Practical Neurophysiological Analysis of Readability as a Usability dimension

Inês I. Oliveira[1], Nuno M. Guimarães[2]

[1]SITILabs, Lusófona University, Portugal
ines.oliveira@ulusofona.pt
[2]LaSIGE/ISCTE-IUL, University Institute of Lisbon, Portugal
nmcfg@iscte.pt

Abstract. This paper discusses opportunities and feasibility of integrating neurophysiologic analysis methods, based on electroencephalography (EEG), in the current landscape of usability evaluation methods. The rapid evolution and growing availability of low-cost, easier to use devices and the accumulated knowledge in feature extraction and processing algorithms allow us to foresee the practicality of this integration.
The work presented in this paper is focused on reading and readability, identified as a key element of usability heuristics, and observable in the neurophysiologic signals' space. The experiments are primarily designed to address the discrimination of the reading activity (silent, attentive and continuous) and the verification of decreasing readability, associated with the user's mental workload analysis. The results obtained in the series of experiments demonstrate the validity of the approach for each individual user, and raise the problem of inter-subject variability and the need for designing appropriate calibration procedures for different users.

Keywords. Usability analysis, neurophysiologic signals, EEG

1 Introduction

The current usability evaluation methods range from interpretative to predictive, the former based on the observation and study of the actual use of an artifact during its development cycle, and the latter exploring external analyses performed by experts equipped with standards, heuristics and modeling techniques [1]. The methods capture either the behavior and perceptions of users or the interactive attributes of the artifact, and converge to evaluation conclusions based on the qualitative or quantitative analysis of empirically collected data.

Alternative, or complementary, methods based in the measurement of physical and physiological signals of the human user (e.g. eye-movements, heart rate (HR) and heart rate variation (HRV), skin conductance (SC), or electroencephalography (EEG)) have been used more frequently in contexts with critical requirements for human performance [2], and tested in dedicated labs, as opposed to the use of the former empiri-

cal methods, typically adopted by usability labs participating in the design and development of interactive artifacts for the general user or consumer.

The evolution of the technological landscape leads us reassess the opportunities for expanding and improving the set of tools for usability evaluation. First, capture devices are cheaper, more reliable, less intrusive, and usable with increasing autonomy. Examples of this evolution are HR or SC measurement devices, which have become portable and wireless and even EEG systems have left the controlled conditions of clinical settings (see for example, the design of dry electrode devices [3]. Second, the capacity to process these signals evolved dramatically, both in computing power and in the understanding of the algorithms that extract meaning from the human data.

As a result of this evolution, we can envision a feasible integration of human physical and physiological information in common interactive artifacts, as a new modality. On the other hand, in the scope of usability evaluation, we can aim at incorporating those human signals in the analysis setup, namely in the common usability lab environment, thus reinforcing the mature and widely adopted empirical methods with easier to use physical measurement methods.

In this paper we assess the effectiveness of a brain signal analysis in usability evaluation, with a focus on readability - the ability of a user being able to read a text. This is not a sufficient condition for system usability, but it is, in many interactive artifacts, a necessary one. Readability is affected by interface design decisions, with a particular relevance on presentation choices. Readability has an impact in usability heuristics, discussed below.

Our focus is therefore the detection and analysis of the cognitive activity of reading. Again, readability is not a sufficient condition for a user to read but it is a necessary one. The detection of reading should be a good indicator of readability, provided that variables that create obstacles to reading such as high text complexity, foreign language, distraction, fatigue, or even cultural bias, are eliminated. In this work, we considered that a strong correlation between reading and readability should be sought in continuous attentive reading activities. This avoids the need of discrimination between recognizing word and text as grammatically appropriate character sequences, and recognizing isolated words as individual learned symbols.

The next section reviews a number of works that have used physiological signal processing in the scope of the analysis of the usability of some system or tool. In the following two sections, we scope this work along the baseline aspects: (i) reading as an activity correlated with readability, which is in turn directly affected by user interface design decisions and has an impact on heuristically defined usability criteria, and (ii) the set of computational techniques used in the processing of brain signals. The concrete experimental framework is then presented, followed by the results of a set of experiments. The implications for future work towards an effective integration of brain signal processing in usability evaluation are elaborated in the conclusion.

2 Related Work

Beer et al. refer that "the usability lab of the future" must integrate analysis tools based in physiological measures, including the EEG [4]. These signals are potentially valuable for measuring users' emotional valence and vigilance during the interaction [5]. As the data generated with these methods comes directly from the users' physical processes, without intermediation of an observer or expert, it can reveal, for example, social masks, when users avoid giving negative answers in interviews. Still, this integration is preliminary. The studies quoted below compare physiological based analysis with classical usability methods, e.g. questionnaires.

Reference	Main Goals	Analysis Methods	Main Conclusions
#1 [6] 5 participants	Distinguish emotional states using 3 distinct menus: - regular &familiar - illegible - error	**Physiological**: EEG Theta, alpha and beta rhythms of the best 2 out of 10 electrodes **Classical**: - Questionnaire - Task Difficulty classification - A/V and eye tracking recordings	- Attested the correlation between EEG and inquiry data (1 user); - Attest the correlation between EEG and difficulty classification (5 users)
#2 [7] 43 participants	Study the emotional response to 2 alternative prototypes for an e-government site, (with/without anthropomorphic Web Assistant)	**Physiological**: ESR[1] differential analysis with (max-min)/min, max and 1st peak value **Classical**: Inquiries, SMEQ[2] Scale Inquiry	- Attest the correlation between ESR and classical methods - Users preferred Web Assistant version - Differences between both sites were statistically relevant with both measures
#3 [8,9] 20 participants	- Performance test to evaluate an university learning system - Non moderated time limited tasks Different system experienced users.	**Physiological**: ESR and HRV[3] **Classical**: - Extended NPL[4] inquiry - TAP[5] - A/V recordings - SUS[6] questionnaire	- Attest the correlation between users performance and emotional state; - Both groups reveled emotional differences
#4 [10,11] 10 participants	Study mental workload differences in a set of equivalent tasks in MS Excel 2003 and 2007	**Physiological**: EEG Alpha and beta rhythms' average Normalized PSD. **Classical**: SUS questionnaire	- Attest the correlation between both types of measures software experience and alpha and beta ratio variation - Users preferred (and showed less mental workload) Excel 2007

