Engaged or Frustrated? Disambiguating Emotional State in Search

Ashlee Edwards
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
ashlee.a.edwards@gmail.com

Diane Kelly
University of Tennessee
Knoxville, TN, USA
dianek@utk.edu

ABSTRACT

One of the primary ways researchers have characterized engagement during information search is by increases in search behaviors, such as queries and clicks. However, studies have shown that frustration is also characterized by increases in these same behaviors. This research examines the differences in the search behaviors and physiologies of people who are engaged or frustrated during search. A 2x2 within-subject laboratory experiment was conducted with 40 participants. Engagement was induced by manipulating task interest and frustration was induced by manipulating the quality of the search results. Participants' interactions and physiological responses were recorded, and after they searched, they evaluated their levels of engagement, frustration and stress. Participants reported significantly greater levels of engagement when completing tasks that interested them and significantly less engagement during searches with poor results quality. For all search behaviors measured, only two significant differences were found according to task interest: participants had more scrolls and longer query intervals when searching for interesting tasks, suggesting greater interaction with content. Significant differences were found for nine behaviors according to results quality, including queries issued, number of SERPs displayed and number of SERP clicks, suggesting these are potentially better indicators of frustration rather than engagement. When presented with poor quality results, participants had significantly higher heart rates than when presented with normal quality results. Finally, participants had lower heart rates and greater skin conductance responses when conducting interesting tasks than when conducting uninteresting tasks. This research provides insight into the differences in search behaviors and physiologies of participants when they are engaged versus frustrated and presents techniques that can be used by those wishing to induce engagement and frustration during laboratory IIR studies.

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

ACM ISBN 978-1-4503-5022-8/17/08…$15.00
http://dx.doi.org/10.1145/3077136.3080818

CCS CONCEPTS

• **Information systems** → **Information retrieval**; *Users and interactive retrieval*

KEYWORDS

Interactive information retrieval, search behavior, emotion, engagement, frustration, physiological signals

ACM Reference format:

Ashlee Edwards and Diane Kelly. 2017. Engaged or Frustrated? Disambiguating Emotional State in Search. In *Proceedings of ACM SIGIR conference, Tokyo, Japan, August 2017 (SIGIR'17)*, 10 pages.

DOI: http://dx.doi.org/10.1145/3077136.3080818

1 INTRODUCTION

The study of engagement in interactive information retrieval is one area where the integration of subjective experience and interaction measures is needed. To date, search engagement has been overwhelmingly characterized by a focus on behavioral signals as proxies for engagement [10, 19, 20]. These measures have primarily been frequency-focused: frequency of clicks and queries, as well as greater frequency in overall activity per session. While behavioral signals can be useful, other work has contradicted this frequency-based notion of engagement. For example, O'Brien and Lebow [24] investigated the relationship between engagement and search behavior and found participants who rated their engagement highest had the lowest reading times overall, lowest browsing times, and lowest total session times. Participants who rated their engagement highest also visited the fewest pages and used the least recommended links, suggesting that participants who were the most engaged had the least amount of search interaction [24]. This calls into question work that has suggested that engagement is signaled by an increase in interaction measures. One possibility is that instead of experiencing increased engagement, people who click and query frequently are actually frustrated. Feild, Allan and Jones [13], for example, found that frustrated participants also exhibit periods of high search activity.

Opportunities for disambiguating engagement and frustration during search may lie in linking behavioral signals with other measures, specifically physiological measures. It is well known that frustration generally produces higher stress responses [29]. However, there is conflicting evidence about how engagement manifests itself physiologically. Work in psychology has found that as engagement increases, stress hormones also linearly

increase [23]. However, O'Brien and Lebow [24] found negative correlations between electrodermal activity, heart rate, and self-report engagement. In this work, we address the following research questions: 1) To what extent do search behaviors differ when people are engaged or frustrated? (2) To what extent do physiological signals help disambiguate engaging and frustrating search episodes?

2 PREVIOUS WORK

Historically, IR research has emphasized the functional aspects of search systems such as performance and usability; however, in recent years, there has been growing interest in creating engaging search experiences for searchers. O'Brien and Toms [25] define engagement as a "category of user experience characterized by attributes of challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control" (p.7).

Researchers in IR have primarily examined engagement by focusing on signals that can be extracted from search logs such as clicks and dwell time. Generally, the underlying assumption is greater frequency of certain behaviors, such as clicks and dwell times, indicates more engaged users. Lehmann and colleagues have completed an extensive body of research on engagement within the context of large-scale search logs [19, 20, 21]. Lehmann et al. [21] proposed and evaluated three interaction-based models of engagement: a general model, a time-based model and a user-based model. Their general model of engagement focused primarily on popularity and clicks on a site, while the time-based model was more focused on loyalty, and the user-based model was more focused on an individual user's behavioral patterns. Lehmann et al. [19] continued this work by proposing the concept of *networked user engagement*, referring to engagement within a network of websites and later focused on user engagement during multitasking episodes using dwell time and page views [20]. These basic signals were transformed into metrics like attention shift, attention range, and cumulative actions. Dupret and Lalmas [10] investigated the usefulness of absence time, or the time between two user visits, as a metric of user engagement. This was based on the assumption that users who are engaged will return to a site sooner, meaning that their absence times will be shorter. Ortiz-Cordova and Jansen [27] sought to identify those who were more engaged with site content and advertisements, and defined engagement in terms of new visits, number of pages visited during the duration of the session, time spent on a site, click-through rate, ads clicked, ad impressions, and rate of return to the site.

