ENGAGED OR FRUSTRATED? DISAMBIGUATING ENGAGEMENT AND FRUSTRATION IN SEARCH

Ashleé Edwards

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Information and Library Science.

Chapel Hill
2015

Approved by:

Diane Kelly

Ioannis Arapakis

Jaime Arguello

Heather O'Brien

Robert Capra

© 2015
Ashleé Edwards
ALL RIGHTS RESERVED

ABSTRACT

Ashleé Edwards: Engaged or Frustrated? Disambiguating Engagement and Frustration in Search
Under the direction of Diane Kelly

One of the primary ways researchers have characterized engagement is by an increase in search actions. Another possibility is that instead of experiencing increased engagement, people who click and query frequently are actually frustrated; several studies have shown that frustration is also characterized by increases in clicking and querying behaviors. This research seeks to illuminate the differences in search behavior between participants who are engaged and frustrated, as well as investigate the effect of task interest on engagement and frustration. To accomplish this, a laboratory experiment was conducted with 40 participants. Participants completed four tasks, and responded to questionnaires that measured their engagement, frustration, and stress. Participants were asked to rank eight topics based on interest, and were given their two most interesting and two least interesting tasks. Poor search result quality was introduced to induce frustration during their most interesting and least interesting tasks.

This study found that physiological signals hold some promise for disambiguating engagement and frustration, but this depends on the time frame and manner in which they are examined. Frustrated participants had significantly more skin conductance responses during the task, while engaged participants had greater increases in skin conductance during the first 60 seconds of the task. Significant main and interaction effects for interest and frustration were found for heart rate in the window analysis, indicating that heart rate fluctuations over time can be most effective in distinguishing engagement from frustration. The multilevel modeling of engagement and frustration confirmed this, showing that interest contributed significantly to the model of skin conductance, while frustration contributed significantly to the model of heart rate.

This study also found that interest had a significant effect on engagement, while the frustrator effectively created frustration. Frustration also had a significant effect on self-reported stress. Participants exhibited increases in search actions such as clicks and scrolls during periods of both engagement and frustration, but a regression analyses showed that scrolls, clicks on documents, and SERP clicks were most predictive of a frustrating episode. A significant main effect for interest was found for time between queries, indicating that this could be a useful signal of engagement. A model including the physiological signals and search behaviors showed that physiological signals aided in the prediction of engagement and frustration.

Findings of this research have provided insight into the utility of physiological signals in distinguishing emotional states as well as provided evidence about the relationship among search actions, engagement and frustration. These findings have also increased our understanding of the role emotions play in search behavior and how information about a searcher's emotional state can be used to improve the search experience.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Diane Kelly, for being a source of support, guidance, and mentorship throughout my doctoral program. Her advice and encouragement have been invaluable throughout this process.

I would also like to thank Dr. Robert Capra for lending me his experimental search system. I’d also like to thank the other members of my committee, Drs. Ioannis Arapakis, Jaime Arguello, and Heather O’Brien, for their helpful feedback and guidance in developing and executing this dissertation.

I would like to thank Dean Gary Marchionini, and Associate Dean Barbara Wildemuth, for their support as I transitioned from the bachelor to doctoral program, and for their acknowledgement and encouragement of my potential. I would also like to thank my managers at ibiblio, Paul Jones and Cristóbal Palmer, for providing me with valuable years of work experience, flexibility, and the best work environment I could ever hope to work in. I’d like to specifically thank my coworker and friend, David Cowhig, for his help with troubleshooting many of the technical aspects of the project and providing a listening ear.

I would like to thank my friends and colleagues who have supported me in this dissertation process, specifically Samantha Kaplan, Kathy Brennan, Amanda Click-Drewry, Emily Vardell, Brenda Linares, Angela Murillo, and Sandeep Avula. You have provided listening ears and general sources of support and merriment in the years we’ve spent together.

Lastly, I would like to thank my family for their support, as well as my partner of five years, Justin Brinegar, for adapting to the distance between Chapel Hill and Cupertino, middle-of-the-night-must-get-up-and-type-this dissertation breakthroughs, the general grumpiness of graduate students, and cooking me a hot meal whenever I needed it.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Interactive information retrieval is the study of how the user interacts with a search system (Kelly, 2009). Work in this area is concerned with modeling both the search system as well as the cognitive and affective states of the user. According to Savolainen (1995), cognitive orientation in information seeking refers to “an analytic and systematic approach to problems” and research has identified cognitive structures as crucial to understanding search behavior (Ingwersen, 1996). The importance of the study of cognition in search has been demonstrated in many areas such as relevance judgments (Saracevic, 2007; Brennan, Kelly & Arguello, 2014), mental models (Vakkari, 2001; Zhang, 2008) and information-seeking behavior (Savolainen, 1995). Cognitive attributes of search have also gained prominence in many foundational models of search (Kuhlthau, 1993; Borlund, 2003). Affect and emotional states are also important to interactive information retrieval, as these offer utility in understanding concepts such as motivation and self-efficacy in search. Affective components are also present in models of search (Nahl & Bilal, 2007), and represent the emotional state of the user throughout the search.