[1] Electrodermal Skin Response
[2] Subjective Mental Effort Questionnaire
[3] Heart rate variation
[4] National Physics Laboratory
[5] Think Aloud Protocol
[6] System Usability Scale

Reference	Main Goals	Analysis Methods	Main Conclusions
#5 [12] 4 participants	Performance versus workload test in a simple game with 3 distinct difficult levels	**Physiological**: - EEG: Avg. PSD[7] (in alpha, beta, theta, delta and gamma), cross spectrum and coherence - fNIR[8]: Normalized Oxygenation Variation **Classical**: User performance (game score)	- Performance is proportional with level's difficulty - Accuracy classification depends on the used measure (better with fNIR) - fNIR may interfere with EEG sensors
#6 [13] 10 participants	Evaluate user preferences and emotional states regarding 4 car company web-sites	**Physiological**: - EEG (ERPs, PSD in beta and theta) - HR (Std, Greater and minor freq. ratio) **Classical**: - Preference questionnaire - Error and Task completion rates - Task Execution time - A/V and screen recordings	- 100% of correlation between classical versus EEG results - 60% of correlation between classical versus ECG results
#7 [14] 36 participants	Game UX evaluation of an immersive game using 2 input devices and consoles: - Standard gamepad in PSP2 - Wii Remote in Nintendo	**Physiological:** EEG - Normalized PSD in alpha, beta, delta, theta and gamma **Classical:** Questionnaires - GEQ (Game Experience Questionnaire) - Auto-localization, to evaluate perception of physical location and action options in VR environment	- Attest the correlation between both types of measures; - WII Remote scores better in questionnaires and also causes a greater mental activity

Table 1. Studies comparing physiological-based methods with classical empirical methods for analysis of specific usability dimensions.

These studies show the potential of physiological measures in usability and user experience evaluation and demonstrate the correlation between physiological and traditional methods. The test cases are however very constrained situations, and several open issues are identified. First, capture devices are expensive, intrusive and complex to handle, making it possible to generate emotions that are not directly related with the interaction [5]. Secondly, it is difficult to generalize the results, because of the various degrees of variation, such as gender, age and culture. Finally, the interpretation of the measures is complex, even when the cause and effect are known, because it strongly depends on the social and interpersonal context [15].

[7] Power Spectrum Density
[8] Functional Near Infrared Spectroscopy

3 Reading in Usability

Usability heuristics have become generalized tools to evaluate usability of interactive products, systems or services. These heuristics are empirically consolidated reflections of the structural coupling [16] requirements that a usable user interface implicitly meets. This coupling between a user and an artefact is maintained as long as the properties of the artifact are compatible with the user's cognitive or physical abilities. On one side we have characteristics of the artifact like legibility, language, visibility or aesthetics, and on the other we have human cognitive processes like reading, understanding, memory or different emotional reactions.

As an example, let us consider the universally acknowledged Nielsen heuristics [17] listed in **Table 2**. The table relates the heuristics with cognitive processes through a number of (not exhaustive and non orthogonal) determining user interface characteristics or interaction mechanisms. This tentative mapping is partially justified by the experimentations reported in the previous section (quoted in column 4 of **Table 2**).

Heuristic (1)	Determining UI Characteristics (2)	Cognitive processes (3)	Related work(4)
1. Visibility of system status	Legibility, visual[9] expression, feedback (e.g. icons)	Perception incl. Reading	#2
2. Match system and real world	Semiotic design (e.g. metaphors)	Understanding, Memory, **Emotion**	#2
3. User control and freedom	Multitasking, escapability, recovery (e.g. undo/redo)	Memory, Workload, Learning, **Emotion**	#3, #7
4. Consistency, standards	Visual design, Homogeneity (e.g. layout, menus)	Perception, Memory	#1
5. Error prevention	Legibility, visual feedback (e.g. data formats)	Perception incl. Reading, Memory, Workload	#1
6. Recognition rather than recall	Semiotic design, legibility (e.g. menus)	Perception incl. Reading**,** Memory	#1
7. Flexibility, efficiency of use	Attention requirements, Adaptability (e.g. shortcuts)	**Workload**, Memory, Learning	#3, #4, #5, #7
8. Aesthetic, minimalist design	Legibility, visual design (e.g. look)	Perception, Reading, **Emotion**	#6
9. Help users recognize, diagnose, and recover from errors	Legibility, visual feedback (e.g. error messages)	Perception, Reading, Problem solving	

Table 2. Relations between (de facto standard) usability heuristics and cognitive processes.

The above associations express a path between usability heuristics and cognitive processes that can be observed through their external manifestations in physical and physiological signals. This path is conceptually important for an integrated perspective of the analysis and evaluation methods.