On the other hand, there is evidence that searchers exhibit more search behaviors when they are frustrated or experiencing difficulties. For example, Feild, et al. [13] found that frustrated participants engaged in increased search behavior. Aula et al. [5] gave participants easy and difficult tasks and found that participants attempted longer queries and had more query reformulations when completing difficult tasks. Hassan, White, Dumais and Wang [15] classified search behaviors as either struggling or exploring, and found participants who were struggling tended to submit less diverse queries than participants who were exploring. They also found that dwell time was lower for participants who were struggling. Also in contrast to studies that have equated increased behavior with increased engagement is the work of O'Brien and Lebow [24] who found participants with the highest levels of engagement had the lowest display times overall, lowest browsing times, and lowest total session times. Participants who rated their engagement highest visited the fewest pages and used the least recommended links. Thus, the challenge is that increases (or decreases) in behavioral signals can be (and have been) used to indicate frustration, task difficulty, task interest, relevance, satisfaction and so forth. Though behavioral signals are useful in that they can potentially serve as indicators of what the user is thinking and feeling, using them alone to define engagement is problematic because what they indicate is often ambiguous.

One potential way to better determine if a person is engaged or frustrated during search is to ask them. Several of the studies described above, including Feild, et al. [13], Aula et al. [5] and O'Brien and Lebow [24] collected self-report data to better understand the user's experiences. Another potential way is to use physiological measures. While the use of physiological measures is not widespread in IR, several recent papers have demonstrated their usefulness [22, 12, 6, 4] and a recent workshop at the *2015 SIGIR Conference* called for more work integrating these measures with other standard IR measures[1]. Physiological measures include things such as skin conductance, heart rate, skin resistance, muscle tension and neuronal activity. One of the underlying motives for using such measures is they provide information about a person's emotional experiences, including those that might be subconscious (and thus, unlikely to be reported by the person) [6].

In IR, physiological measures have been combined with other measures such as search behaviors and self-report measures to better understand the search process. Moshfeghi and Jose [22] combined facial expression, heart rate, galvanic skin response, and EEG signals with dwell time to predict relevance. Arapakis, Konstas, and Jose [4] combined galvanic skin response, heart rate, temperature, and facial expression analysis with machine learning to predict which documents or video snippets would be relevant. Arapakis, Jose, and Gray [2] linked facial expressions with subjective reports of emotions experienced and discovered that irritation and anxiety were linked to perceptions of difficulty and feelings of fatigue experienced during a task. Barreda-Angeles et al. [6] used physiological signals to examine the unconscious effects of response latency on participants, and found indications of physiological effects even at small increases in latency.

Physiological measures have also been used to better understand frustration [29], engagement with content [22, 4] and reactions to search interfaces [12]. Scheirer, Fernandez, Klein, and Picard [29] discriminated between frustrated and non-frustrated states by linking increases in galvanic skin response

[1] https://sites.google.com/site/neuroir2015/home

(GSR) and blood volume pressure (BVP) to frustration as a result of delayed mouse clicks. O'Brien and Lebow [24] found that in an online news context, participants who were more engaged had lower levels of electrodermal activity and electromyography (facial movement). Arapakis, Lalmas and Valkanas [3] investigated the usefulness of mouse gestures and gaze behavior as possible indicators of engagement with search results and found a correlation between gaze behavior and engagement, specifically, participants had more fixations when reading an interesting article and spent more time with the content.

The studies cited above focused on engagement or frustration, but did not attempt to distinguish between these two emotions. We found one study that attempted to use physiological measures to distinguish between these two states [14]. Grafsgaard, et al. [14] used facial expression and skin conductance to differentiate between engaging and frustrating episodes in the context of online tutoring. Students were given a modified version of the *endurability* sub-scale of O'Brien and Toms' [26] User Engagement Scale (UES) and also completed portions of the NASA-TLX. Grafsgaard, et al. found that certain facial expressions could be used to distinguish between engagement and frustration. For example, endurability was predicted by more frequent inner brow raising and frustration by more frequent brow lowering. While this study provided interesting results regarding facial expression, it did not examine some of the other physiological measures reported in previous IR studies, which complicates cross-study comparisons.

Our review of the literature shows that while a number of studies have used increases in search behaviors such as clicking, querying and time spent as indicators of engagement, other work has found these same measures are associated with constructs such as frustration, and at least one study found that fewer (not more) search behaviors were associated with more engaging experiences. While there has been some work integrating physiological measures into the study of engagement and frustration, to our knowledge there has been no study which has examined how the combination of physiological signals, search behaviors and self-report measures can potentially be used to distinguish between engaging and frustrating search experiences. Moreover, none of the studies above examined frustration and engagement during a single study and within the same participant. In this work, we attempt to see if, for the same participant, we can distinguish between engaging and frustrating searches with respect to search behaviors and physiological responses.

3 METHOD

A laboratory experiment was conducted, with two within-subject experimental conditions, each with two levels. These conditions are described in more detail below. Unless otherwise reported, two-way repeated measures ANOVAs and Bonferroni's post-hoc tests were used for analyses. A total of 40 undergraduate participants were recruited through a university email list. Participants' average age was 20.45 (*SD*=1.77), and most were female (n=31). Participants came from a variety of majors: the natural sciences (n=13), humanities (n=11), social sciences (n=8) and professional schools (n=7). Most participants reported having between 7-9 years of search experience, and reported conducting online searches for information more than 7 times a day. This study was approved by our university IRB (#15-0956).

3.1 Search Tasks

The tasks used in this study were based on tasks created and evaluated by Kelly, et al. [17]. Specifically, we modeled our tasks after those labeled as *evaluative* because participants in this previous study rated these types of tasks as engaging and spent about ten minutes completing them. The *evaluate* task type requires participants to search for information in order to evaluate several options, make a selection based on this information, and then justify the selection (Figure 1). A total of eight tasks were used (see http://bit.ly/2r6w5w2). Tasks were from four domains (health, science, technology and entertainment). Participants were asked to bookmark pages that helped them address the tasks. Participants were not given any task time limits.

> You recently heard a story on National Public Radio about the use of biomass as fuel. Biomass refers to material created from living organisms. What are different types of biomasses that are used as fuels and how are they created? How do biomass fuels compare with fossil fuels when it comes to environmental impact? Which do you think is better? Why?