Engagement is one cognitive and affective component of interactive information retrieval that has not been modeled as extensively as others. Engagement has been defined differently in different areas; in organizational psychology engagement is defined as when employees are filled with “vigor” and dedication” (Schaufeli et al., 2008). In cognitive psychology, engagement is defined as a state of greater goal orientation, perceived ability, and motivation (Shernoff et al., 2003). A common thread in the study of engagement is the creation of a positive experience for the user in order to encourage them to continue performing a particular activity. In contrast to this subjective approach, engagement in interactive information retrieval has been studied primarily from a behavioral perspective, with a focus on defining engagement through search actions (Jiang, He, & Allan, 2014; Teevan, Collins-Thompson, White, Dumais, & Kim, 2013; Lehmann, Lalmas, Dupret & Baeza-Yates, 2013). These measures have primarily been frequency-focused: frequency of unique clicks, issuing of more queries and more query reformulation, as well as greater frequency in overall activity per session. This research is beneficial because it demonstrates how engagement may manifest itself behaviorally.

Research in interactive information retrieval could benefit from more investigation of the cognitive and affective components involved in frustration. Frustration has a much more universal definition than engagement; it is most often defined as a “response to impediment of progress towards a goal” (Amsel, 1992). Work in information retrieval has shown that frustrated participants tend to exhibit both decreases in performance and increases in negative emotion (Aula, Khan & Guan, 2010; Poddar & Ruthven, 2010). Research has also shown that frustrating experiences occur when the participant experiences difficulties using the search system (Hoppmann, 2009), and that frustration can occur at different points in a search session (Hertzum, 2010). Frustration is typically also characterized by an increase in search actions, specifically an increase in clicks (Field, Allan & Jones, 2010).

The relationship between engagement and frustration has also been explored. In psychology, frustration and engagement are often linked through goal identification; people become frustrated when they are prevented from achieving their goals, and become engaged when they make progress towards their goals (Csikszentmihalyi, 2014). Motivation is also important as it determines the strength of goal orientation and thus the valence of emotion when progress is either made or blocked. Work in engagement and flow theory has also shown that goal orientation is key; Pace (2004) stated that directed attention (a component of flow and engagement) occurs when there is congruence between novelty, interest, and the goals of the searcher. Pace also stated that when a person either fails to find an item of interest or fails to do so quickly, they may become frustrated, emphasizing how easily engagement can shift to frustration. Similarly, Amsel (1992) defined frustration as a response to a system of "intermittent reward and non-reward" as a person progresses towards a goal. In frustrated states, participants can experience "behavioral activation" or an increase in behavioral response. Given the similarity in the cognitive and affective basis of these two states, it is possible that instead of experiencing engagement, people who click and query frequently are actually frustrated.

While behavioral signals can be useful, other work has offered greater focus on subjective measures. O'Brien and Toms have made significant advancements in shifting the perception of engagement as completely interaction-based to a focus on the cognitive and affective experience of engagement. O'Brien and Toms (2008) offer a definition of engagement which states it is 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). With this definition O'Brien and Toms successfully capture engagement as an interaction between system and user variables. This more holistic understanding of engagement has contradicted other frequency-based notions. O'Brien and Lebow (2013) investigated the relationship between engagement, search behavior, and stress, 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 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.

There is a theoretical basis for the similarities and differences between engagement and frustration. O'Brien and Toms (2008) identified challenge as a key component of engagement, however, other work has shown that frustration can occur when a participant perceives a challenge, but believe they do not have the resources to overcome it (Amsel, 1992). This difference in responses when one identifies a challenge versus an obstacle is encapsulated in the inverted-U theory (Yerkes & Dodson, 1908). The inverted-U theory of performance states some amount of stress can improve performance, while too much stress inhibits it (Muse, Harris, & Feild, 2003). This finding has been replicated in some of the literature on task performance in human factors. For example, Helton, Shaw, Warm, Matthews, and Hancock (2008) found that as participants moved from low to high workload (as they might do with a mentally challenging task), their level of engagement increased, but as participants moved from high to low workload, their level of engagement decreased. Matthews, Warm, Reinerman, Langheim, and Saxby (2010) suggested that task engagement is closely linked to energetic arousal and cognition as well as a need for success. This study showed that for more challenging tasks, participants who were more engaged were more likely to appraise the task as challenging and more likely to use task-based coping strategies. They also found significant correlations between cognitive indicators of stress such as self-reported stress and post-task engagement.

