Unconscious Physiological Effects of Search Latency on Users and Their Click Behaviour

Miguel Barreda-Ángeles
Barcelona Media
Barcelona, Spain
miguel.barreda@barcelonamedia.org

Ioannis Arapakis
Yahoo Labs
Barcelona, Spain
arapakis@yahoo-inc.com

Xiao Bai
Yahoo Labs
Barcelona, Spain
xbai@yahoo-inc.com

B. Barla Cambazoglu
Yahoo Labs
Barcelona, Spain
barla@yahoo-inc.com

Alexandre Pereda-Baños
Barcelona Media
Barcelona, Spain
alexandre.pereda@barcelonamedia.org

ABSTRACT

Understanding the impact of a search system's response latency on its users' searching behaviour has been recently an active research topic in the information retrieval and human-computer interaction areas. Along the same line, this paper focuses on the user impact of search latency and makes the following two contributions. First, through a controlled experiment, we reveal the physiological effects of response latency on users and show that these effects are present even at small increases in response latency. We compare these effects with the information gathered from self-reports and show that they capture the nuanced attentional and emotional reactions to latency much better. Second, we carry out a large-scale analysis using a web search query log obtained from Yahoo to understand the change in the way users engage with a web search engine under varying levels of increasing response latency. In particular, we analyse the change in the click behaviour of users when they are subject to increasing response latency and reveal significant behavioural differences.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Web search engine; response latency; user study; physiological signals; user behaviour; query log analysis; click behaviour

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.
© 2015 ACM. ISBN 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767719.

1. INTRODUCTION

Improving the efficiency of search systems has been an active research area in the last few decades. So far, most research in this direction had a very system-oriented viewpoint, specifically focusing on reducing the financial costs involved in the search process. Until recently, the impact of efficiency improvements on users' searching behaviour and experience were left unexplored. However, with the ever-growing competition in online search systems for attracting users and the increasing emphasis on both short- and long-term user engagement, this has started to change. In particular, a new line of research was born to investigate the impact of search response latency on user behaviour and engagement. Our work is an addition to this line of research. Herein, we specifically focus on the physiological effects of latency increase on users and are especially interested in the cases where the latency increase is not very large.

Research on human information processing has consistently demonstrated that human beings are not consciously aware of the mental processes determining their behaviour [21, 23]. Such unconscious influences do not need to be restricted to basic or low-level mental processes, but can also reach high-level psychological processes like motivations, preferences, or complex behaviours [5]. This has obvious implications when it comes to the assessment of user experience in human-computer interaction (HCI) contexts. For example, previous research in the context of web search has shown that response latency values lower than a certain threshold are unnoticeable by the users and, therefore, inconsequential in terms of user experience [1]. However, the conclusions drawn in earlier studies are based on self-reported methods (i.e., methods in which the users are explicitly asked to assess their experience), which are inherently limited as, obviously, the users cannot verbally report information that is not consciously available to them. Therefore, the possibility that even small latency increases, not consciously perceived by users, could affect the experience of using search engines cannot be completely dismissed.

Psychophysiological methods can be very helpful in unveiling attentional and emotional reactions that are not consciously available to us. During the last years, the use of this kind of methods has spread over different fields such as psychology, marketing, and HCI. Besides the fact that they provide information on unconscious processes, other advantages of these methods include their high temporal and spacial resolution, and their robustness against cognitive biases (e.g., social desirability bias). Hence, they always provide "honest" responses. Nevertheless, when using psychophysiological methods, since there is not a direct question to the subject, there is not a direct answer either. The information on the research questions has to be inferred from the variations on the physiological signals and the way they are related to more or less complex psychological constructs. Therefore, physiological data is often used to draw a picture of the cognitive or emotional state of the user.

In this study, we are interested in the impact of response latency increase on user behaviour in web search and, more specifically, in smaller latency increases that may not be consciously perceived by the users. To this end, we employ two different approaches: a small-scale controlled user study and a large-scale query log analysis. In our controlled study, we use physiological methods (electro-dermal activity, electromyography) to get information on the subtle emotional reactions of the users and how they are affected by increasing response latency. We follow the standard approach to research with physiological methods. This approach relies on the bifactorial model of emotions [7], involving two factors: arousal (the intensity of the emotion) and valence (how positive or negative the emotion is). Although this approach requires the use of physical sensors, the sensors do not invade user's privacy and can capture short-term changes not measurable by other means. In our query log analysis, we focus on the variation in the click behaviour of users. The large size of the query log we deal with enables us to observe effects, which were not easy to observe through a small-scale user study. In general, the two approaches are complementary and allow for a more accurate and holistic investigation of the effects of response latency in web search.

The following summarizes our contributions and findings.

- We perform a small-scale controlled experiment and demonstrate the effects of small increases in response latency, using physiological measures of arousal and valence. We compare and contrast these physiological signals with information gathered from self-reports and show that the former are more effective in capturing the attentional and emotional reactions to increasing response latency. We show that there are sizeable effects on users' physiological reactions, even though such effects are not observed in self-reported measures.
- We conduct a large-scale analysis using query logs obtained from Yahoo Web Search and gain insights about the change in the click behaviour of users. In particular, we analyze how a user's click behaviour is affected when the response latency differs upon submission of the same query by the user at different times (under the constraint that retrieved results are identical). This analysis shows that, on average, even small increases in response latency leads to a statistically significant decrease in click likelihood.

The rest of the paper is organised as follows. Section 2 provides a brief summary of the related work on the user impact of search latency. The details of our controlled user study and query log analysis, together with the experimental results, are presented in Sections 3 and 4, respectively. We conclude the paper in Section 5.

2. RELATED WORK

A recent line of work focused on modelling the effort users spent in searching and the corresponding gains in interactive information retrieval [2, 3, 28]. However, in these studies, the response latency was not the main subject under investigation. A relatively older line of research looked into the impact of page load time on the web browsing behaviour of users [9, 11, 12, 13, 20, 24, 30]. A more relevant line of research studied the impact of response latency on user behaviour in the web search context [1, 8, 18, 26, 31]. Herein, we survey only this latter line of research as it has a much higher overlap with the scope of our work.

Schurman and Brutlag [26] conducted a bucket test with live search engine traffic, exposing web search users to different response latency levels. Their study revealed that users who were exposed to higher response latency issued fewer search queries compared to a control group with no added latency. In particular, the authors observed that increasing the response latency by 400ms led to a drop of 0.59% in the daily search volume of affected users. They also observed that the negative effect on the search volume was persistent for a certain period of time even after the latency had returned to the original levels. The study of [26] relied on server-side delays, essentially modelling only the additional server processing time. In our user study, we use client-side delays allowing us to have a greater control on the end-to-end latency actually experienced by the users. Moreover, in our query log analysis, we analyse the impact of latency on click behaviour of users, instead of their querying behaviour.