Figure 1. Example search task in the science domain

3.2 Experimental Conditions

Our study rests on people being engaged while they search. Manipulating engagement during a laboratory study is not an easy task and there is little guidance in the IIR literature about how to do this. O'Brien and Toms [25], as well as researchers in other fields [29, 14] have shown that interest is a key component of engagement. Thus, in this study we attempt to induce engagement through *task interest*. To identify tasks that did and did not interest participants from the set of eight we created, we asked them to rank the tasks according to their interest the day before they completed the experiment using an online form. The presentation order of the topics was randomized to avoid selection bias. During the experiment, participants were given four tasks to complete: the two tasks they ranked as most interesting (ranks 1 and 2) and the two tasks they ranked as least interesting (ranks 7 and 8). Participants' rankings of tasks can be found online at: http://bit.ly/2r6w5w2. Thus, in this study, task interest has two levels: *interesting* and *uninteresting*.

To induce frustration, we manipulated search engine results page (SERP) quality. This was accomplished by modifying the source code of the search system so that results were presented beginning at rank 500. During pilot testing, we tried other starting points (100, 200), but found we needed to start at 500 to ensure consistently poor result quality. Thus, in this study, *SERP quality* has two levels: *normal* and *bad*. In total, participants

completed four search tasks: (1) interesting/normal SERP; (2) interesting/bad SERP; (3) uninteresting/normal SERP and (4) uninteresting/bad SERP. These combinations were counterbalanced.

3.3 Search System & Cover Story

We used a search system developed by members of our lab, which used the Bing API. The interface for this system was similar to a standard Web search interface although there were no advertisements. Participants were presented with a query box and a list of search results with snippets (10 per page). The search task was also printed at the top of the page just above the search box. Participants could perform normal activities such as querying, clicking and bookmarking pages. A cover story was used to keep participants from becoming suspicious of the poor results. Participants were told they were evaluating four different search systems. Participants were debriefed at the end of the experiment regarding the deception and manipulation.

3.4 Search Behaviors

Search behaviors were logged and we extracted queries, clicks, and time. These behaviors were chosen because they have been used in other work to examine engagement [24], to determine whether a user is struggling or exploring [15] and to study frustration during search [13].

3.5 Self-Report Measures

3.5.1 Task Evaluations: Interest and Difficulty

Participants completed pre-search questionnaires, which asked them about their prior knowledge and experience searching for the task topic, task relevance, their interest in the task (manipulation check) and how difficult they expect it will be to search. Participants completed post-task questionnaires after each task, which asked them about experienced difficulty and success, and their perceptions of their own skill, and the ability of the system to retrieve documents. All of these items were evaluated with 5-point scales and coded so that higher values indicated more of the construct being evaluated (e.g., more difficult, more success).

3.5.2 Engagement

Engagement was measured using the User Engagement Scale (UES) [26], and defined earlier in this paper. The scale used in [26], which is a specialized version of the UES designed for online search situations (namely Wikipedia), includes several subscales: aesthetics (5 items), perceived usability (8 items), felt involvement (3 items), novelty (2 items), endurability (5 items), and focused attention (5 items), for a total of 28 items. The aesthetics subscale was not used in this study because the interface was constant throughout. We also changed the term "Wikipedia," which was evaluated in the original study, to "search system." Finally, we dropped one item from the felt involvement subscale because it did not make sense in the context of our study. Thus, in total we had 22 items. All of these items were evaluated with a 5-point Likert scale, where 1=strongly disagree and 5=strongly agree. Responses were coded so that higher values indicated more of the construct being measured. Each subscale score represents an average of the items along each dimension.

3.5.3 Frustration & Stress

Frustration was defined in this study using Amsel's [1] definition of primary frustration, or "a temporary state that results when a response is nonreinforced in the presence of reward expectancy" (p. 2). Peters, O'Connor and Rudolf's [28] three-item scale was used to measure frustration. The scale contains 5-points, where 1=not at all and 5=extremely (higher values equal more frustration). An example item from this scale is "Trying to complete this task was a frustrating experience." This measure served to verify that our manipulation of frustration (i.e., SERP quality) worked.

In addition to engagement and frustration, we also measured self-reported stress as a way to better understand physiological measures. Stress was defined as "a particular relationship between the person and environment that is appraised by the person as taxing or exceeding his or her resources" [18] (p. 40). Stress was measured using the 24-item Short Stress State Questionnaire (SSSQ) [16]. The SSSQ has three factors: distress, worry, and engagement. Since we already had a scale in place to measure engagement, these items were removed from the questionnaire. *Distress* is a measure of negative affect, and contains items such as "I feel depressed" or "I feel irritated." *Worry* is a measure of self-perception, and contains items such as "I feel concerned about the impression I am making," and "I thought about how I would feel if I were told how I performed." The scale contains 5-points, where 1=not at all and 5=extremely (higher values equal greater stress). Items were changed to reflect past tense and "this task" was changed to "this search task."

3.6 Physiological Measures

Skin conductance and heart rate were measured using BioPac. Previous work has shown that these are useful signals of stress [7]. To measure skin conductance, electrodes were attached to the thenar and hypothenar eminences of each participants' non-dominant palm. To measure heart rate, electrodes were attached to the collarbone and ribs of each participant. Braithwaite, et al. [8] offer recommendations for using the BioPac and accompanying software to perform electrodermal measurement, which we followed. After the electrodes were attached to participants, a three-minute resting period was observed for each participant before he/she was instructed to start the experiment.

Data was gathered at a rate of 1sample/msec. The electrodermal activity signal was recorded with a high band pass filter of 0Hz, and a low band pass filter of 35Hz, with a gain of 1000Hz. Skin conductance values are reported in microsiemens (μs). The electrocardiography signal was recorded with a high band pass filter set at 0.5 and a low band pass filter set at 35Hz, with a gain of 1000Hz. Heart rate was extracted from the raw ECG signal and is reported in beats per minute (bpm). Both the skin conductance and heart rate data were visually inspected for

any errors or potential aberrant artifacts. A Shannon entropy analysis was conducted to look for aberrations or potential artifacts in the data, as was done in other work [4]. Both signals were examined for entropy values across conditions, and the differences were non-significant. The average entropy for skin conductance signals was 3.24, and the average entropy for heart rate data was 3.64. These values indicate very little aberrant fluctuation in either measure.