[9] "Visual" is mentioned here in the broad sense of an external, perceptually intelligible representation, but it can be based in any other modality like audio or haptic forms.

Reading, especially continuous, attentive, and silent reading, as not been analyzed in this context. In fact, reading in the common user interface is generally associated with word recognition for most of the textual elements of the user interface (e.g. labels, menus, icons or forms). Its performance can be severely affected by several design factors including: typeface and text features problems (e.g. inappropriate color and text or line spacing), poor contrast between background and text, uncomfortable screen distance, design and formatting problems (e.g. too wide or too narrow text and center or right justification) [18]. Readability can therefore be considered as a usability guideline among the heuristics referred above [19].

4 Brain Signals

In this section, we briefly describe the hardware and software platform used to capture and process the brain signals. The selected and used computational techniques are also briefly enumerated, as a more detailed explanation and discussion of the processing techniques can be found elsewhere [20, 21].

The signal acquisition and selection was based in two fundamental requirements. First, the acquisition devices should be low-end devices, usable in the environment of a usability lab, as opposed to a clinical environment. The low-end devices, with a limited number of channels, do still require some level of expertise to set up an acquisition session (electrode placement and impedance adjustment) but are manageable by an experienced technician or researcher, and are in line with the expected evolution of more usable and portable acquisition devices.

Second, the signals to observe should avoid highly demanding synchronization requirements. Since the goal is to detect and discriminate reading activity, the focus of the analysis was the variation of brain waves instead of the brain responses to discrete stimuli. This focus leads to the choice of the analysis of brain rhythms (alpha (α), beta (β), delta (δ), theta (θ), and gamma (γ)), instead of ERP (Event Related Potentials), which require controlled synchronization conditions between stimuli and acquired signal (a few milliseconds). While the later signals provide the proper information for studying brain responses and are actually the main source of "input" in BCI (Brain Computer Interfaces), see [22-24], the former analysis is appropriate for cognitive process detection and better suited for future lightweight devices.

4.1 Signal Aquisition

The signal acquisition was made in an open space human-computer interaction lab with MindSet MS-1000, a sixteen (16) channel digital capture device, and its proprietary software acquisition tool named MindMeld. All sensors were attached to an ElectroCap cap and positioned accordingly with the 10-20 International System: six (6) in the frontal area, four (4) in the temporal, and the remaining six (6) in the parietal, central and occipital (two in each area).

The capture was performed at 256Hz using referential electrodes placed in ear lobes. All requirements defined by EEG capture experts and the devices fabricants

were met: all users were connected to a ground wire to reduce the electrical noise peak; hair was brushed with a wooden brush to reduce electrostatic; scalp sensors place was previously cleaned with alcohol; conductive gel was applied in all electrodes and impedance was maintained below 5000kΩ in all electrodes.

4.2 Signal Processing Chain

The components of the signal processing chain we applied to the brain signals are shown in **Fig. 1**. In the figure, the thicker connectors mean "mandatory path"; traced connectors mean "alternative or optional", and the thinner connection shown the contribution of feature selection to the indicated steps in the chain.

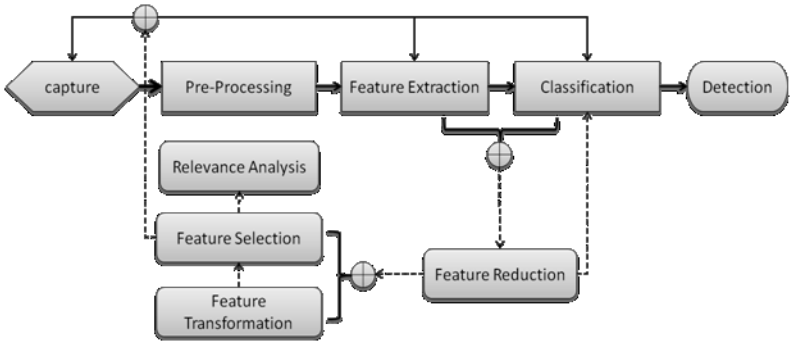


Fig. 1. The Signal Processing Chain.

The sixteen (16) signals, one per channel, are pre-processed individually to reduce non EEG artifacts (e.g. electrical noise), and transformed into feature vectors, where each vector is composed by 16x5(80) real values – the estimated average power spectrum in delta, theta, alpha, beta and gamma rhythms, determined in one second periods (also called *windows*), overlapped in 0.5 seconds. A feature vector can then be classified in certain classes related to user cognitive processes, such reading and non-reading. In the process, feature vector dimension can previously be reduced, to increase the chain and classification overall performance, through transformation and/or selection methods [25]. The feature selection can also be based on relevance analysis. This analysis ranks features according to their contribution to the discrimination of the classes under observation [21].

All the results presented and discussed in this paper were obtained using PSD (Power Spectrum Density) feature extraction and an SVM (Support Vector Machines) classifier. PSD measures the energy of the signal in a certain frequency [21]. It is a frequency feature, so to get its variation over time, the PSD is determined in one second length rectangular windows, overlapped in 0.5 seconds. SVM are supervised classifiers proven to be successful in EEG analysis [26]. SVMs divide the solution space in hyper-planes through discriminating functions. In our case we use the kernel trick that makes SVMs, originally linear, nonlinear classifiers. This kernel uses a Gaussian radial basis function (RBF), generally used in BCI research.

Preprocessing methods that were used in the results presented in this paper only include a Notch (narrow band) filter, which allows reducing electrical noise peek at 50Hz, still present after grounding users. No method has been used to reduce eye and movement artifacts. Our goal is to build a method robust enough to handle these interferences that naturally result from user interaction.