Arapakis et al. [1] conducted a controlled user study, where participants were exposed to different response latency levels while querying two web search engine front-ends. The study revealed interesting differences in the way users perceived the latency. For example, the users of a fast search engine were found to be more likely to notice the added delays than the users of a slow search engine. In most cases, the users could not notice added latency delays below 500ms. However, when the added delays were above 1,000ms, the users were very likely to notice them. Significant differences were also observed in the reported positive and negative affect scores at post-task, between the fast and slow search engines. The authors also reported a noticeable decrease in the perceived system usability as latency values increased. Hence, the tendency to overestimate or underestimate system performance biased users' interpretations of search interactions and system usability. The work of Arapakis et al. [1] also involved a large-scale query log analysis that shows the change in the click behaviour of users due to increasing latency. The authors showed that, when the response latency is too high, people prefer to browse the returned results on the search engine result page instead of issuing new queries to the search engine. In our query log analysis, we adopt a slightly different experimental methodology than the one used in [1] and also focus on the effects at much lower increases in latency.

A study conducted by Teevan et al. [31] involved a query log analysis on the impact of increasing response latency on user engagement. Similar to the finding in Arapakis et al. [1], with increasing response latency, the authors observed a decrease in the likelihood that the user will click on the search results. Moreover, the authors observed an increase in the time the users issue the first click on the search result page as the response latency increases. Both findings point out the negative consequences of slow response on user engagement.

Brutlag et al. [8] conducted a user study where the participants interacted with two search interfaces serving their results at controlled latency values (one with low latency and one with high latency). The users stated their preferences between the two interfaces through a questionnaire. The findings of the study were mostly inconclusive regarding the impact of response latency on user preferences. Our work complements the work of Brutlag et al. by showing that self-reported studies are not adequate to demonstrate the effects of increasing latency, but physiological effects on users do exist and can be measured by other means.

Maxwell and Azzopardi [18] verified the validity of five different hypotheses (taken from information foraging and search economic theories) about how users' search behaviour should change when faced with delays. The study involved 48 participants who interacted with four different search interfaces with different types of delays (no delay, only query response delay, only document download delay, and both query response and document download delays). The study found strong support for the three of the hypotheses.

With respect to the above-mentioned bucket test [26], user studies [1, 8, 18], and query log analyses [1, 31], the main novelty of our work is in the reported psychophysiological measurements that aim to reveal the unconscious effects of small response latency increases on users. Our work shows that there are sizeable effects on users' physiological reactions to increasing response latency and physiological measurements are more effective in capturing these reactions than the information gathered by self-reported studies. To the best of our knowledge, no such measurements were reported in literature before.

3. CONTROLLED USER STUDY

To understand the impact of increasing response latency on users' search behaviour, we carried out a controlled experiment in which we examined users' interactions with Yahoo Web Search. The goal of the study was to demonstrate the effects on users' experience (more specifically, on their engagement with the search engine and satisfaction with the provided service). We note that most of the response latency values used in this study were lower than what the literature has claimed as the threshold for conscious detection of the delay [1]. The reason for that is because we want to show that, even if these small latencies are unnoticeable, they produce observable reactions in the users. Applied to our object of interest, it seemed logical to expect that if high response latency produces a worse emotional experience in the users, we would find a more negative (in the valence domain) and more intense (in the arousal domain) emotional reaction. Moreover, we combined the measurement of physiological aspects of emotions with self-reported measures of affect, attention, and usability of the system, all known constituent of engagement [19, 22]. This approach allowed us to explore both conscious and "not-necessarily" conscious aspects of the user experience and obtain an accurate picture of how lower or higher latency values affect it.

3.1 Experimental Design

The experiment used a repeated-measures design with one independent variable: search latency (with four levels in milliseconds: 0, 500, 750, and 1,000). The search latency was controlled using a client-side script that adjusted the latency by a desired amount of delay. The dependent variables were (i) experienced positive and negative affect, (ii) level of focused attention, (iii) perceived system usability, and (iv) participants' physiological responses.

Each of the participants carried out four search tasks, one for each of the four latency conditions. To control for order effects, we randomised the task assignment. The search tasks involved submitting as many navigational queries as possible, selected out of a list of 200 web domains, and retrieving the associated URL within eight minutes.[1] Compared to other types of queries (e.g., informational or commercial), navigational queries introduce a smaller cognitive load to the searcher and promote a convergence in the search intent across all users. An additional advantage is that they do not require native-level knowledge of the language used. Therefore, by mitigating the effort of query formulation, our participants were able to experience the latency effect more effectively. Other types of search queries and levels of task complexity may interact differently with the impact of search latency. However, we leave this investigation for future work. The web domains for constructing the navigational queries were selected from Alexa.[2]

3.2 Apparatus

We used a desktop computer equipped with an LCD monitor, a keyboard, and a mouse. In the background, we ran a custom-made JavaScript code that allowed us to control the search latency. Moreover, the script captured a series of browser events (e.g., mouseover, click), and recorded the timestamps of every submitted query and search result page rendered in response to a query. The script was deployed using the Greasemonkey extension[3] in a Mozilla Firefox web browser. The physiological data were recorded using a Biopac MP-150 Data Acquisition System,[4] using a sampling rate of 1000Hz. The AcqKnowledge 4.2 software was used for the storage and processing of the data while an in-house application, written in Python, was used to synchronise the physiological data with the query submission times.

3.3 Psychophysiological Measures of Engagement

To analyse the emotional reactions of participants, we recorded two types of physiological measures: (i) electrodermal activity (EDA) and (ii) electromyography (EMG). EDA, which refers to the conductivity of the skin, varies according to the activation of the sympathetic branch of the autonomous nervous systems and has been commonly used as an index of emotional arousal [10, 25]. Given this, different aspects of the EDA signal such as the general level of skin conductance or the amplitude or latency of momentary skin conductance responses (SCRs) can be used [10]. In our study, we examined the SCRs elicited in an interval of up to 10 seconds after each query submission. Considering that the SCRs associated to a certain stimulus can take up to 6 seconds from stimulus onset to the point where it reaches the peak of the response, this 10-second time window was considered enough to capture the SCRs related to queries.

[1] We define navigational queries as those that seek a single website or the web page of a single entity.