4 RESULTS

4.1 Self-Report Measures

We presented preliminary results regarding task interest in an earlier short paper [11]. This paper did not report results of the SERP quality manipulation and was instead focused on presenting the method we used to identify and assign search tasks that interested participants, since this is an important methodological concern for IIR researchers. Results show the interest manipulation was successful (Table 1): participants were significantly more interested in tasks they ranked as more interesting, reported significantly higher prior knowledge, relevance and search frequency, and expected these tasks to be significantly less difficult than tasks ranked as less interesting.

Table 1. Means, standard deviations, and paired-sample t-test results of pre-task evaluations, $^{***}p<.001$

	Interesting Tasks	Uninteresting Tasks	*t*-values *df*=79	Cohen's *d*
Interest	4.01 (0.96)	2.77 (1.15)	7.39***	1.17
Prior Knowledge	3.00 (1.13)	2.20 (1.80)	5.27***	0.53
Relevance	3.61 (1.23)	2.24 (1.23)	7.06***	1.11
Search Frequency	2.44 (1.20)	1.64 (0.94)	4.69***	0.74
Difficulty	3.35 (1.15)	3.81 (0.93)	2.72***	0.44

Tables 2 and 3 show participants' post-task evaluations. Although repeated measures ANOVAs were conducted, there were no significant interaction effects, so we present data in separate tables to simplify the presentation. There were significant main effects for SERP quality for all items lending support that our frustration manipulation was successful. When SERP quality was bad, participants rated tasks as significantly more difficult, rated their skills lower, rated system performance lower and reported less success. There were no significant main effects for task interest and no significant interaction effects.

Table 2. Means, standard deviations, and F-test results of post-task evaluations according to task interest

	Interesting Tasks	Uninteresting Tasks	*F*-values *df*=1, 39	η^2
Difficulty	3.61 (0.83)	3.38 (0.89)	1.76	0.04
Skill	3.85 (0.70)	3.62 (0.85)	2.47	0.06
System	3.49 (0.86)	3.35 (1.04)	0.07	<0.01
Success	3.77 (0.69)	3.80 (0.77)	0.03	<0.01

Table 3. Means, standard deviations, and F-test results of post-task evaluations according to SERP quality, $^{***}p<.001$

	Bad SERP Quality	Normal SERP Quality	*F*-values *df*=1, 39	η^2
Difficulty	3.33 (1.26)	1.66 (0.91)	56.27***	0.59
Skill	3.29 (1.02)	4.19 (0.84)	23.57***	0.37
System	2.49 (1.50)	4.26 (1.19)	43.31***	0.53
Success	3.14 (1.28)	4.44 (0.90)	32.99***	0.46

Participants reported significantly greater felt involvement, focused attention, and novelty when completing interesting tasks (Table 4). When SERP quality was bad, they reported significantly lower perceived usability, felt involvement, and endurability (Table 5). There were no interaction effects.

Table 4. Means, standard deviations, and F-test results of engagement according to task interest, $^{*}p<0.05$; $^{**}p<0.01$

	Interesting Tasks	Uninteresting Tasks	*F*-values *df*=1, 39	η^2
Perceived Usability	3.63 (0.60)	3.56 (0.79)	0.24	<0.01
Focused Attention	2.87 (0.79)	2.64 (0.66)	7.21**	0.16
Felt Involvement	3.61 (0.50)	3.27 (0.72)	5.35*	0.12
Endurability	3.41 (0.58)	3.20 (0.76)	3.41	0.08
Novelty	3.72 (0.67)	3.09 (0.90)	19.66**	0.34

Table 5. Means, standard deviations, and F-test results of engagement according to SERP quality, $^{**}p<0.01$; $^{***}p<0.001$

	Bad SERP Quality	Normal SERP Quality	*F*-values *df*=1, 39	η^2
Perceived Usability	2.93 (1.05)	4.25 (0.82)	56.17***	0.59
Focused Attention	2.73 (0.82)	2.78 (0.85)	0.22	<0.01
Felt Involvement	3.16 (1.00)	3.72 (0.72)	19.87***	0.34
Endurability	2.83 (1.02)	3.78 (0.79)	36.26***	0.48
Novelty	3.15 (1.20)	3.66 (0.97)	8.65**	0.18

Tables 6 and 7 show the results of the frustration and stress scales according to task interest and SERP quality, respectively. There were no differences in participants' reported frustration or stress according to task interest. Participants reported experiencing significantly more frustration when completing tasks where the SERP quality was bad, suggesting a successful manipulation of frustration. Participants also reported experiencing significantly more distress and overall stress when completing tasks where SERP quality was bad. There were no interaction effects.

Table 6. Means, standard deviations, and F-test results of frustration and stress according to task interest

	Interesting Tasks	Uninteresting Tasks	*F*-values *df*=1, 39	η^2
Frustration	2.42 (1.04)	2.52 (0.97)	0.02	0.01
Stress: Distress	1.56 (0.62)	1.60 (0.73)	0.26	0.01
Stress: Worry	1.52 (0.51)	1.50 (0.54)	0.16	<0.01
Stress: Overall	1.55 (0.45)	1.56 (0.53)	0.03	<0.01

Table 7. Means, standard deviations, and F-test results of frustration and stress according to SERP quality, $^{***}p<.001$

	Bad SERP Quality	Normal SERP Quality	*F*-values *df*=1, 39	η^2
Frustration	3.18 (1.10)	1.75 (0.94)	70.52***	0.97
Stress: Distress	1.58 (0.67)	1.27 (0.47)	34.54***	0.47
Stress: Worry	1.56 (0.56)	1.46 (0.48)	2.35	0.57
Stress: Overall	1.75 (0.50)	1.35 (0.39)	30.18***	0.44

4.2 Search Behaviors

There was no difference in the number of queries submitted by participants for interesting tasks, or in the length of these queries, or number of clicks made per query (Table 8). Participants displayed similar numbers of SERPs, clicked on a similar number of results and saved a similar number of results when completing interesting and uninteresting tasks. There were significantly more scrolls made to SERPs for interesting than uninteresting tasks.