5 Experimental Procedures

An experimental session (or trial) is composed by a sequence of several distinct experiments, all related with silent reading in a screen. A session takes approximately twenty-five (25) minutes. All experiments were separated by thirty (30) second resting periods, during which users were asked about the text topic (they just read), when they stopped reading, and their overall mental and emotional state.

All the results presented in this paper were performed with six (6) users, three (3) women and three (3) man, ages between 20 and 45, without relevant neurological or sight or visual known conditions, three (3) using glasses, and one (1) left-handed. All users concluded successfully a higher educational degree and frequently read literary, technical and/or news texts in both paper and digital format. There was no previous training, but all users repeated the experiences in different days, with distinct texts and images.

All texts were written in the native language and never repeated for the same user. An event generator application was used to build and display an event script such as a slide show [20]. We built twenty-three (23) different scripts for these experiments. All content was displayed in a 15.4'' laptop LCD colored screen, with a 1280x800 resolution and 4,295E+9 colors. The laptop was set in a regular desk, with about 70-80 centimeters height. Users sat in front of the screen, at a distance between 50-60 centimeters. In general, all texts were displayed with Arial 21px font in a black foreground over a white background, unless experiments thus required. The experiments whose results are described in this paper are the following:

1. **Text versus Blank Screen:** Users read two texts (with news) for 30 seconds each, interleaved with a blank screen for 20 seconds. This was a preliminary experiment, to evaluate whether it was possible distinguishing silent reading cognitive state from another basic visual state, such as looking at a blank screen, to tune algorithms and also to study the variation between sessions and users.
2. **Text versus Drawings:** this is a similar experience, but, instead of a blank screen, it interleaves text with black and white unfilled drawings for 30 seconds. It allows assessing the possibility of distinguishing silent reading cognitive state regarding a simple visual, non verbal, stimulus.

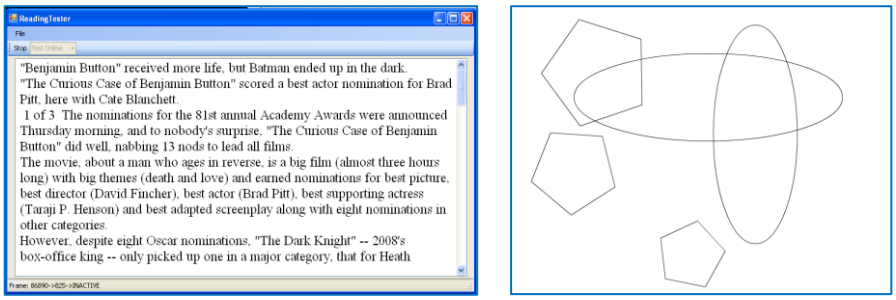


Fig. 2. Example of news and drawing display used in the first two experiments.

3. **Text Size Decrease.** Users read a 70 words news text, one word at a time, each lasting one second. Every ten seconds the size of text decreases by 3px, varying between 21px and 3px.

Fig. 3. Text size variation simulation used in Text Size Decrease experiment.

4. **Background/Text Contrast Decrease (by varying background).** Users read news text during seventy (70) seconds while every ten (10) seconds the background is darkened about 16% in relation to white. This was preceded by the same experience without text to serve as control.
5. **Background/Text Contrast Decrease (by varying text).** Users read news text during seventy (70) seconds while every ten (10) seconds the text is lighten about 16% in relation to black.
6. **RGB Difference Background/Text Decrease.** Users read news text during seventy (70) seconds while every ten (10) seconds the RGB difference between background and text is reduced between 5 to 15%.

All these experiments concern directly and indirectly the usability heuristics (1) visibility, (4) standards, or (8) aesthetic and minimalist design, mentioned in **Table 2**. As previously mentioned, text visual characteristics influence readability, and consequently reading performance. This indirectly influences all other heuristics where the reading cognitive process plays a role, such as Recognition and Recall.

5.1 W3C (World Wide Web Consortium) Thresholds

To guarantee that a text can be read in a screen by users with color deficits , the W3C proposes a minimum font size for screens of 9px [28] and a contrast and RGB difference thresholds of 125 and 500 respectively [27]. The last two sizes (6px and 3px) in experiment 3 violate these W3C recommendations, and the intensity and RGB values that were used in experiments 4-6 are shown in the next tables.

Experimental Step	Back. Intensity (B)	Text Intensity (T)	(B)-(T)	Compliance with W3C threshold
1	255	0	255	**Yes**
2	214	0	214	**Yes**
3	173	0	173	**Yes**
4	132	0	132	**Yes**
5	91	0	91	**No**
6	51	0	51	**No**

Table 3. Experiment 4. Intensity values and compliance with W3C thresholds.

Experimental Step	Back. Intensity(B)	Text Intensity(T)	(B) –(T)	Compliance with W3C threshold
1	255	0	255	**Yes**
2	255	40	215	**Yes**
3	255	81	174	**Yes**
4	255	122	133	**Yes**
5	255	163	92	**No**
6	255	204	51	**No**

Table 4. Experiment 5. Intensity values and compliance with W3C thresholds.

Experi-mental Step	Background			Text			RGB Dif-ference	Compli-ance w/ W3C threshold
	Red	Green	Blue	Red	Green	Blue		
1	255	255	255	0	0	0	765	**Yes**
2	255	247	215	20	0	40	657	**Yes**
3	235	239	215	20	0	120	549	**Yes**
4	235	223	215	20	0	200	453	**No**
5	195	223	215	60	0	200	373	**No**
6	175	223	175	80	8	200	335	**No**

Table 5. Experiment 6. RGB values and compliance with W3C thresholds.