[2] `http://www.alexa.com/topsites`

[3] `http://www.greasespot.net`

[4] `http://www.biopac.com/`

Table 1: I-PANAS-SF [32]

Positive Affect items	Negative Affect items
active	afraid
alert	ashamed
attentive	hostile
determined	nervous
inspired	upset

Another physiological method commonly used to quantify the valence of the stimuli is the activity of facial muscles, measured through EMG techniques. More specifically, the activity over the *corrugator supercilii* (EMG-CS), a muscle on the eyebrow responsible for frowning, was used as a proxy for the negative valence of the experienced emotions [25]. Since the muscular response appears to be immediate after the stimuli onset, we considered a time window of 3 seconds after each query submission as good enough to capture the possible changes in the EMG activity. As we anticipated that higher latency values would lead to an emotionally negative experience for the users, and this would manifest itself as higher levels of arousal and as negative valence, we also predicted higher levels of SCR and more intense EMG-CS activity in response to higher latency values.

3.4 Self-Reported Measures of Engagement

In our study, we used two types of questionnaires. The first questionnaire was introduced at the beginning and inquired about demographic information. The second questionnaire was administered at post-task and included the User Engagement Scale (UES) and the Computer System Usability Questionnaire (CSUQ). UES [22] is a multi-dimensional survey instrument that measures user engagement with technology. More specifically, it examines the cognitive (felt involvement, focused attention, perceived usability) and affective (positive and negative affect) aspects of interactions. Affect accounts for the hedonic experiences, as well as the motivations that influence and sustain our engagement during computer-mediated activities. The Felt Involvement items gauge users' feelings of being drawn into the experience, while the Focused Attention (FA) scale pertains to being absorbed or losing the track of time. CSUQ [17] is a multi-dimensional user satisfaction questionnaire, designed for use in scenario-based usability evaluations. Out of the four scales it contains, we considered only the scores from system usefulness (SYSUSE). Combined together, UES and CSUQ probe users' perceptions of the pragmatic and hedonic qualities of their interactions, as well as their perceptions of the search engine, all of which are considered key aspects of the user experience [15]. The questions were all forced-choice type and appeared in a random sequence to reduce potential ordering effects.

I-PANAS-SF. The International Positive and Negative Affect Schedule Short Form (I-PANAS-SF) [32] was used to measure the affect at pre-task and post-task (Table 1). I-PANAS-SF is a 10-item version of PANAS [33] that measures affect changes. It includes 5 items measuring positive affect (PAS) and 5 items measuring negative affect (NAS). Participants were asked to respond on a 7-point Likert scale (1: very slightly or not at all, . . . , 7: extremely) their agreement to the statement: "You feel this way right now, that is, at the present moment", for each item. Although I-PANAS-SF may not be as efficient and accurate for capturing temporal micro-resolutions of emotional responses, there are several examples of studies from the domain of library and information science [14, 22], where PANAS has been successfully applied for measuring searchers' affect between search tasks. Given that the duration of our search tasks is comparable to those in the aforementioned studies, and considering the effectiveness of self-report methods in general, we believe that our experimental approach for measuring emotions was reasonably accurate.

Table 2: Focused attention scale [22]

1. I forgot my immediate surroundings while performing the search task.
2. I was so involved in my search task that I ignored everything around me.
3. I lost myself in this search experience.
4. I was so involved in my search task that I lost track of time.
5. I blocked out things around me when I was completing the search task.
6. When I was performing this search task, I lost track of the world around.
7. The time I spent performing the search task just slipped away.
8. I was absorbed in my search task.
9. During this search task experience I let myself go.

Focused attention. FA is a 9-item subscale, part of a larger scale for measuring user engagement [22]. FA has been used in past studies [19, 22] to evaluate users' perceptions of time passing and their degree of awareness about what took place outside of their interaction with the given task. Herein, it is adapted to the context of our search tasks. Given the context of our work, FA was a more meaningful dimension, at least compared to other subscales of engagement (e.g., aesthetics, novelty) that were not relevant enough or were addressed by the other questionnaires employed in our study (I-PANAS-SF, CSUQ). To measure FA, the participants were instructed to report on a 7-point Likert scale (1: strongly disagree, . . . , 7: strongly agree) their agreement to each item shown in Table 2.

System usability. CSUQ [17] was developed by IBM to measure the perceived usability of systems in realistic scenarios. A 7-point Likert scale of agreement (1: strongly disagree, . . . , 7: strongly agree) that quantifies system usefulness is used for each of the eight statements in the SYSUSE subscale. Two example statements are "I am able to complete my work quickly using this search site" and "I am able to efficiently complete my work using this search site".

3.5 Participants

There were 19 participants (female=2, male=17) aged from 18 to 41 and free from any obvious physical or sensory impairment. The participants were of mixed nationality (e.g., Spanish, Turkish, Russian), came from a variety of educational backgrounds (10.5% had a high school diploma, 31.5% had a BSc or college degree, 34.8% had an MSc degree and 21.0% had a PhD degree), and were all proficient with the English language (10.5% intermediate level, 84.2% advanced level, 5.2% native speakers).

3.6 Procedure

The user study was carried out in a laboratory setting. At the beginning of each session, the participants were informed about the conditions of the experiment and were asked to complete a demographics questionnaire. Upon completing the demographics questionnaire and the pre-task I-PANAS, the electrodes for the measurement of the physiological sig-

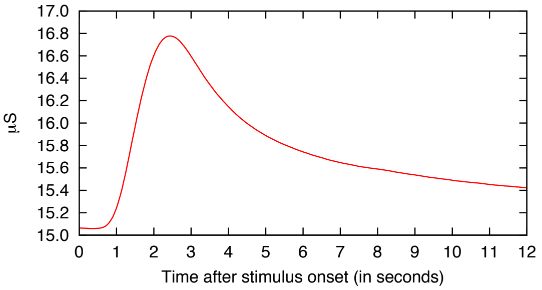

Figure 1: A typical SCR (taken from a participant).

nals were fitted. Two electrodes were placed on two fingers of the non-dominant hand in order to record the EDA signal, while two more electrodes were placed over the eyebrow in the non-dominant side, at the positions described in [29].

Each participant performed four search tasks, one for each latency condition. During the search tasks, the participants were presented with two web browser windows: the first window displayed the search engine used for the search task while the second window displayed the questionnaire. For each navigational query, participants were instructed to locate the associated URL among the first ten results of the search result page and copy-paste it in the corresponding box of the questionnaire. The queries were submitted to the search site the same way as in a realistic search scenario, i.e., by typing and clicking. 1-minute breaks were introduced at the end of each search task, after completing the I-PANAS, CSUQ, and FA scales. At the end of the session, the electrodes were removed, and the participants were debriefed. To motivate the participants, they were informed that they would receive a gift card whose value ranges from 5€ to 10€ depending on the number of queries they manage to submit. In the end, all participants received a 10€ gift card.