Cognitive indicators of stress and their accompanying physiological signals hold promise for disambiguating engagement and frustration. While it is well known that frustration generally produces higher stress (Scheirer et al., 2002), there is conflicting evidence about the relationship between engagement and stress. Work in psychology has found that as engagement increases, stress hormones also linearly increase (Nes, Segerstrom, & Sephton, 2005), which is supported by the underlying belief that cognitive activity expresses itself physiologically as stress (Cannon, 1927). However, O'Brien and Lebow (2013)'s work found negative correlations between electrodermal activity, heart rate, and self-report measures, indicating that as engagement increases, physiological stress decreases. Even if we place engagement and frustration at opposite ends of a spectrum, it is not clear how participants shift from being engaged to frustrated, or how this change is reflected in their stress response and search actions.

My research seeks to investigate the relationships among task interest, engagement and frustration, stress, and search actions. Task interest is introduced in this study as a means of potentially fostering engagement, and poor search result quality is introduced as a means of creating frustration. Figure 1 illustrates the constructs, relationships and questions for this study.

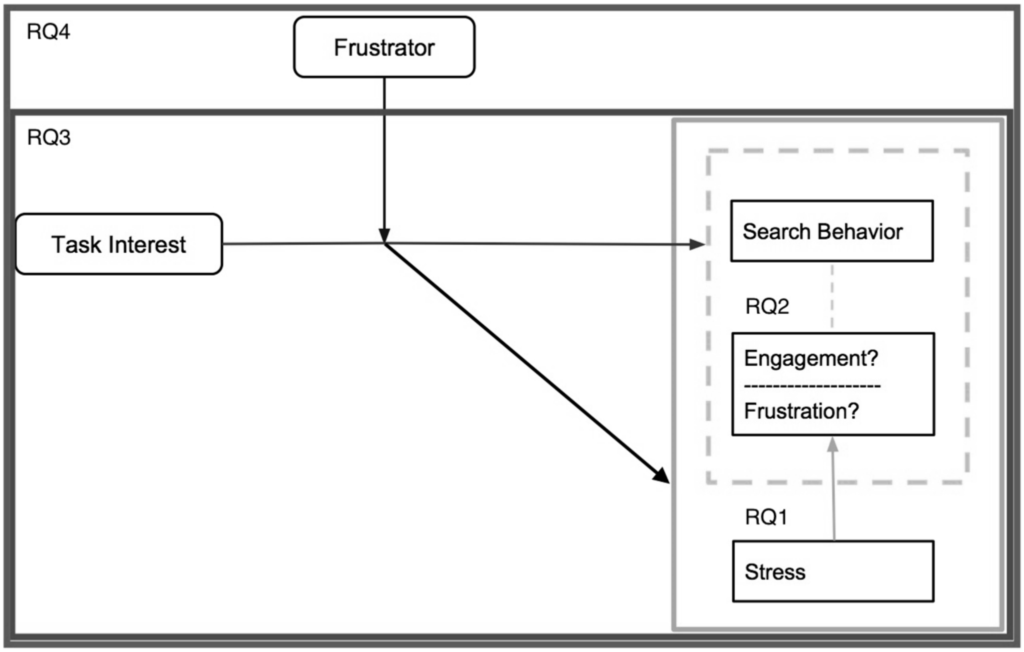


Figure 1. Proposed Model, Relationships and Research Questions.

1.1. Research Questions

The research questions for this study are:

RQ1: To what extent can physiological signals of stress help disambiguate the emotional states of engagement and frustration?

O'Brien and Lebow (2013) found that for an online news context, participants who were more engaged had lower levels of electrodermal activity and electromyography (facial movement), which contradicts other studies that suggest that the cognitive activation (such as that which is experienced during engagement) may lead to higher physiological signals of arousal (Yun, Shastri, Pavlidis, & Deng, 2009). Studies have shown that frustration involves an increase in physiological signals (Scheirer et. al, 2002; Partala & Surakka, 2004), and that negative affect is reflected in an increase in anxiety and stress (Fowles, 1987). Therefore, I hypothesized that:

H1: Physiological signals of stress during frustrating experiences will **exceed** those experienced during engaging experiences.

