59	23	22	12	54%	1	2
Static	3.6	45	20	20	9.8	49%	0.55	0.15
Temporal	3.8	53	25	23	13	50%	0.48	−0.63
Twisting	5.7	49	24	24	10	44%	0.51	0.17
Underlining	4.5	41	24	25	9.4	40%	−0.02	−0.78

Table A.18: Descriptive analysis of the MEMORIZATION value for all MENU-TYPE.

Menu	Minimum	Maximum	Mean	Median	Std. dev.	Coeff. var.	Skewness	Kurtosis
Activation area	561.48	1680.57	755.66	650.63	248.80	33%	2.4	6.3
Blinking	411.89	1167.71	619.97	548.95	170.67	28%	1.7	2.7
Bolding	497.37	1479.91	737.56	676.74	205.82	28%	1.9	4.0
Boxing	478.88	1316.57	654.96	610.48	162.29	25%	2.2	6.7
Colored	471.96	1379.12	665.48	594.24	200.27	30%	1.8	3.6
Ephemeral	434.37	1130.56	622.91	574.15	166.89	25%	1.5	2.0
Glowing	479.78	1098.80	658.17	617.63	159.72	24%	1.1	0.78
Grayscaling	539.90	1573.49	809.51	769.74	192.08	24%	2.0	5.8
Highlighting	483.57	1798.96	681.69	598.19	243.23	36%	2.9	11
Italicizing	605.27	1778.64	854.97	796.69	251.11	29%	2.6	7.8
Leaf	785.55	1623.86	1023.88	981.05	194.14	19%	1.5	2.4
Morphing	558.82	2208.74	791.34	706.72	302.04	38%	3.1	12
Out-of-context	485.26	1288.93	685.73	651.35	162.26	24%	2.3	6.5
Pulsing	471.11	1326.10	660.26	561.24	197.21	30%	1.9	3.2
Rotating	575.40	2529.84	824.35	709.98	340.89	41%	3.6	16
Split without repl.	481.58	1471.51	713.84	657.90	232.18	33%	1.8	3.0
Static	461.36	1377.32	626.55	570.26	193.49	31%	2.6	7.4
Temporal	521.89	1194.05	691.84	663.07	137.41	20%	1.5	3.4
Twisting	464.09	1952.79	646.06	598.44	245.85	38%	4.3	21
Underlining	542.29	2314.29	777.22	724.28	292.09	38%	4.0	20

Table A.19: Descriptive analysis of the MEMORIZATION DTW distance for all MENU-TYPE.

Group 1	Group 2	Rank sum diff.	p-value	Cohen's d	Glass's Δ	Interp.
Leaf	Static	220.0	.0061**	0.14	0.76	Very small

Table A.20: Dunn's pair-wise comparison test with multiplicity adjusted p-value: significant differences between MEMORIZATION values of the various MENU-TYPE. Interpretation of Cohen's d [90]: d<0.2=Very small.

Group 1	Group 2	Rank sum diff.	p-value	Cohen's d	Glass's Δ	Interp.
Activation area	Blinking	234.0	.0019**	0.63	0.54	Medium
”	Ephemeral	216.0	.0085**	0.62	0.53	Medium
”	Leaf	-254.0	.0003***	1.20	1.07	Large
”	Static	225.0	.0040**	0.57	0.51	Medium
Blinking	Bolding	-214.0	.0100**	0.62	0.69	Medium
”	Grayscaling	-321.0	<.0001****	1.04	1.11	Medium
”	Italicizing	-359.0	<.0001****	1.09	1.37	Large
”	Leaf	-488.0	<.0001****	2.20	2.36	Large
”	Morphing	-241.0	.0010***	0.69	1.00	Medium
”	Rotating	-297.0	<.0001****	0.75	1.19	Medium
”	Underlining	-259.0	.0002***	0.65	0.91	Medium
Bolding	Ephemeral	196.0	.0403*	0.61	0.55	Medium
”	Leaf	-274.0	<.0001****	1.42	1.38	Large
”	Static	205.0	.0203*	0.55	0.53	Medium
Boxing	Grayscaling	-218.0	.0072**	0.87	0.95	Large
”	Italicizing	-256.0	.0002***	0.94	1.23	Large
”	Leaf	-385.0	<.0001****	2.06	2.27	Large
”	Rotating	-194.0	.0468*	0.63	1.04	Medium
Colored	Grayscaling	-238.0	.0013**	0.73	0.72	Medium
”	Italicizing	-276.0	<.0001****	0.83	0.95	Large
”	Leaf	-405.0	<.0001****	1.82	1.79	Large
”	Rotating	-214.0	.0100**	0.56	0.79	Medium
Ephemeral	Grayscaling	-303.0	<.0001****	1.03	1.11	Large
”	Italicizing	-341.0	<.0001****	0.93	1.23	Large
”	Leaf	-470.0	<.0001****	2.05	2.28	Large
”	Morphing	-223.0	.0048**	0.55	0.83	Medium
”	Rotating	-279.0	<.0001****	0.62	1.03	Medium
”	Underlining	-241.0	.0010***	0.50	0.74	Medium
Glowing	Grayscaling	-230.0	.0026**	0.86	0.95	Large
”	Italicizing	-268.0	<.0001****	0.93	1.23	Large
”	Leaf	-397.0	<.0001****	2.05	2.28	Large
”	Morphing	-296.0	<.0001****	0.55	0.83	Large
”	Rotating	-206.0	.0188*	0.62	1.03	Medium