3.7 Physiological Data Filtering, Reduction and Analysis

3.7.1 EDA Signal

A smoothing filter with a time window of 200ms was applied to the EDA signal, and a visual inspection was performed to confirm that no artifacts were present in the EDA recordings. With the remaining data, a temporal series was constructed for every physiological signal by averaging the recorded data of every 1-second period, thus producing a temporal series of 480 points for each physiological signal, per participant and per condition. Then, each 10-second period of the EDA signal following a query submission was visually inspected to determine whether an SCR was produced. For determining the presence of an SCR, we examined the data for spikes on the EDA levels. Such spikes would need to begin at the first seconds following a query submission and match the typical form of an SCR, which is characterised by a sharp increase during a time window of length 1 to 3 seconds [10] with a much softer decrease in subsequent seconds (Fig. 1).

A total of 132 SCRs were identified. Participants presented between 1 to 20 SCRs ($M = 7.76, SD = 5.71$) with a mean of 1.94 SCRs by task ($SD = 1.84$). A repeated measures ANOVA showed no significant difference in the number of SCRs by latency condition, ($F(3, 48) = 0.63, p = .6, \eta_p^2 = .04$). Data not related to an SCR following a query submission were excluded from the analysis. Hence, the remaining signal contains ten observations (one observation per second, from seconds 1 to 10) for each of the 132 SCR instances. Next, the EDA value for the timestamp for which the SCR is rendered was used as a baseline and compared with each of the seconds following the query submission to inspect the amplitude of the SCR and to make different SCRs comparable. Thus, the EDA value at second 0 (the moment at which the query is submitted) was subtracted from the EDA values of each posterior second (1 to 10) of the query.

In our data analysis, we applied mixed multilevel models. This type of models present several advantages over other traditional statistical methods (e.g., ANOVA). One such advantage is that they can deal with nested and highly autocorrelated data, such as physiological data [16]. Moreover, the structure in different levels and the possibility of adding random terms allows us to deal with the large variability observed in psychophysiological data. Our model construction approach was similar to the so-called growth modelling [6], in which first null models without predictors are fitted and then both random and fixed factors are progressively introduced to the model. After adding each predictor, a likelihood test is conducted to check whether the new predictor has increased the model fitting [6]. If the model fitting has increased significantly, then the predictor is kept. If the increase is not significant, it is removed. This process is repeated for every factor that can act as a predictor until the model that best fits the data is constructed.

In the case of the EDA data, a visual inspection suggested that, apart from the latency condition, other factors such as the individual differences in participants' mean EDA, the order of presentation of the tasks, as well as the time spent within the tasks could have effects on the EDA signal. Moreover, since the SCR implies an increase followed by a decrease on the EDA levels, the time within each SCR is also likely to have had a significant effect. Thus, all these factors were considered in the construction of EDA models. The final EDA model included a random intercept for participants' means, as well as a random slope for the effects of task presentation order and time within the tasks, indicating that those factors had randomly varying effects between participants. The random intercept suggests that the mean levels of EDA varied among participants. The random slopes for time and order of presentation indicate that the effect of these factors is different (random) among participants. The fixed factors of the model, those that had a fixed effect on the participants, were the latency condition as well as the time within each SCR. The fixed effects are shown in Table 3.

3.7.2 EMG-CS Signal

A band-pass filter between 30–500Hz was applied to the EMG signal [29], and then it was integrated. Visual inspection confirmed the presence of artifacts due to electrode movement or bad adherence to the skin. For those cases, the data of one participant were removed along with the data of another participant, for a specific condition. Then, the EMG-CS signal was averaged for every 1-second window, as in the case of the EDA signal. However, since there is not a particularly defined EMG response, as it is in the case of SCR, we directly included in the analysis the EMG-CS data for the entire 3-second period after each query submission. EMG-CS data were analysed using mixed multilevel models, as in the case of the EDA data.

Table 3: Fixed effects for the EDA and EMG models

EDA model		EMG model	
Fixed factors	Coefficients	Fixed factors	Coefficients
Intercept	$-.31^{*}$	Intercept	$.0188^{***}$
Latency 500ms	$.50^{***}$	Latency 500ms	$.0019^{***}$
Latency 750ms	$.42^{**}$	Latency 750ms	$.0034^{***}$
Latency 1000ms	$.60^{***}$	Latency 1000ms	$.0010^{*}$
Seg 2	$.11^{***}$	Seg 1	.0000393
Seg 3	$.36^{***}$	Seg 2	$.0002397^{***}$
Seg 4	$.68^{***}$	Seg 3	$.0003163^{***}$
Seg 5	$.88^{***}$		
Seg 6	$.90^{***}$		
Seg 7	$.80^{***}$		
Seg 8	$.74^{***}$		
Seg 9	$.72^{***}$		
Seg 10	$.69^{***}$		

* Correlation is significant at the .05 level (two-tailed).
** Correlation is significant at the .01 level (two-tailed).
*** Correlation is significant at the .001 level (two-tailed).

A random intercept for participants' means, as well as random slopes for the order of presentation and the effect of the time within the task were computed, similar to the analysis of the EDA model. The significant fixed factors for the EMG-CS models were the latency conditions, as well as the effect of time within the first 3 seconds after the query submission. During the model construction, one of the assumptions of our analysis, the homoscedasticity of residuals, was violated. This was confirmed by a visual inspection of the residuals. More specifically, the data of three participants showed clear patterns that deviated from the homoscedastic residuals. Those cases were excluded, and our model was fitted again. Table 3 summarises the final EMG-CS model.

3.7.3 *Entropy Analysis*

We also investigated the entropy feature for its performance in discriminating participants' physiological responses to web search results served at different latencies. Entropy has been extensively used in signal processing and pattern recognition. In information theory, entropy measures the disorder or uncertainty associated with a discrete, random variable, i.e., the expected value of the information in a message. The application of the entropy concept for the classification of our data is based on the assumption that the physiological signals' spectrum is more organised during signal segments where the user is carrying out their task without any significant "break-downs" or delays compared to signal segments where the user is experiencing stress or discomfort due to experienced latency increase. Thus, a signal segment of the former type would be characterised by low entropy. More specifically, we compute two entropy-based features for the EDA and EMG-CS data: Shannon entropy and permutation entropy.