Participants submitted significantly more queries when completing tasks when the SERP quality was bad (Table 9). These queries were significantly shorter and participants made significantly more clicks per query. When the quality was bad, participants displayed significantly more SERPs, clicked more SERPs, made more scrolls and saved fewer results.

Table 8. Means, standard deviations, and F-test results of search behaviors according to task interest, $^{*}p<.05$

	Interesting Tasks	Uninteresting Tasks	*F*-values *df*=1, 39	η^2
Queries Issued	5.65 (4.45)	5.06 (4.13)	0.96	0.02
Query Length	3.50 (1.67)	3.65 (1.79)	0.73	0.02
Clicks per Query	7.77 (4.77)	6.79 (3.97)	2.09	0.05
SERPs Displayed	13.80 (9.16)	12.29 (7.88)	1.83	0.04
SERP Clicks	19.52 (14.30)	17.70 (13.75)	1.39	0.03
SERP Scrolls	31.78 (27.59)	23.02 (18.94)	5.91*	0.13
Pages Saved	3.60 (2.06)	3.42 (1.83)	0.41	0.01

Table 9. Means, standard deviations, and F-test results of search behaviors according to SERP quality, $^{**}p<.01$, $^{***}p<.001$

	Bad SERP Quality	Normal SERP Quality	*F*-values *df*=1, 39	η^2
Queries Issued	7.51 (4.66)	3.20 (4.42)	57.18***	0.60
Query Length	3.53 (1.83)	3.62 (1.64)	52.07***	0.57
Clicks per Query	8.27 (5.03)	6.29 (3.42)	11.18***	0.22
SERPs Displayed	15.52 (9.87)	8.56 (5.90)	39.72***	0.50
SERP Clicks	23.04 (15.57)	14.19 (10.66)	21.15**	0.35
SERP Scrolls	41.15 (32.01)	13.66 (14.52)	54.45***	0.58
Pages Saved	3.14 (1.40)	3.88 (2.31)	7.53**	0.16

Participants spent similar amounts of time completing interesting and uninteresting tasks (Table 10). They also spent similar amounts of time per SERP and per document. When completing interesting tasks, participants had significantly greater query intervals, or time between subsequent queries. There were significant main effects for SERP quality for overall time spent and time per SERP (more time for bad SERP quality) (Table 11).

Table 10. Means, standard deviations, and F-test results of time according to task interest, $^{***}p<.001$

	Interesting Tasks	Uninteresting Tasks	*F*-values *df*=1, 39	η^2
Time per Task	6.86 (3.84)	5.88 (3.06)	3.66	0.09
Time per SERP	2.04 (1.80)	1.74 (1.30)	2.00	0.05
Time per Doc	2.72 (2.25)	2.33 (1.79)	2.07	0.06
Query Interval	5.02 (3.70)	2.96 (3.57)	19.30***	0.33

Table 11. Means, standard deviations, and F-test results of time according to SERP quality, $^{*}p<.01$, $^{**}p<.001$

	Bad SERP Quality	Normal SERP Quality	*F*-values *df*=1, 39	η^2
Time per Task	7.11 (3.66)	5.63 (3.17)	7.26*	0.16
Time per SERP	2.71 (1.71)	1.08 (0.84)	56.13**	0.62
Time per Doc	2.66 (2.16)	2.40 (1.91)	0.63	0.02
Query Interval	4.36 (3.95)	3.62 (3.53)	2.91	0.07

4.3 Physiological Measures

Physiological data from 39 participants is reported because of a logging failure. The overall skin conductance values were similar for both task interest and SERP quality (Tables 12 and 13, rows 1 and 2). There were no significant main or interaction effects.

A window analysis was performed by splitting the data along ten-second intervals, and averaging those windows for each participant for each task in an effort to better capture any fluctuations that might happen over time (Tables 12 and 13, rows 3 and 4). This resulted in 42 ten-second-time windows for each task. These windows impose an artificial seven-minute length for each participant, although task times varied; this was done to

compare participants for a given time period. There were no significant main effects for skin conductance according to task interest or SERP quality, but there was a significant interaction effect. Post-hoc tests showed participants had the highest skin conductance when conducting interesting tasks with normal SERP quality, the second highest when conducting uninteresting tasks with bad SERP quality. The lowest skin conductance was observed when participants completed uninteresting tasks with normal SERP quality. Participants had significantly higher heart rates for uninteresting tasks and for tasks where bad quality SERPs were presented. There were also significant interaction effects: participants completing interesting tasks with normal SERP quality experienced significantly lower heart rates than any other combination (Table 14, rows 1 and 2).

Skin conductance responses (SCRs) are characterized by an increase in electrodermal response followed by a decrease in response, usually involving an increase of one or more microsiemens [7]. Each participant had between 0 to 15 SCRs per task (Tables 12 and 13, row 5). Participants had significantly more SCRs when completing interesting tasks than uninteresting tasks. They also experienced significantly more SCRs when the SERP quality was bad. There were no interaction effects.

The data were also examined for changes within the first 60 seconds of the task, where the stimulus is likely to have the most dramatic effect [7] (Tables 12 and 13, rows 6 and 7). There was a significant effect for SERP quality, with those receiving normal SERP quality experiencing greater skin conductance. There was also a significant interaction effect, with the highest skin conductance being found for those completing interesting tasks with normal SERP quality and the lowest for those completing interesting tasks with bad SERP quality (Table 14, row 3).

Participants experienced similar heart rates during interesting and uninteresting tasks within the first 60 seconds (Tables 12 and 13, row 7). A significant main effect for SERP quality was found for heart rate in the first 60 seconds, with those experiencing normal SERP quality having slightly higher heart rates. There was also a significant interaction effect, with those completing interesting tasks with normal SERP quality having the highest heart rates and those completing interesting tasks with bad SERP quality the lowest (Table 14).