Table A.21: Dunn's pair-wise comparison test with multiplicity adjusted p-value: significant differences between Memorization DTW distances of the various Menu-Type. Interpretation of Cohen's d [90]: $d \geq 0.8$=Large, $0.5 \leq d < 0.8$=Medium, $0.2 \leq d < 0.5$=Small, $d < 0.2$=Very small. (Part 1/2)

Group 1	Group 2	Rank sum diff.	p-value	Cohen's d	Glass's Δ	Interp.
Grayscaling	Highlighting	214.0	.0100**	0.58	0.66	Medium
”	Pulsing	239.0	.0012**	0.77	0.78	Medium
”	Static	312.0	<.0001****	0.95	0.95	Large
”	Twisting	275.0	<.0001****	0.74	0.85	Medium
Highlighting	Italicizing	-252.0	.0004***	0.70	0.71	Medium
”	Leaf	-381.0	<.0001****	1.55	1.40	Large
Italicizing	Out-of-context	210.0	.0137*	0.80	0.67	Large
”	Pulsing	277.0	<.0001****	0.86	0.77	Large
”	Split without repl.	200.0	.0298*	0.58	0.56	Medium
”	Static	350.0	<.0001****	1.01	0.90	Large
”	Temporal	199.0	.0322*	0.80	0.64	Large
”	Twisting	313.0	<.0001****	0.84	0.83	Large
Leaf	Morphing	247.0	.0006***	0.91	1.20	Large
”	Out-of-context	339.0	<.0001****	1.89	1.74	Large
”	Pulsing	406.0	<.0001****	1.86	1.87	Large
”	Split without repl.	329.0	<.0001****	1.44	1.59	Large
”	Static	479.0	<.0001****	2.05	2.04	Large
”	Temporal	328.0	<.0001****	1.97	1.71	Large
”	Twisting	442.0	<.0001****	1.70	1.94	Large
”	Underlining	231.0	0.0029**	0.99	1.27	Large
Morphing	Static	232.0	.0022**	0.64	0.54	Medium
”	Twisting	195.0	.0434*	0.52	0.48	Medium
Pulsing	Rotating	-215.0	.0092**	0.58	0.83	Medium
Rotating	Static	289.0	<.0001****	0.71	0.57	Medium
”	Twisting	251.0	.0004***	0.59	0.52	Medium
Static	Underlining	-250.0	.0004***	0.60	0.77	Medium
Temporal	Underlining	-213.0	.0108*	0.37	0.62	Small

Table A.22: Dunn's pair-wise comparison test with multiplicity adjusted p-value: significant differences between Memorization DTW distances of the various Menu-Type. Interpretation of Cohen's d [90]: d≥0.8=Large, 0.5≤d<0.8=Medium, 0.2≤d<0.5=Small, d<0.2=Very small. (Part 2/2)

Menu	Minimum	Maximum	Mean	Median	Std. dev.	Coeff. var.	Skewness	Kurtosis
Activation area	1926	24 094	6626	4189	5330	80%	1.9	3.3
Blinking	1340	9183	3432	2954	1709	50%	2.6	1.5
Bolding	1336	26 813	5211	2738	6168	118%	2.5	5.5
Boxing	1614	9816	3697	2834	1919	52%	1.5	2.0
Colored	1149	11 196	3115	2443	2041	66%	2.4	6.5
Ephemeral	1254	11 934	3336	2498	2087	63%	2.2	6.3
Glowing	1402	10 023	3114	2484	2015	65%	2.2	4.5
Grayscaling	1512	29 484	4883	3093	5323	109%	3.4	13
Highlighting	1754	29 695	4736	3533	4578	97%	4.5	23
italicizing	1699	34 496	8219	6531	7393	90%	2.3	5.8
Leaf	2199	31 786	10 387	8195	6447	62%	1.3	1.9
Morphing	1254	50 765	5612	3570	8151	145%	4.7	25
Out-of-context	1696	18 473	3791	2914	3046	80%	3.6	15
Pulsing	1656	13 484	4648	3500	3222	69%	1.8	2.3
Rotating	1941	44 281	6519	4088	7658	117%	3.7	16
Split without repl.	1063	22 789	4334	2484	4603	106%	2.6	7.0
Static	1430	24 606	4807	3125	5090	106%	3.1	10
Temporal	1426	25 312	3895	2873	3907	100%	4.6	24
Twisting	2273	27 289	4534	3476	3977	88%	5.1	29
Underlining	1730	45 539	5948	3802	7267	122%	4.4	23

Table A.23: Descriptive analysis of the Completion-Time for all Menu-Type.