Shannon entropy: Shannon entropy [27] allows to estimate the average minimum number of bits needed to encode a string of symbols in binary form (if *log* base is 2) based on the alphabet size and the frequency of symbols. Given a finite time series $X(t) = (x_t\colon 1 \leq t \leq T)$, the Shannon entropy can be expressed as

$$H(X) = \sum_i P(x_i) I(x_i) = -\sum_i P(x_i) \log_b P(x_i), \qquad (1)$$

where I is the information content of X, $I(x)$ is the random variable, and b is the base of the logarithm used.

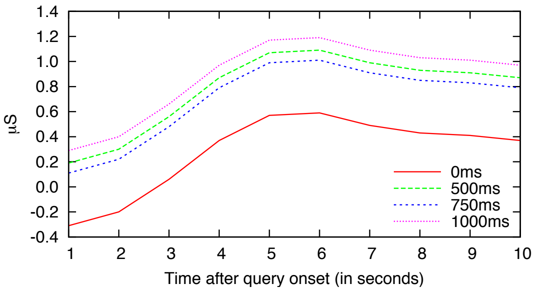

Figure 2: Graphical representation of the values of the fixed coefficients for the EDA model.

Permutation entropy: Permutation entropy [4] provides a fast and robust method for estimating the complexity of time series, by considering the temporal order of the values. More specifically, it calculates the variety of different permutations appearing at the components of a time series.

Let us consider a time series $X(t) = (x_t\colon 1 \leq t \leq T)$. S_n is the set of all possible $n!$ permutations π of order n. For each $\pi \in S_n$ we determine the relative frequency

$$p(\pi) = \frac{|\{t \mid 0 \leq t \leq T-n, (x_{t+1}, \ldots, x_{t+n}) \text{ has type } \pi\}|}{T-n+1}, \qquad (2)$$

which estimates the frequency of π as good as possible for a finite time series. The permutation entropy of order $n \geq 2$ is defined as

$$H(n) = -\sum_{\pi \in S_n} P(\pi) \log p(\pi) \qquad (3)$$

The permutation entropy can be calculated for arbitrary real-world time series, and particularly in the presence of dynamical and observational noise. We compute the entropy for all permutations of order $n \in \{2, 3, 4, 5\}$.

3.8 Results

3.8.1 *Physiological Data*

As we can see in the EDA model (Table 3), the fixed coefficients of the model for the effect of the time after the query describe a curve that approximately fits the form of a typical SCR (Fig. 2). This is not surprising since, during data filtering, queries that are not followed by an SCR were removed, while the remaining data that was analysed contained an SCR after each query submission. What is of interest is that the model can also predict significant increases in the EDA levels associated with the increased latency conditions compared to the no-latency (0ms) condition. The coefficients showed increases of 0.5μS and 0.42μS for the 500ms and 750ms latency conditions, respectively, while a higher increase of 0.60μS was observed in the 1,000ms latency condition. These increases were statistically significant compared to the 0ms condition, which was taken as the baseline in the model for the calculation of the coefficients. Nevertheless, even if the increase associated with the 1,000ms condition seemed to be higher than the increases associated with the 500ms and 750ms conditions, recalculations of the model using different levels of latency as a baseline condition showed that the difference between the coefficients for the three latency conditions (500ms, 750ms, 1,000ms) did not reach statistical significance ($p > .05$), i.e., while the three

Table 4: Means (M) and standard errors (SE) for the post-PAS, post-NAS, FA, CSUQ scales

	Latency condition			
Scale	0ms	500ms	750ms	1,000ms
post-PAS	17.21 ± 1.46	18.21 ± 1.57	18.68 ± 1.51	17.53 ± 1.79
post-NAS	6.42 ± 0.41	6.32 ± 0.54	6.47 ± 0.48	5.95 ± 0.36
FA	29.32 ± 1.71	28.26 ± 2.29	27.95 ± 2.18	26.32 ± 2.00
CSUQ	28.16 ± 1.91	29.37 ± 2.09	27.63 ± 1.81	29.05 ± 1.86

latency conditions are significantly different from the 0ms condition, there is no significant difference between them.

In the case of the EMG-CS data, the model showed that the three higher latency conditions (500ms, 750ms, 1,000ms) resulted in a significant increase in the EMG-CS activity. The largest increase was observed for the 750ms latency condition, while more moderate increases were associated with the 500ms and 1,000ms latency conditions. Recalculations of the model, by changing the condition taken as baseline, showed that the increase in EMG-CS activity related to the 750ms condition was significantly different from the increases associated with the 500ms ($p < .01$) and 1,000ms conditions ($p < .001$), and that there was no significant difference between these two conditions ($p > .05$). Here, we observe that all three latency conditions are associated with emotional experiences that are characterised by a more negative valence. Surprisingly, the 750ms latency condition was found to be associated with the most intense activation of the EMG-CS (i.e., negative emotional valence) compared to the 1,000ms latency condition. However, the EDA and EMG-CS data provide support for the hypothesis that even latencies at low ranges can have a significant effect on the user experience, resulting in intensified, negative emotions.

Finally, we applied the Friedman's ANOVA test to the Shannon entropy and weighted permutation entropy scores computed for the EDA and EMG-CS data ($n \in \{2, 3, 4, 5\}$). With respect to the EDA data, the obtained Shannon entropy scores did not change significantly across the latency conditions ($\chi^2(3) = 3.40, p > .05$). Similarly, the weighted permutation entropy scores did not exhibit a significant difference over the latency conditions, for permutations of order $n = 2$ ($\chi^2(3) = 2.20, p > .05$), $n = 3$ ($\chi^2(3) = 0.33, p > .05$), $n = 4$ ($\chi^2(3) = 0.80, p > .05$), or $n = 5$ ($\chi^2(3) = 0.80, p > .05$). When examining the EMG-CS data, the obtained Shannon entropy scores did not significantly change across the latency conditions ($\chi^2(3) = 0.46, p > .05$). Finally, the weighted permutation entropy scores did not differ significantly across the latency conditions for any of the permutations of order $n = 2$ ($\chi^2(3) = 5.06, p > .05$), $n = 3$ ($\chi^2(3) = 2.86, p > .05$), $n = 4$ ($\chi^2(3) = 0.20, p > .05$), or $n = 5$ ($\chi^2(3) = 0.26, p > .05$). These preliminary findings suggest that the computed entropy features did not contribute significantly to the accurate discrimination of the EDA and EMG-CS data, for the latency conditions we examined. This could be due to the lack of sensitivity towards subtle changes in the data, perhaps measurable more accurately by other means or being affected by additive noise.