Table 12. Means, standard deviations, and F-test results of physiological measures according to task interest, $^{*}p<.05$, $^{***}p<0.001$

	Interesting Tasks	Uninteresting Tasks	*F*-values	η^2
Skin Conduct. (SK)	7.19 (4.57)	7.15 (4.73)	0.02	<0.01
Heart Rate (HR)	79.57 (16.79)	79.88 (16.03)	0.24	0.01
Window SC	7.15 (0.47)	7.03 (0.60)	3.68	0.08
Window HR	77.70 (2.21)	79.36 (1.22)	35.97***	0.47
SK Responses	4.50 (4.18)	3.45 (2.99)	4.12*	0.10
First 60sec SC	7.38 (0.40)	7.39 (0.31)	0.59	0.01
First 60sec HR	80.43 (2.26)	80.17 (1.90)	2.63	0.04

Table 13. Means, standard deviations, and F-test results of physiological measures according to SERP quality, $^{***}p<.001$

	Bad SERP Quality	Normal SERP Quality	*F*-values	η^2
Skin Conduct. (SK)	7.22 (4.52)	7.12 (4.78)	0.12	<0.01
Heart Rate (HR)	79.38 (16.33)	80.07 (16.50)	2.93	0.07
Window SC	7.08 (0.33)	7.10 (0.70)	0.05	<0.01
Window HR	78.75 (1.45)	78.14 (2.33)	9.48***	0.19
SK Responses	4.79 (4.21)	2.91 (2.71)	16.40***	0.30
First 60sec SC	7.20 (0.37)	7.58 (0.40)	128.65***	0.69
First 60sec HR	79.78 (2.26)	80.82 (1.77)	27.56***	0.32

Table 14. Interaction effects for physiological measures, $^{*}p<0.05$, $^{***}p<.001$

	SERP Quality	Interesting	Uninteresting	F-values (η^2)
Window SC	Bad	7.04 (0.37)	7.11 (0.28)	14.45***
	Normal	7.26 (0.53)	6.94 (0.80)	(0.26)
Window HR	Bad	78.42 (1.64)	79.34 (1.05)	29.12***
	Normal	76.97 (2.48)	79.37 (1.38)	(0.41)
First 60sec SC	Bad	6.96 (0.32)	7.43 (0.23)	156.41***
	Normal	7.80 (0.29)	7.36 (0.36)	(0.73)
First 60sec HR	Bad	80.14 (2.66)	79.42 (1.72)	4.26*
	Normal	80.72 (1.76)	80.92 (1.78)	(0.07)

4.3.1 Multilevel models of physiological data

A mixed-effects multilevel growth model was performed to further investigate the relationship and variation between skin conductance, heart rate, interest and frustration. Growth models are a type of multilevel modeling predicated on time-ordered data. These models are especially suited for modeling physiological data (such as the ones gathered in this experiment) over time [7]. In this experimental setup, interest, frustration, and task order are fixed effects, while task time is a random effect. The coefficients of these effects essentially represent values by which skin conductance and heart rate can be predicted. Tables 15 and 16 detail the models of skin conductance and heart rate, respectively. A log likelihood ratio, or "L-ratio" was computed for each model. Log likelihood refers to the log taken of the "likelihood" score produced by the model, which serves as an indication of the probability of the observed values given certain parameters. In multilevel modeling, log likelihood scores are compared to test for significant differences. Thus, the L-ratio represents a test statistic by which significance is determined.

The best fitting model for skin conductance was Model 3, which only had interest and frustration as fixed effects. Model 3, when examined, showed a significant effect for interest, while there was no significant effect for frustration. The best fitting model for heart rate was the model with interest, frustration, and task order (Table 16). In the 3rd and 4th models, interest was non-significant while frustration was significant, indicating that

frustration contributed more to heart rate than interest. In the 4th model for heart rate, task order also significantly contributed to the model.

Table 15. Models of skin conductance data, $^{*}p<0.05$

Model	Fixed Effects	Random Effects	Coefficients	L-ratio
1	Interest (I)	Task Time	-0.51	3.45
2	Frustration (F)		-0.10	
3	Interest + Frustration		I: -0.63* F: -0.26	4.59
4	Interest + Frustration + Task Order (TO)		I: -0.59 F: -0.26 TO: -0.19	3.56

Table 16. Models of heart rate data, $^{*}p<0.05$

Model	Fixed Effects	Random Effect	Coefficients	L-ratio
1	Interest (I)	Task Time	0.32	0.61
2	Frustration (F)		-0.70	
3	Interest + Frustration		I: 0.15 F: -0.67*	1.20
4	Interest + Frustration + Task Order (TO)		I: 0.24 F: -0.72* TO: -1.08	4.89

5 DISCUSSION

The goal of this study was to investigate the differences in search behaviors and physiologies of participants when they are engaged or frustrated during online information search. This study was motivated by previous findings that seemingly contradicted one another: in some cases, increases in search behaviors such as clicking, querying and time spent, have been associated with engagement, while in other studies, these same measures have been associated with constructs such as frustration. While there has been some work integrating physiological measures into the study of engagement and frustration, to our knowledge there has been no study which has examined how the combination of physiological signals, search behaviors and self-report measures can potentially be used to distinguish between engaging and frustrating search experiences. This study was also novel in that it studied engagement and frustration simultaneously, and within the same participant. There were a number of significant results, which we summarize and discuss below. Importantly, many of the results had strong effect sizes.

Our study rested on our being able to induce engagement and frustration in a laboratory setting. Both the engagement and frustration manipulations were successful, and one contribution of this study is the method used to induce interest and frustration in the laboratory, which could be used by other IIR researchers. Participants reported significantly greater interest in, prior knowledge of, and relevance of tasks they ranked as more interesting than those they ranked as less interesting. They also indicated they had searched more frequently in the past for tasks they ranked as more interesting, and expected to experience significantly greater difficulty when completing less interesting tasks, although their post-search ratings showed no significant differences in experienced difficulty. There were also no differences in participants' post-search ratings of their own skills completing tasks, the system's ability to complete tasks, and their overall success with each task. These results provide support for the method used to induce interest and also show that participants' prior interest in a task may not have a strong impact on their perceived abilities to successfully complete the task in a laboratory setting. There were also no differences in the levels of frustration and stress reported by participants, which provides evidence that task interest may not be related to these affective components of search.