RQ2: To what extent do search actions differ for participants who are engaged versus frustrated?

Research has shown that there is a relationship between search behavior and emotional state. Aula et al. (2010) found that participants attempted longer queries and more frequent query reformulation when struggling with difficult tasks. Other work has shown that frustrated participants exhibit distinct click behaviors (Field et al., 2010). O'Brien and Lebow (2013)'s work on engagement and stress showed that participants with the highest levels of engagement performed the fewest search actions. Hassan, White, Dumais and Wang (2014) tried to classify search actions as either "struggling" or "exploring," and found that query similarity was a good indicator of struggling, as participants who were struggling tended to have less query diversity than participants who were exploring, as supported by Aula et al. (2010). They also found that dwell time was lower for participants who were struggling, and (though not significant) participants clicked more in exploratory sessions than struggling ones. Using this information, they found that behavioral information in the aggregate increased prediction accuracy; for example, query information up to and after the second query was highly predictive of a struggling vs. exploring session. This suggests that there are distinct patterns of behavior associated with positive and negative search experiences.

However, there are conflicting findings in each study. Where Aula et al. found that participants had greater dwell time when they were frustrated and struggling, Hassan et al. found there were shorter dwell times for participants who were struggling. Aula et al. also found that participants expanded their query terms by adding more words to form natural language queries, while Hassan et al. found that participants who struggled were more likely to substitute terms than add them. There were methodological differences in these studies that may have explained the opposing results; Aula et al. performed a user study, while Hassan et al. examined log data. Therefore, it was difficult to formulate a hypothesis about the relationship between engagement, frustration, and search actions. Therefore, this will be treated as an exploratory research question.

RQ3: How does interest in the task affect engagement, frustration, and stress? What effect does interest in the task have on search actions?

Previous work has shown that interest is critical to engagement (Csikszentmihalyi, 1991; Pace, 2004), and that users feel more frustrated during tasks that are important to them (Hertzum, 2010). My manipulation of engagement is constructed around these findings, but it will be necessary to check if this manipulation works, which is part of what this research question addresses. Research has also shown that some arousal is experienced during both engagement (O'Brien & Lebow, 2013) and frustration (Scheirer et al., 2002), but I believe it is likely that the arousal and stress accompanying frustration will exceed that of engagement based on the inverted-U theory (Muse, Harris & Feild, 2003) and its relationship to activation (Russell, 2003), meaning that the participants will experience more stress than they are able to cope with and will ultimately become frustrated. Therefore, I proposed the following hypothesis:

H3: Engagement will be greater during tasks that participants rated as interesting. I believe this is because participants will be more motivated to complete these tasks, given that we know that motivation, interest and goal orientation are integral to engagement (Csikszentmihalyi, 1991; Pace, 2004). This motivation will be reflected in increases in search actions.

RQ4: How does the presence of a frustrator moderate the relationship between interest in the task and search behavior, engagement, frustration, and stress?

Research has shown that motivation and goal-orientation are integral to the experience of engagement (Matsumoto & Sanders, 1988). Work has also shown participants feel most frustrated when they face an obstacle (Hoppmann, 2009), during tasks that are important to them (Hertzum, 2010), and that poor search result quality can induce search abandonment (Song, Shi & Fu, 2013). Therefore, I hypothesized that:

H4: Participants will also feel more frustrated (as induced by reversed search result rankings) during tasks that are interesting versus tasks that are not interesting. This frustration will be exhibited through physiological signals of arousal that are above and beyond those of engagement, as well as higher reported frustration.

1.2. Implications

This research will offer insight into the utility of physiological signals in search evaluation. This study will also provide a larger methodological comment on ways to effectively use physiological signals in information retrieval evaluation. Many other studies have made use of these techniques, and this dissertation will add to this body of research. This research will also provide evidence about the relationships between search actions, engagement and frustration. Other studies have characterized both engagement and frustration by an increase in search actions, and this research, through its linkage of subjective evaluation and search actions, can clarify the intricacies of these relationships. Lastly, this study will add to a larger understanding of the role emotions play in search behavior, and how information about a searcher's emotional state can be used to improve the search experience.

CHAPTER II: LITERATURE REVIEW

This chapter reviews previous work related to engagement, frustration, simulated work tasks, and the study of physiological signals. This review first examines theories of emotion, including engagement and frustration, and how these have been studied in interactive information retrieval. Lastly, the review concludes with a review of literature on physiological signals and recommendations for measurement, as well as a review of the literature on the development and use of simulated work tasks.