6 Results and Discussion

The organization of this section is the following: In **section 6.1.** are presented the results regarding experiments 1-2, where is analysed silent continuous reading distinction towards other simple visual states. Here are discussed topics such as user and session generalization.

The results of the remaining experiments, which vary text or background aspects that affect readability, are discussed in **section 6.2**. Based on the work of Kimura and Masaki [10,11], we analyze the correlation between classical evaluation based measures, such as inquiries, and an EEG based measure. This measure was the beta/alpha PSD ratio, considered to be indicative of *mental workload* [10,11].

The presented results consider the following data corpus subsets:

- **Intra-user data set (A).** Includes 13 sessions of a single unique user.
- **Inter-user data set (B).** Includes 12 sessions of six users, two sessions each.

All the procedures and metrics were initially tuned using data set A. The following section discusses the results of processing these experiments using both data sets.

6.1 Silent Reading Distinction

The results of silent reading distinction (with both data set A and B) are shown in **Fig. 4**. F-Measure is a classification performance measure that averages precision and recall geometrically in a single value. It varies between 0%, the worst possible result and, 100%, the best, where all mental states were 100% correctly classified.

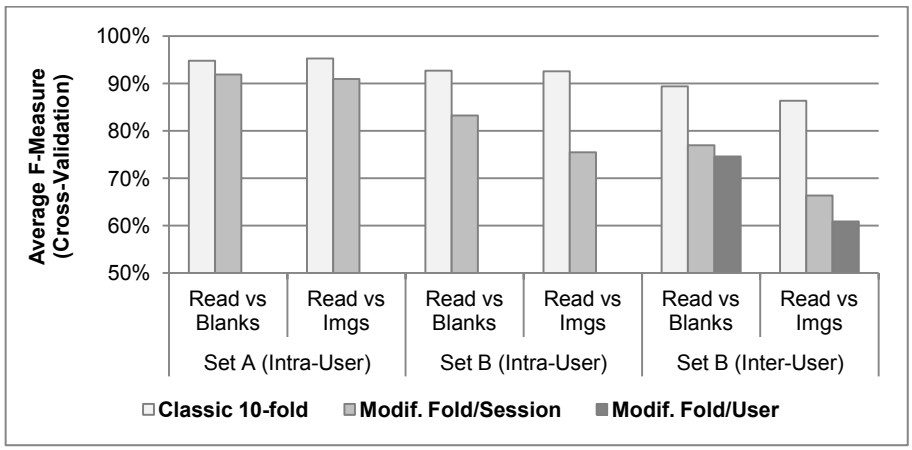


Fig. 4. Silent Reading Detection Results.

We used both classical and modified versions of cross-validation. Cross-validation is an evaluation procedure, commonly applied while using supervised classifiers. These classifiers learn upon some correctly class labeled training data set, requiring a previous training step. This causes the results to depend on the selection of the training and test sets. Cross-validation[10] minimizes this dependency, making this selection

[10] Cross-validation splits randomly the available data set in similarly sized sub-sets, called folds, 10 in our case. It then performs 10 runs of train and test classification procedures, training with 9 folds and testing with the remaining one. The final result is the average of all classification iterations.

more trustworthy. Modified schemes test the result generalization to distinct sessions and users[11].

Classic cross-validation schema results in Set A, 94,77% and 95,22%, suggest that there is a clear distinction between the silent reading mental state and alternative patterns. This also attests that the feature vectors of both classes clearly fall into different areas of the solution space, which is split in two by SVM. Session generalization results, 91,90% and 90,93%, show a slight decay of the classifier performance, but are promising towards the possibility of training the classifier with previously recorded sessions.

Inter-user Set B classic 10-fold results were above 85%, which indicates that the reading pattern is also detected in this case but session and user generalization performed poorly. User generalization can only be possible if both mental states are very similar in all users. Session generalization is also affected by this, since we have sessions belonging to distinct users.

6.2 Classical Measures Correlation

Our main goal is to successfully relate EEG based measures with classical methods measures, such as reading performance and inquiry data, while varying readability aspects such as text size and background/text contrast (Experiments 3 to 6).

We considered that a possible measure could be **mental workload** − beta/alpha ratio, which has been proved to be related with mental workload and user discomfort [10,11]. Beta rhythm (13-30Hz) has been related with mental activity and alpha (8-13Hz), to the mental rest in usability related experiments. So when user's mental workload is high the amount of alpha rhythm decreases, and the amount of beta and beta/alpha increases [10,11].

Mental Workload measure was first determined in each channel, using one seconds segments with 0.5 seconds overlapping, and then averaged twice. First it was spatially averaged in all channels to obtain the mean overall mental workload. Next it was temporally averaged in all samples belonging to a distinct experimental step, which we consider to be a reading situation where the text characteristics, such as size, contrast or color difference, remain constant. Next two sections discuss the results obtained while relating this measure with classical method measures, starting by performance based heuristics, directly determined from the text characteristics.

Performance Based Measures Correlation.

We considered that the following two heuristics could approximate reading performance by changing some text or background relevant property (e.g. text size):

- **Reading performance,** inversely proportional to the considered aspect, that is:

$$\text{Reading\_performance}(\text{step}_i) = \text{aspect}(\text{step}_i) / \text{aspect}(\text{step}_1) \quad (1)$$

[11] The folds of these schemes coincide with sessions and users, which means that we are testing if the classifier can be trained with sessions and users, distinct from the sessions and users tested.