3.8.2 *Self-Reported Data*

Participants' responses to the 5-item PAS, 5-item NAS, 9-item FA, and 8-item CSUQ-SYSUSE scales were summed to obtain the final scores (Table 4). Results are reported at a statistical significance level of .05. To take an appropriate control of Type I errors in multiple pair-wise comparisons, we applied the Bonferroni correction. The ANOVA test was applied to determine if there were statistically significant differences between the scores assigned to different latency conditions. The results showed no significant differences in the CSUQ ($F(3, 54) = .76, p = .52, \eta_p^2 = .1$) or the FA scores ($F(3, 54) = .79, p = .5, \eta_p^2 = .04$). Similarly, no significant effect was observed for post-NAS ($F(3, 54) = .93, p = .43, \eta_p^2 = .05$) or post-PAS ($F(3, 54) = 1.23, p = .33, \eta_p^2 = .19$). These results suggest that, contrary to the clear effects found in the physiological data, the users did not consciously perceive the experience as better or worse, depending on the latency of the search engine response. However, the absence of significant effects in the users' responses to the self-reports should not be considered as a definite proof of the absence of latency effects on the user experience, as clearly indicated by the physiological data findings. In addition, the data from the FA scale reveals a pattern (Fig. 3) that suggests, although not statistically significant, a reduced engagement as the latency increases to higher values. But, given a larger sample, it is possible that a weak effect of the search latency on the reported FA scores may become more evident.

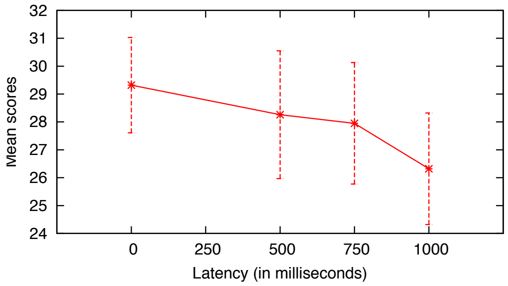

Figure 3: FA levels across all latency conditions.

4. QUERY LOG ANALYSIS

By exploiting physiological measures, the previous study has demonstrated the presence of certain latency effects that are unconsciously experienced by the users. In this section, we investigate the effects of similar small latency increases on the engagement of users with search engine result pages. In particular, we observe the change in users' likelihood of clicking a search result when presented with a fast or slow response. Unlike the previous study, which involved a small number of participants, this analysis is performed using prerecorded search click data obtained from Yahoo Web Search. Hence, it allows us to demonstrate, at large scale, that even small latency increases can lead to a statistically significant decrease in the engagement of a user with a search engine.

4.1 Setup

Click data. We randomly sampled a large number of queries from the web search query log (all queries were submitted on the same day). From this sample, we further selected queries that were (i) issued from desktop computers located in the US and (ii) processed in a specific search data centre in the US. The motivation behind the first filtering was to eliminate a potential bias due to device differences (e.g., desktop, mobile, or tablet) while the second filtering aimed to reduce the bias due to variation in geolocation of users. No other filtering or normalization (e.g., case conversion or stemming) was applied on queries. The resulting set after the filtering had about 30 million queries. For every query in this final set, we also extracted the associated click data. Here, we limited ourselves to clicks issued on algorithmic search results as the only sign of engagement with the search engine result page (e.g., we have not included clicks on vertical search results or advertisements).

User data. The sampled queries were issued by about 6 million users, all located in the US. The number of males and females in the sample was quite similar, whereas there was some skew in the age distribution towards older users. However, each age group involved at least 100K users. The age and gender information were self-reported by the users in their account settings and were available in the query log.

Latency measurement. The latency values used in our analysis correspond to the time difference between the submission of the query by a user and the display of the search result page in the user's browser. This value is measured by a client-side JavaScript code running on the web browser and is communicated to a logging server in the search engine, letting us record the exact user-perceived latency value for every query. This end-to-end latency measurement includes the query processing time in the search engine, the network latency due to transfer of the query and results, and the time spent by rendering retrieved results in the user's browser. Since the rendering time is included in the measurement, we do not have a hidden bias arising from client's processing power, memory, or workload. Hence, we do not expect to have a hidden bias due to potentially uneven distribution of computational capacity by user demographics either.

4.2 Methodology and Metrics

Methodology. A naive method to quantify the effect of increasing latency on click behaviour would be to compute the click-through rate at different latency intervals and to show a negative correlation between the click-through rate and latency. Unfortunately, this method is not feasible because there are external factors that affect both the click-through rate and latency at the same time, most important being the quality of retrieved results [1]. For example, some search result pages may attract more clicks because they contain relevant results, and it may be the case that such pages are more likely to be cached and thus served with low response latency. The methodology we adopted in our analysis is specifically designed to eliminate this kind of potential bias due to the differences in search quality. Our methodology is based on measuring the impact of latency by observing the differences in click behaviour for pairs of "identical" query instances. More specifically, we compare the presence of clicks for two given query instances that are (i) submitted by the same user, (ii) having the same query string, and (iii) matching the same search results. That is, the same user submits the same query at two different times and is exposed to identical search results (but, potentially with different response latency values).[5] Therefore, the quality of presented results is identical, and the differences in click behaviour are more likely to stem from the differences in perceived response latency. We note that our methodology differs from the methodology adopted in [1] in that we also require the user to be identical. This way, our analysis becomes compatible with our controlled user study since the slower instance in a pair corresponds to the case where the user is exposed to an increased latency.

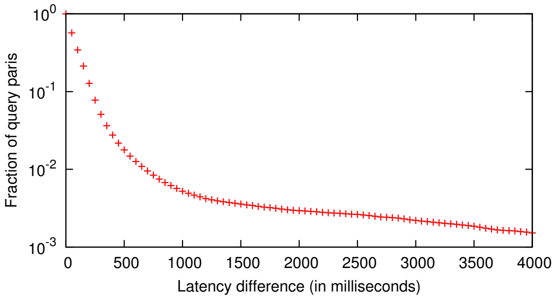

Figure 4: Complementary cumulative distribution of latency differences associated with query pairs.

Metrics. Let us assume a query pair $(q_{\text{fast}}, q_{\text{slow}})$, such that $\ell(q_{\text{fast}}) \leq \ell(q_{\text{slow}})$, where $\ell(q)$ denotes the response latency for query q. Our hypothesis is that, on average, users are more likely to engage with the search result page corresponding to q_{fast} than that of q_{slow} since the latter query is served with higher response latency. To verify this hypothesis, we place query instance pairs in coarse-grain buckets according to the latency difference $\ell(q_{\text{slow}}) - \ell(q_{\text{fast}})$ and, for each bucket, observe two different metrics, which indicate whether the users were engaged more with the results of the fast or slow query, using users' clicks on search results as a proxy for engagement:

Click presence: We compute the fraction of query instance pairs where at least one fast query result was clicked while no slow query result was clicked. We also compute the fraction of query instance pairs where at least one slow query result was clicked while no fast query result was clicked. We refer to these two fractions as *click-on-fast* and *click-on-slow*, respectively. For each latency difference bucket, we then compute the ratio between *click-on-fast* and *click-on-slow*, higher ratios (larger than 1) indicating a user preference towards faster query response.