There were significant differences in participants' ratings of three of the five user engagement subscales, with participants reporting greater focused attention, felt involvement and novelty when searching for tasks they had previously ranked as more interesting. There were no differences in perceived usability, which is not surprising since the system components were stable (i.e., there were equal numbers of instances of poor and normal search quality for each of the interesting and uninteresting tasks). The lack of significant results regarding endurability suggest a person's lasting impression of a system might be more greatly impacted by factors other than task interest. Overall, however, these results support the idea that task interest is an important component of engagement, especially regarding dimensions that are more related to cognitive absorption, and that task interest can have a positive influence on the extent to which a person becomes engaged during search, which is consistent with, and extends, past research [9]. Researchers wishing to induce engagement in laboratory settings might continue to explore how interest can be exploited. Of course, one potential risk of allowing participates to pre-select tasks that interest them is a skew towards particular tasks. In this study, online communication was the most popular task, while vehicle purchasing and tattoo removal were the least popular tasks. The over- and under-representation of tasks can complicate data analysis and may impact the generalizability of the results.

Our manipulation of frustration was also successful. Participants reported significantly higher levels of frustration during search tasks with bad SERP quality, as well as significantly more distress and overall stress. There was no difference in the level of worry reported between groups, which suggests that while participants might have become frustrated and distressed during search episodes, they did not find their experiences worrisome. This could be a limitation of measuring these constructs in a laboratory setting – in real world search

situations, a person might find bad SERP results more worrisome.

Other post-search ratings also provide evidence that the frustration manipulation was successful. When completing tasks where poor results were presented, participants rated the tasks as significantly more difficult, rated their skill and the system's ability as significantly worse, and reported significantly less task success. Finally, the user engagement ratings provide evidence of how frustration impacts a participant's experiences: felt involvement, perceived usability and endurability were all rated significantly lower when poor search results were presented.

There were no significant differences between interesting and uninteresting tasks for nearly all measures of search behavior, with the exception of scrolls on the SERP and query intervals. The increased amount of scrolling on SERPs when conducting interesting tasks could reflect increased scrutiny of the search results. The significant main effect for time between queries also suggests that participants cognitively, rather than physically, engaged more with search results for tasks they found interesting (this time includes both time viewing SERPs and time examining documents).

While participants generally issued more queries, made more clicks, displayed more SERPs, and spent longer amounts of time when completing tasks that interested them, most of these differences were not significant. These negative results potentially have implications for the design of IIR studies where search tasks are assigned. A typical concern when assigning search tasks is that participants might exhibit different amounts of effort depending on motivation and interest. In this study, interest did not have a huge impact on user effort, at least as measured by things such as clicks, queries and time. However, we did not explicitly measure motivation and it might be that this has a larger impact on effort than interest, and that the relationship among motivation, interest and effort in laboratory studies is complicated and in need of further study.

There were a number of significant differences in participant search behaviors based on the quality of the search results. When presented with bad SERP quality (when participants were frustrated), participants submitted significantly more queries, made more clicks per query, displayed more SERPs, made more clicks on the SERP, made more scrolls and saved fewer documents. Most of these differences were quite strong, suggesting that physical measures of effort might provide greater evidence of frustration, rather than engagement. Participants also spent significantly more time completing tasks and more time per SERP, but they did not spend more time per document or more time between queries, which continues to support the notion that these measures might be better indicators of engagement. These findings complement other work on engagement and frustration in that they support findings such as the relationship between engagement and dwell time [15] and Feild et al.'s [13] finding that frustrated participants exhibit periods of high search activity.

With respect to the physiological measures, there were no differences in overall skin conductance or heart rate according to whether the task was interesting or not, or according to the quality of the search results. However, an examination of the data at different levels of granularity (window analysis) and at different points in time (first 60 seconds) did reveal several interaction effects for both skin conductance and heart rate. Participants had the highest skin conductance when completing interesting tasks with normal SERP quality, which suggests participants experienced some positive arousal in this condition since they did not self-report frustration or stress during these tasks. Participants in this condition also experienced greater skin conductance in the first 60 seconds of the search. The lowest skin conductance was exhibited by participants conducting uninteresting tasks with normal SERP quality, which further shows that the participant's level of task interest is what caused the positive arousal. Interestingly, the second highest skin conductance was exhibited by participants completing uninteresting tasks with poor SERP quality, which also suggests heightened arousal, only in this case it can be interpreted as negative when combined with the self-report data.

Participants completing uninteresting tasks had the highest heart rates, and those conducting interesting tasks with normal SERP quality has the lowest rates, which suggests that heart rate might be more associated with negative arousal, while skin conductance with positive arousal. The results of the multilevel modeling further support this: task interest provided the best predictor of variations in skin conductance, while frustration provided the best predictor of variations in heart rate.

6 CONCLUSIONS

This research offers insight into the utility of physiological signals in search evaluation, and provides evidence about the relationships between search behaviors, engagement and frustration. Specifically, it demonstrated the successful induction of engagement and frustration via experimental manipulation, as well as a novel approach to understanding competing emotional states present within the same participant. It also showed that physiological signals can be difficult to use as tools of disambiguation when examined during emotional states with high arousal. Lastly, it demonstrated that increased search behaviors were stronger indicators of frustration rather than interest. Other studies have characterized both engagement and frustration by an increase in search behaviors, and this research, through its linkage of subjective evaluation and search behaviors, clarifies some of the intricacies of these relationships.

This work is well situated among other work that has incorporated physiological signals in measuring subjective experience as well as objective measures such as search behavior [6]. It has confirmed that physiological signals can be useful as indicators of engagement, though their interpretation and application must be carefully handled. Other studies have characterized both engagement and frustration by an increase in search behaviors [15], and this research, through its linkage of subjective evaluation and search behaviors, helps clarify the intricacies of these relationships. The utility of the other physiological measures in discrimination is unclear, and it is also unclear how interest and frustration combine to produce

particular levels of arousal. It is possible that task interest creates a more pronounced response early in the session, and then as a person becomes frustrated, responses associated with this emotion begin to dominate.