- **W3C related ranking**, where the difference of the aspect regarding W3C orientations was quantified as

$$\text{W3C related ranking}(\text{step}_i)=\begin{cases} 0,\text{if diff}(\text{state}_i) < 0\% \\ 1,\text{ if } 0\% \le \text{diff}(\text{state}_i) < 35\% \\ 2,\text{if } 35\% \le \text{diff}(\text{state}_i) < 70\% \\ 3,\text{if } 70\% \le \text{diff}(\text{state}_i) < 105\% \\ 4,\text{if diff}(\text{state}_i) \ge 105\% \end{cases} \quad (2)$$

where

$$\text{diff}(\text{step}_i)=\text{aspect}(\text{step}_i)/\text{W3C threshold} - 1 \quad (3)$$

Both heuristics consider that the initial step always assures a better readability that the following steps.

Next table shows the results of the correlation analysis between these performance heuristics and the mean workload in all sessions. The use of averaging in EEG is widely disseminated in order to reduce variability, and in this case results are not definitive but can show a trend.

Experiment		Heuristic	Reading Performance	W3C Ranking
Text Size Decrease		CORR	☑0,794	☑0,906
		PVAL	0,033	0,005
Back./Text Contrast Decrease	**Varying Back.**	CORR	☒ -0,134	☒-0,097
		PVAL	0,800	0,856
	Varying Text	CORR	☒-0,196	☒-0,107
		PVAL	0,710	0,840
Back./Text RGB Difference Decrease		CORR	☑0,722	☑0,716
		PVAL	0,105	0,109

Table 5. Correlation of the mean step workload with performance heuristics (inter-user Set B).

Correlation measures the probability of existing a linear relation between two measures; when it is close to 1 or -1, it is considered very strong; when is 0, it doesn't exist. In this case we are using Pearson correlation: PVAL is the p-value or the probability of the correlation being null; when it is below 0.05 or at least 0.1 it means that the conclusion is probabilistically relevant. Correlation and un-correlation evidence is signaled in the table through ☑ and ☒; the remaining values are inconclusive but point towards a possible correlation.

The results displayed show that mean mental workload is highly correlated with text size aspect variation, but the same conclusion does not hold in the remaining experiments. Both contrast experiments show significantly uncorrelated values, which means that mental workload very likely background/text contrast decrease is not linear related with the contrast difference itself. However, RGB difference shows a no significant correlation with both performance metrics.

Inquiry Based Measures Correlation.

The same methodology was applied to the inquiry data that was registered both previously and during experimental sessions. This includes indicators of user fatigue and reading stop, which were approximated with the following two heuristics:

- **Fatigue State:** rates the user perceived fatigue state:

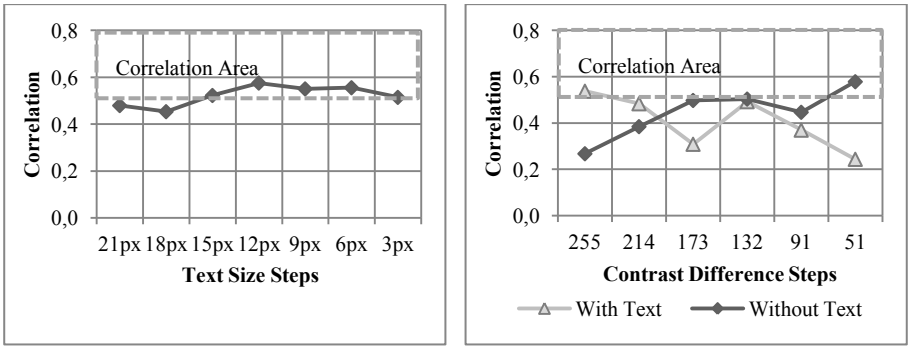
$$\text{Fatigue\_State}(\text{trial}_i)=\begin{cases}1, \text{completely awake}\\ 2, \text{awake}\\ 3, \text{lightly tired}\\ 4, \text{tired}\\ 5, \text{very tired}\end{cases} \quad (4)$$

- **Reading (Occurrence) State**: signals whether reading has occurred or not in a certain experimental step:

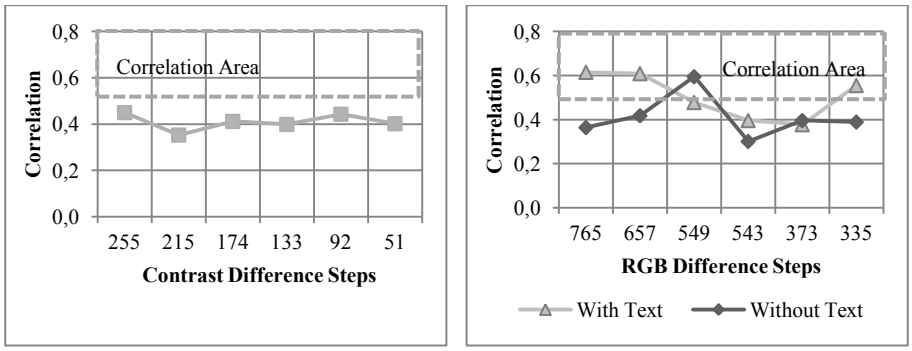
$$\text{Reading\_State}(\text{step}_i)=\begin{cases}1, \text{read}\\ 0, \text{didn't read}\\ 0.5, \text{read partially}\end{cases} \quad (5)$$

Figure 5 presents graphically the correlation variation between the fatigue heuristic and the average mental workload in each step of experiments 3-6, also in inter-user Set B. For simplicity sake we omit the p-value (PVAL) and just display correlation values. All grayed areas signal correlations above 0.5. The significance of the correlations will be referred in context.