Click count: We compute the fraction of query instance pairs in which the users have issued a larger number of clicks on the fast query results than the slow query results. We also compute the fraction of query instance pairs in which the users have issued a larger number of clicks on the slow query results than the fast query results. We refer to these two fractions as *click-more-on-fast* and *click-more-on-slow*, respectively.[6] As in the previous metric, we then compute the ratio between *click-more-on-fast* and *click-more-on-slow*, higher ratios (larger than 1) indicating a user preference towards faster query response.

4.3 Results

Fig. 4 shows the complementary cumulative distribution of latency differences across all query instance pairs. We observe that, for more than 99% of the pairs, the latency difference associated with the queries in the pair is less than

[5] Two search result pages are said to be identical if their top 10 results have the same URLs, ranked in the same order.

[6] In this metric, we consider only the query pairs where at least one result was clicked for each query in the pair.

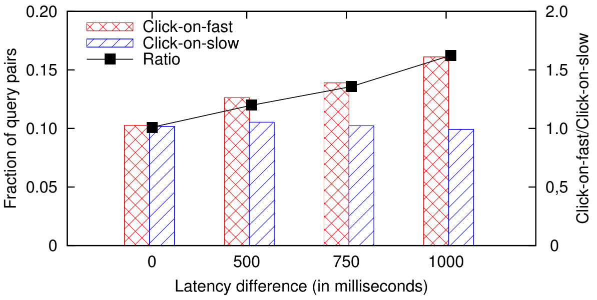

Figure 5: Fast or slow query response preference according to the click presence metric.

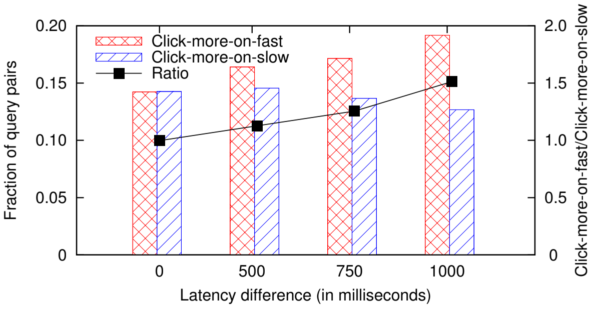

Figure 6: Fast or slow query response preference according to the click count metric.

1,000ms. This is due to two reasons. First, the queries in a pair were issued from a desktop computer and processed in the same search data centre. Hence, the associated network overheads are similar. Second, the submitted query string and retrieved search results were identical for both query instances. Hence, the response times of the search engine are similar, and the differences in latency potentially stem from the differences in the rendering of results in user's web browser. In any case, the range of latency differences we explore is consistent with the latency increments used in the controlled user study presented in Section 3.

In Fig. 5, we show the fraction of query pairs falling under the *click-on-fast* and *click-on-slow* sets, as well as the ratio between the two fractions. Each latency bucket contains the query pairs in which the latency difference between the queries is not less than a certain threshold latency value (the values shown on the x axis: 0ms, 500ms, 750ms, and 1,000ms). According to this figure, when the latency difference increases, the *click-on-fast* set grows quickly while the *click-on-slow* set remains relatively stable. The ratio between the *click-on-fast* and *click-on-slow* sets is always larger than 1, and it keeps increasing as the latency difference increases. To test the statistical significance of the results reported in Fig. 5, we apply the McNemar test to determine whether the marginal frequencies of the binary outcomes are equal. More specifically, we compute a 2 × 2 contingency table for each latency difference shown in Fig. 5, using the raw counts of query pairs that fall in each possible outcome (*click-on-fast*, *click-on-slow*, *click-on-both*, *click-on-neither*). As expected, the McNemar test does not reveal a statistically significant difference for the 0ms latency difference (the control group). However, the test indicates a significant difference in the proportion of query pairs allocated in the possible outcomes for the 500ms ($\chi^2(1) = 22.41, p < .001$), 750ms ($\chi^2(1) = 31.13, p < .001$), and 1,000ms ($\chi^2(1) = 51.40, p < .001$) latency differences. Hence, we reject the null hypothesis of marginal homogeneity, i.e., the marginal probabilities for each outcome are the same. In other words, given a pair of identical queries and search result pages, users show a clear preference to engage with the search results that are served with lower latency.

Fig. 6 shows the fraction of query pairs falling under the *click-more-on-fast* and *click-more-on-slow* sets, as well as the ratio between the two fractions. We observe that, as the latency difference increases, the size of the *click-more-on-fast* set increases while the size of the *click-more-on-slow* set decreases. As before, the ratio between the size of the *click-more-on-fast* and *click-more-on-slow* sets is always larger than 1 and it keeps increasing as the latency difference increases. We applied the McNemar test to measure the statistical significance of the results reported in Fig. 6. In this case, the outcomes of the 2 × 2 contingency table are *click-more-on-fast*, *click-more-on-slow*, *click-same-on-both-at-least-one-click*, and *click-same-on-both-zero-click*. As in the previous experiment, we do not observe any statistically significant difference for the 0ms latency difference (the control group). When applying the McNemar test to the remaining latency differences, we note a significant difference in the proportion of users allocated in the possible outcomes for the 500ms ($\chi^2(1) = 13.09, p < .001$), 750ms ($\chi^2(1) = 22.06, p < .001$), and 1,000ms ($\chi^2(1) = 46.37, p < .001$) latency differences. This leads us to reject the null hypothesis of marginal homogeneity in favour of the alternative hypothesis, i.e., the marginal proportions are significantly different from each other. Hence, this experiment shows that the users tend to click on a larger number of result links if they are served with lower latency.