This work also poses several questions, in particular: what advantage does disambiguating engagement and frustration offer? Both states could signal a need for customized search experiences, so characterizing these states in the context of behavior and tasks is the first step in deciding whether to intervene. Future work could include trying to identify a threshold over which participants pass from being engaged to unengaged, or frustrated to not frustrated. There also may be other ways to disambiguate engagement and frustration, such as incorporating more physiological signals to create a more comprehensive model of user state, or encouraging greater engagement in the laboratory by having participants bring in genuine tasks.

ACKNOWLEDGMENTS

This work was conducted when the first author was a doctoral student at the University of North Carolina at Chapel Hill. The authors wish to thank Ioannis Arapakis, Jaime Arguello, Rob Capra and Heather O'Brien for their feedback on this work as members of the first author's dissertation committee. The authors also wish to thank the School of Information and Library Science at UNC for its support of this research.

REFERENCES

[1] Abram Amsel. 1958. The role of frustrative nonreward in noncontinuous reward situations. *Psychological bulletin*, *55*, 102-119.

[2] Ioannis Arapakis, Joemon Jose, and Phillip D. Gray. 2008. Affective feedback: an investigation into the role of emotions in the information seeking process. In *Proceedings SIGIR*, 395-402.

[3] Ioannis Arapakis, Mounia Lalmas, George Valkanas. 2014. Understanding within-content engagement through pattern analysis of mouse gestures. In *Proceedings CIKM*, 1439-1448.

[4] Ioannis Arapakis, Ioannis Konstas, and Joemon Jose. 2009. Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *Proceedings of ACM Multimedia*, 461-470.

[5] Anne Aula, R.M. Khan, Zhiwei Guan. 2010. How does search behavior change as search becomes more difficult? In *Proceedings of SIGCHI*, 35-44.

[6] Miguel Barreda-Ángeles, Ioannis Arapakis, I., Xiao Bai, B. Barla Cambazoglu, and Alexandre Pereda-Baños. 2015. Unconscious Physiological Effects of Search Latency on Users and Their Click Behaviour. In *Proceedings of SIGIR*, 203-212.

[7] Wolfram Boucsein. 2012. *Electrodermal activity*. Springer Science and Business Media.

[8] Jason J. Braithwaite, Derrick G. Watson, Robert Jones, Mickey Rowe. 2013. A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments. In *Psychophysiology, 49*, 1017-1034.

[9] Mihaly Csikszentmihalyi. 2014. Flow. In *Flow and the Foundations of Positive Psychology*, 227-238.

[10] Georges Dupret and Mounia Lalmas. 2013. Absence time and user engagement: evaluating ranking functions. In *Proceedings of WSDM*, 173-182.

[11] Ashlee Edwards and Diane Kelly. How does interest in a work task impact search behavior and engagement? 2016. In *CHIIR 2016*, 249-252.

[12] Ashlee Edwards, Diane Kelly, and Leif Azzopardi 2015. The impact of query interface design on stress, workload and performance. In *Advances in Information Retrieval*, 691-702.

[13] Henry A. Field, James Allan, and Rosie Jones. 2010. Predicting searcher frustration. In *Proceedings of SIGIR*, 34-41.

[14] Joseph F. Grafsgaard, Joseph B. Wiggins, Kristy Elizabeth Boyer, Eric N. Wiebe, and James Lester. 2013. Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*, 8 pages.

[15] Ahmed Hassan, Ryen W. White, Susan Dumais, and Yi-Min Wang. 2014. Struggling or exploring?: disambiguating long search sessions. In *Proceedings of WSDM*, 53-62.

[16] William S. Helton. 2004. Validation of a short stress state questionnaire. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1238-1242.

[17] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-Ching Wu. 2015. Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *Proceedings of ICTIR*, 101-110.

[18] Richard S. Lazarus and Susan Folkman. 1984. Stress, appraisal, and coping. New York: Springer Pub. Co.

[19] Janette Lehman, Mounia Lalmas, Ricardo Baeza-Yates, and Elad Yom-Tov. 2013. Networked user engagement. In *Proceedings of the 1st Workshop on User Engagement Optimization*, 7-10.

[20] Janette Lehmann, Mounia Lalmas, Georges Dupret, and Ricardo Baeza-Yates. 2013. Online multitasking and user engagement. In *Proceedings of CIKM*, 519-528.

[21] Janette Lehmann, Mounias Lalmas, Elad Yom-Tov, and Georges Dupret. 2012. Models of user engagement. In *User Modeling, Adaptation, and Personalization*, 164-175.

[22] Yashar Moshfeghi, and Joemon Jose. 2013. An effective implicit relevance feedback technique using affective, physiological and behavioural features. In *Proceedings of SIGIR*, 133-142.

[23] Lise Solberg Nes, Suzanne C. Segerstrom, and Sandra E. Sephton. 2005. Engagement and arousal: Optimism's effects during a brief stressor. In *Personality and Social Psychology Bulletin*, *31*, 111-120.

[24] Heather L. O'Brien and Maria Lebow. 2013. Mixed-methods approach to measuring user experience in online news interactions. In *JASIST*, *64*, 1543-1556.

[25] Heather L. O'Brien and Elaine G. Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. In *JASIST*, *59*, 938-955.

[26] Heather L. O'Brien and Elaine G. Toms. 2013. Examining the generalizability of the User Engagement Scale (UES) in exploratory search. In *IP&M*, 49, 1092-1107.

[27] Adan Ortiz-Cordova and Bernard J. Jansen. 2012. Classifying web search queries to identify high revenue generating customers. In *JASIST*, *63*, 1426-1441.

[28] Lawrence H. Peters and Edward J O'Connor. 1980. Situational constraints and work outcomes: The influences of a frequently overlooked construct. In *Academy of Management Review*, *5*, 391-397.

[29] Jocelyn Scheirer, Raul Fernandez, Jonathan Klein, Rosalind W. Picard. 2002. Frustrating the user on purpose: a step toward building an affective computer. In *Interacting with Computers*, *14*, 93-118.