The obtained results show that the correlation with reported fatigue varies with the readability aspect being considered. In general texts resize experiment steps are significantly correlated, contrast steps, uncorrelated and color difference, insignificantly correlated. For example, Graphic a) shows a significant trend of a positive correlation between fatigue reported state and mental workload when text size gets smaller. This also means that minor size text reading implied a greater mental workload in users that reported to be more fatigue.



a) Text Resize Experiment b) Back./Text Contrast (Varying Back.)



c) Back./Text Contrast (Varying Text) d) Back./Text RGB Difference

Fig. 5. Fatigued state-Mental workload correlation results (inter-user Set B).

Regarding RGB difference, it revealed some positive correlation in some color combinations, and contrast variation experiments were generally uncorrelated with fatigue reported state. When RGB difference lowers below the W3C defined threshold (in step 4) correlation visibly decreases, indicating this may cause a greater mental workload in more fatigue users (see Graphic d).

Additionally, graphics b) and c) show a similar correlation values in the initial steps, in line with the fact that both experiments start with the same text and background colors. And also in graphics b) and d) one can verify that there are mental workload differences when user reads or does not read, even when the backgrounds are the same.

Table 6 presents the correlation results regarding the remaining inquiry based heuristic: reading (occurrence) state. These were performed in some specific steps (usually the last) of the described experiments also in inter user Set B. In text size experiment we use the last but one step, because in the last step all users reported that they couldn't read the text[12]; in RGB difference, users declared to read in all steps.

Experiment		Step	CORR	PVAL
Text Size		6) 6px	☑0,732	0,0162
Background/Text Contrast	Varying Background	6) 51	☒0,027	0,933
	Varying Text	6) 51	☒0,028	0,933
RGB Difference		NA		

Table 6. Read occurrence-Mental workload correlation results (inter-user Set B).

These results point that there is a correlation between reading occurrence and mental workload only in text variation experiment. As it was told before this requires further study because this can be text size specific.

[12] Correlation with constant functions cannot be mathematically determined.

7 Conclusions and Future Work

The integration of low cost and feasible neurophysiologic methods in the usability analysis framework is possible and can be effective in some specific conditions. This is the main conclusion that can be drawn from the suite of experiences reported in this paper.

Based on the mapping of usability heuristics or guidelines onto specific cognitive processes, such as reading in our case, and on the appropriate feature extraction and measurement, we can observe neurophysiologic changes of the user that are directly correlated with the manipulation of typical usability conditions of an interface. In other words, the analysis of the EEG signals performed by an appropriate and computationally sensible processing chain can make convincingly discriminating decisions concerning reading states and corresponding readability analysis.

The effectiveness of this method for a single user is clearly demonstrated in the classification results that were presented. Its generalization for different users is still an open question, and suggests further work in, at least two directions. The first, aiming at generalization, is the increase of the corpus size that may lead to the discovery of a significant average pattern for the cognitive state under observation. The second, in the opposite direction of personalization, is the design of a calibration procedure that will lead to different baselines for different users, and therefore different classification thresholds.

This paper also discussed the relation of mental workload, an EEG based measure, and reading performance heuristics or inquiry based measures, based in readability aspects variation such as text size, contrast and color. The results obtained so far are promising, indicating that it is possible to successfully relate both types of measures in some of the experiments, and also in some legibility aspects. A greater corpus, with more users and more sessions with the same users are required to deal more effectively with EEG variability.

Acknowledgements

This work was partially supported by Fundação para a Ciência e Tecnologia (FCT), Portugal, Grants SFRH/BD/30681/2006 and PPTDC/EIA-EIA/113660/2009.

References

1. Dix, A., Finlay, J.E., Abowd, G.D. and Russell, B.: Human Computer Interaction, 3rd Ed, Pearson Education Limited, ISBN-13: 978-0-13-046109-4 (2004).
2. Schmorrow, D., Kruse, A., Reeves, L., Bolton, A.: Augmenting Cognition in HCI: 21st Century Adaptive System Science and Technology, In *The Human–Computer Interaction Handbook*, 2nd Ed, Sears, A. and Jacko, J.A. (eds), pp 1172-1188, Lawrence Erlbaum Associates, ISBN-13: 978-0-8058-5870-9 (2008).