5. CONCLUDING REMARKS

The results of our controlled study revealed that, as the response latency of the search engine reaches higher values, the arousal and the negative valence of the experienced emotions increase as well. Although those effects did not produce changes on the self-reported data, their impact on users' physiological responses was evident. Thus, even if such short latency increases of under 500ms are not consciously perceived, they have sizeable physiological effects that can contribute to the overall user experience. This highlights the need for a more inter-disciplinary approach to the evaluation of human information processing in HCI research, as there are plenty of relevant effects that we might be missing if we rely solely on self-report measures. The fact that latency conditions tested here presented effects with little conscious awareness cannot be taken as a proof that they do not affect user's online behaviour, preferences, or choices. Research in psychology has indicated that our motivations and preferences are not always determined by conscious objectives or reasons. Moreover, it leads to the question of what is the actual effect that such delays might have on the engagement of users. With a large-scale query log analysis, we ascertained the effect on the clicking behaviour of users. As a result of this analysis, we revealed a significant decrease in users' engagement with the search result page, even at small increases in latency.

We believe that the users show high variation in the way they perceive the response latency of a web service as the effect depends on demographics, context, and potentially many other factors. This subjective nature may create an opportunity for search engines. Search results can be served to each user at custom latencies depending on the estimated behavioural impact of latency on the user. For example, if no degradation is expected in user experience, the priority of the user query may be reduced or the search results may be computed using less hardware resources, eventually serving the query with a higher latency. Serving results at higher latencies may bring financial benefits to search engines in the form of decreased hardware investments and reduced energy consumption, also saving time and resources for time-critical queries. To this end, we need to devise mechanisms for accurate prediction of user-perceived response latency as well as the impact of latency on user experience. We can then come up with proper models for personalising response latency on a per-user basis, eventually aiming to achieve financial cost savings for the search engine company without hurting users' engagement and satisfaction.

6. REFERENCES

[1] I. Arapakis, X. Bai, and B. B. Cambazoglu. Impact of response latency on user behavior in web search. In *Proc. 37th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 103–112, 2014.

[2] L. Azzopardi. Modelling interaction with economic models of search. In *Proc. 37th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 3–12, 2014.

[3] L. Azzopardi, D. Kelly, and K. Brennan. How query cost affects search behavior. In *Proc. 36th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 23–32, 2013.

[4] C. Bandt and B. Pompe. Permutation entropy: A natural complexity measure for time series. *Physical Review Letters*, 88, 2002.

[5] J. A. Bargh and T. L. Chartrand. The unbearable automaticity of being. *American Psychologist*, 54(7):462–479, 1999.

[6] P. Bliese. Multilevel modeling in R (2.5): A brief introduction to R, the multilevel package and the NMLE package. `http://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf`, 2013.

[7] M. Bradley and P. J. Lang. *Emotion and motivation*, pages 581–607. Handbook of psychophysiology. New York: Cambridge University Press, 2007.

[8] J. D. Brutlag, H. Hutchinson, and M. Stone. User preference and search engine latency. In *Proc. ASA Joint Statistical Meetings*, 2008.

[9] E. S. Davis and D. A. Hantula. The effects of download delay on performance and end-user satisfaction in an Internet tutorial. *Computers in Human Behavior*, 17(3):249–268, 2001.

[10] M. E. Dawson, A. M. Schell, and D. L. Fillion. *The electrodermal system*, pages 159–181. Handbook of psychophysiology. New York: Cambridge University Press, 2007.

[11] B. G. C. Dellaert and B. E. Kahn. How tolerable is delay: Consumers' evaluations of Internet web sites after waiting. *Journal of Interactive Marketing*, 13(1):41–54, 1999.

[12] A. R. Dennis and N. J. Taylor. Information foraging on the Web: The effects of "acceptable" Internet delays on multi-page information search behavior. *Decision Support Systems*, 42(2):810–824, 2006.

[13] D. F. Galletta, R. Henry, S. McCoy, and P. Polak. Web site delays: How tolerant are users? *Journal of the Association for Information Systems*, 5:1–28, 2003.

[14] J. Gwizdka and I. Lopatovska. The role of subjective factors in the information search process. *Journal of the American Society for Information Science and Technology*, 60(12):2452–2464, 2009.

[15] M. Hassenzahl. Funology. In M. A. Blythe, K. Overbeeke, A. F. Monk, and P. C. Wright, editors, *The Thing and I: Understanding the Relationship Between User and Product*, pages 31–42. Kluwer Academic Publishers, 2004.

[16] S. D. Kristjansson, J. C. Kircher, and A. K. Webb. Multilevel models for repeated measures research designs in psychophysiology: An introduction to growth curve modeling. *Psychophysiology*, 44(5):728–736, 2007.

[17] J. R. Lewis. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, 1995.

[18] D. Maxwell and L. Azzopardi. Stuck in traffic: How temporal delays affect search behaviour. In *Proc. 5th Information Interaction in Context Symp.*, pages 155–164, 2014.

[19] L. McCay-Peet, M. Lalmas, and V. Navalpakkam. On saliency, affect and focused attention. In *ACM SIGCHI Conf. Human Factors in Computing Systems*, pages 541–550, 2012.

[20] F. F.-H. Nah. A study on tolerable waiting time: How long are web users willing to wait? *Behaviour and Information Technology*, 23(3):153–163, 2004.

[21] R. E. Nisbett and T. D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, 1977.

[22] H. L. O'Brien and E. G. Toms. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69, 2010.

[23] W. Prinz. Why don't we perceive our brain states? *European Journal of Cognitive Psychology*, 4(1):1–20, 1992.

[24] J. Ramsay, A. Barbesi, and J. Preece. A psychological investigation of long retrieval times on the World Wide Web. *Interacting with Computers*, 10(1):77–86, 1998.

[25] N. Ravaja. Contributions of psychophysiology to media research: Review and recommendations. *Media Psychology*, 6(2):193–235, 2004.

[26] E. Schurman and J. Brutlag. Performance related changes and their user impact. In *Velocity – Web Performance and Operations Conf.*, 2009.

[27] C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[28] M. D. Smucker. Towards timed predictions of human performance for interactive information retrieval evaluation. In *Proc. Symp. Human-Computer Information Retrieval*, 2009.

[29] L. Tassinary, J. T. Cacioppo, and E. J. Vanman. *The skeletomotor system: Surface electromyography*, pages 267–299. Handbook of psychophysiology. New York: Cambridge University Press, 2007.

[30] N. J. Taylor, A. R. Dennis, and J. W. Cummings. Situation normality and the shape of search: The effects of time delays and information presentation on search behavior. *Journal of the American Society for Information Science and Technology*, 64(5):909–928, 2013.

[31] J. Teevan, K. Collins-Thompson, R. W. White, S. T. Dumais, and Y. Kim. Slow search: Information retrieval without time constraints. In *Proc. Symp. Human-Computer Interaction and Information Retrieval*, pages 1:1–1:10, 2013.

[32] E. R. Thompson. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of Cross-Cultural Psychology*, 38(2):227–242, 2007.

[33] D. Watson, L. A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070, 1988.