Group 1	Group 2	Rank sum diff.	p-value	Cohen's d	Glass's Δ	Interp.
Activation area	Blinking	246.0	.0007***	0.80	0.59	Large
”	Bolding	244.0	.0008***	0.24	0.26	Small
”	Boxing	198.0	.0360*	0.73	0.54	Medium
”	Colored	338.0	<.0001****	0.86	0.65	Large
”	Ephemeral	263.0	.0001***	0.81	0.61	Large
”	Glowing	329.0	.0003***	0.87	0.65	Large
”	Out-of-context	211.0	.0132*	0.65	0.53	Medium
”	Split without repl.	264.0	.0001***	0.46	0.43	Small
”	Static	203.0	.0237***	0.34	0.34	Small
”	Temporal	228.0	.0033**	0.58	0.51	Medium
Blinking	Italicizing	-291.0	<.0001****	0.89	2.80	Large
”	Leaf	-386.0	<.0001****	1.47	4.06	Large
”	Rotating	-205.0	.0203*	0.55	1.80	Medium
Bolding	Italicizing	-289.0	<.0001****	0.44	0.48	Small
”	Leaf	-384.0	<.0001****	0.82	0.83	Large
”	Rotating	-203.0	.0237*	0.18	0.21	Very small
Boxing	Italicizing	-243.0	.0008***	0.83	2.35	Large
”	Leaf	-338.0	<.0001****	1.40	3.48	Large
Colored	Highlighting	-218.0	.0072**	0.45	0.79	Small
”	Italicizing	-383.0	<.0001****	0.94	2.50	Large
”	Leaf	-478.0	<.0001****	1.52	3.56	Large
”	Rotating	-297.0	<.0001****	0.60	1.66	Medium
”	Twisting	-258.0	.0002***	0.44	0.69	Small
”	Underlining	-257.0	.0002***	0.53	1.38	Medium
Ephemeral	Italicizing	-308.0	<.0001****	0.89	2.33	Large
”	Leaf	-403.0	<.0001****	1.47	3.37	Large
”	Rotating	-222.0	.0052**	0.56	1.52	Medium
Glowing	Highlighting	-210.0	.0143*	0.45	0.80	Small
”	Italicizing	-375.0	<.0001****	0.94	2.53	Large
”	Leaf	-470.0	<.0001****	1.52	3.60	Large
”	Rotating	-289.0	<.0001****	0.60	1.68	Large
”	Twisting	-250.0	.0005***	0.45	0.70	Small
”	Underlining	-248.0	.0005***	0.53	1.40	Medium
Grayscaling	Italicizing	-220.0	.0061**	0.51	0.62	Medium
”	Leaf	-315.0	<.0001****	0.93	1.03	Large
Highlighting	Leaf	-260.0	.0002***	1.01	1.23	Large
Italicizing	Morphing	217.0	.0078**	0.33	0.35	Small
”	Out-of-context	256.0	.0002***	0.78	0.59	Medium
”	Pulsing	195.0	.0434*	0.62	0.48	Medium
”	Split without repl.	309.0	<.0001****	0.63	0.52	Medium
”	Static	249.0	.0005***	0.53	0.46	Medium
”	Temporal	273.0	<.0001****	0.73	0.58	Medium
Leaf	Morphing	312.0	<.0001****	0.64	0.74	Medium
”	Out-of-context	351.0	<.0001****	1.30	1.02	Large
”	Pulsing	290.0	<.0001****	1.12	0.89	Large
”	Split without repl.	404.0	<.0001****	1.08	0.93	Large
”	Static	344.0	<.0001****	0.96	0.86	Large
”	Temporal	368.0	<.0001****	1.21	1.00	Large
”	Twisting	220.0	.0061**	1.09	0.90	Large
”	Underlining	222.0	.0054**	0.64	0.68	Medium
Rotating	Split without repl.	223.0	.0048**	0.34	0.28	Medium

Table A.24: Dunn's pair-wise comparison test with multiplicity adjusted p-value: significant differences between Completion-Time of the various Menu-Type. Interpretation of Cohen's d [90]: $d \geq 0.8$=Large, $0.5 \leq d < 0.8$=Medium, $0.2 \leq d < 0.5$=Small, $d < 0.2$=Very small.