3. Zander, T.O., Lehne, M., Ihme, K., Jatzev, S., Correia, J., Kothe, C., Picht, B. and Nijboer, F.: A dry EEG-system for scientific research and brain-computer interfaces, *Frontiers in Neuroscience*, Vol. 5, art. 53 (2011).
4. Beer, R., Lehman, W., Noldus, L., Patèrno, F., Schmidt, E., Hove, W., Theuws, J.: The Usability Lab of the Future, *Human Computer Interaction - INTERACT'03* (2003).
5. Ganglbauer, E., Schrammel, J. e Deutsch, S.; Applying Psychophysiological Methods for Measuring User Experience: Possibilities, Challenges and Feasibility, *Proc. in User Experience Evaluation Methods in Product Development* (UXEM'09) in conjunction with Interact'09(2009).
6. Hu, J., Nakanishi, M., Matsumoto, K., Tagaito, H., Inoue, K., Shima, K., Torii, K.: A Method of Usability Testing by Measuring Brain Waves, *Proc. of the International Symposium on Future Software Technology* (ISFST-2000) (2000).
7. Foglia, P., Prete, C., Zanda, M.: Relating GSR Signals to traditional Usability Metrics: Case Study with an anthropomorphic Web Assistant, *IEEE Intnl Instrumentation and Measurement Technology Conf.,* Victoria, Vancouver Island, Canada, May 12–15 (2008) .
8. Stickel, C., Ebner, M., Steinbach-Nordmann, S., Searle, G. e Holzinger, A.: Emotion Detection: Application of the Valence Arousal Space for Rapid Biological Usability Testing to enhance Universal Access, *Proc. of UAHCI '09 - 5th Intnl Conf. on Universal Access in Human-Computer Interaction, Springer-Verlag* (2009)
9. Stickel, C., Scerbakov, A., Kaufmann, T., Ebner, M.: Usability metrics of time and stress-biological enhanced performance test of a university wide learning management system, *HCI and Usability for Education and Work*, USAB 2008, Springer, p.173-184, ISBN 978-3-540-89349-34 (2008).
10. Kimura, M., Uwano, H., Ohira, M. e Matsumoto, K.: Toward Constructing an Electroencephalogram Measurement Method for Usability Evaluation, J.A. Jacko (Ed.): *Human-Computer Interaction*, Part I, HCII 2009, LNCS 5610, pp. 95–104, Springer-Verlag (2009).
11. Masaki, H., Ohira, M., Uwano, H., e Matsumoto K.: A Quantitative Evaluation on the Software Use Experience with Electroencephalogram, A. Marcus (Ed.): *Design, User Experience, and Usability*, Pt II, HCII 2011, LNCS 6770, pp. 469–477, Springer-Verlag (2011).
12. Hirshfield, L, Chauncey, K., Gulotta, R., Girouard, A., Solovey, E., Jacob, R., Sassaroli, A., Fantini, S.: Combining Electroencephalograph and Functional Near Infrared Spectroscopy to Explore Users' Mental Workload, *FAC '09 Proc. of the 5th Intnl Conf. on Foundations of Augmented Cognition, Neuroergonomics and Operational Neuroscience*, HCI International 2009, pp 239-247, Springer-Verlag (2009).
13. Lee, H., Seo, S.: A Comparison and Analysis of Usability Methods for Web Evaluation: The Relationship Between Typical Usability Test and Bio-Signals Characteristics (EEG, ECG). *Proc. of 2010 DRS (Design Research Society) Montreal Conference*, ISBN 978-2-9811985-2-5 (2010).

14. Nacke, L. :Wiimote vs. Controller: Electroencephalographic Measurement of Affective Gameplay Interaction, *Proc. of the Intnl Conf. on the Future of Game Design and Technology Futureplay*, ACM Press (2010).
15. Ward, R. e Marsden, P.: Affective computing: problems, reactions and intentions, *Journal of Interacting with Computers*, Vol. 16, Issue 4, pp. 707–713, Human Computer Interaction in Latin America, Elsevier (2004).
16. Winograd, T. and Flores, F.: Understanding Computers and Cogniton: a New Foundation for Design, Addison Wesley (1986).
17. Nielsen, J., and Molich, R.: Heuristic evaluation of user interfaces, Proc. ACM CHI'90 Conference, Seattle, WA, April (1990).
18. Shneiderman, B.: Designing the User Interface, Strategies for Effective Human-Computer Interaction, Addison Wesley- 3rd edition, ISBN 0-201-69497-2 (1997).
19. Tognazzini, B.: First Principles of Interaction Design, AskTog – Interaction Design Solutions for the Real World, NN/g – Nielsen Norman Group, 18.01.2012, AskTog: http://www.asktog.com/basics/firstPrinciples.html (2012).
20. Oliveira, I., Grigore, O. e Guimarães N.: Reading detection based on electroencephalogram processing, *WSEAS 13th Intnl Conf. on Computers*, Rhodes, Greece (2009).
21. Oliveira, I., Grigore, O. e Guimarães, N. e Duarte, L.: Relevance of EEG Input Signals in the Augmented Human Reader, *ACM – Augmented Human - AH'10*, April 2010, Megève, France (2010).
22. Wolpaw, J.R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H. , Schalk, G., Donchin, E., Quatrano, L. A., Robinson, C. J., and Vaughan, T. M.: Brain–Computer Interface Technology: A Review of the First International Meeting, *IEEE Transactions on Rehabilitation Engineering*, Vol. 8 (2000)
23. Millán, J.R.: Adaptative Brain Interfaces, *Communications of the ACM*, vol 46, Issue 3, 74:80, ACM, March (2003).
24. Zander, T.O., Gartner, M., Kothe, C., and Vilimek, R.: Combining Eye Gaze Input with a Brain-Computer Interface for Touchless Human-Computer Interaction, *International Journal of Human Computer Interaction*, 27:1, 38-51 (2011)
25. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection, *The Journal of Machine Learning Research*, Vol. 3, pp 1157-1182 (2003).
26. Xu, Q., Zhou, H., Wang, Y. and Huang, J.: Fuzzy support vector machine for classification of EEG signals using wavelet-based features, Medical Engineering & Physics 31, pp. 858–865 (2009).
27. W3C: CSS Techniques for Web Content Accessibility Guidelines 1.0, Obtained in December 2011, from W3: http://www.w3.org/TR/WCAG10-CSS-TECHS (2010)
28. W3C: Fonts, Obtained in December 2011, from W3: http://www.w3.org/TR/CSS2/fonts.html (2011)
29. Fisch, B.: Fisch and Spehlmann's EEG Primer: Basic Principles of Digital and Analog EEG, Elesevier, ISBN 978-0-444-82148-5 (1999).