2.1. Emotion

The study of emotion occurs at three levels: at a physiological level, via endocrine and autonomic processes; at a neurological level, via neural activity and brain structures; at a cognitive level, by examining how people think about, process, and regulate their emotions; or at a behavioral level, by studying how emotion is physically expressed. This means there are many bases for defining emotion.

2.1.1. Definitions of Emotion. "Affect" is an umbrella term that has often come to describe three commonly grouped yet distinct phenomena: emotion, affect, and mood. Though emotion and affect are often used interchangeably, the psychological community understands affect and emotion as distinct concepts. Emotion is made up of seven components: core affect, behavioral response, attention, cognitive appraisal, attribution, and neurophysiological changes in the body in response to the emotion (Russell, 2003). "Affect" most frequently refers to "core affect," which is a combination of "hedonic values" (ranging from pleasure to displeasure) and "arousal values" (ranging from sleepy to activated)

(Russell, p. 154). The primary difference between core affect and emotion is that core affect does not involve the cognition inherent to emotion, such as cognitive appraisal and attribution. Mood, in contrast, is affect that persists over time and is not directed at any particular object. Together, these concepts can describe a person's emotional state. For the purposes of this review, I use the term emotion, since most work in the area focuses on emotion and not core affect.

2.1.2. Theories of Emotion. A common approach to studying emotion and cognition is the constructivist approach (Mandler, 1990), which states that emotion is the result of cognitive analysis and physiological response. William James and Carl Lange, considered the founding fathers of this approach (Cannon, 1927), believed that emotions were the result of a response to a physiological stimulus. Lazarus (1991) elaborated on this further, stating that cognition is a necessary but not sufficient part of emotion, meaning that thoughts can engender emotions (i.e., in cognitive appraisal) but emotions cannot occur without thoughts.

Lazarus' (1991) interpretation of the relationship between cognition and emotion served as the basis for appraisal theories, which state that the brain evaluates events for emotional cues. Reisenzein and Hoffmann (1990) mapped emotions to appraisal of specific events, and found that emotions like pride, shame and guilt are caused by events that are believed to be a result of a person's actions, while emotions like love and anger are caused by events that are believed to be the result of someone else's actions. Roseman, Dhawan, Rettek, Naidu, and Thapa (1995) proposed a theory that positive emotions are elicited when a person appraises an event as consistent with his or her objectives. Negative emotions are elicited when a person appraises an event as inconsistent with his or her objectives. In a search context, if a system presented a user with a relevant document, this would be appraised as consistent with the user’s effort to complete their search effectively. In the search context example, if the user were randomly presented with completely irrelevant documents, they would appraise this as inconsistent with their objective and would experience negative emotions.

In contrast to the constructivist approach, which posits that emotions occur after a person has evaluated a situation, there is the fundamentalist approach which states that there are fundamental emotions which are discrete patterns of behavior and experience. The embodied appraisal approach represents a middle ground between the constructivist and fundamentalist approaches. Prinz (2003) developed the embodied appraisal theory which states that emotions occur as a set of physiological changes, which over time become linked with emotions through learned patterns of behavior. This approach appears to successfully reconcile the multi-process nature of emotion.

2.1.3. Physiological and Behavioral Expression of Emotion. Another theory that links emotion and physiology is the James-Lange theory of emotion (1927), which states that emotion is the manifestation of a response to a physiological stimulus. Schachter and Singer (1962) tested the James-Lange theory of emotion by performing a series of experiments manipulating participants’ physiological states and measuring their emotional responses. They found that participants will label their emotional state if they experience physiological arousal without an immediate, obvious explanation (such as feeling embarrassed when one is flushed); with an appropriate explanation, participants will not label their emotional state. Participants also only described their current state in emotional terms if they experienced some type of physiological arousal.

The relationship between emotion and behavior is less well understood. Laird (1974) found that participants changed their description of their emotion based on what facial expressions they were told to make, indicating that the relationship between emotion and behavior is bidirectional. Baumeister, Vohs, DeWall and Zhang (2007) saw the relationship between unconscious emotion and behavior as dependent upon type of emotion, i.e., core affect could engender a “fight or flight” response, while conscious emotion is a mediated process in which emotion influences cognitive processes which then, in turn, affect behavior. Though Clore (1994) believed that emotions existed entirely in the realm of the conscious, Zemack-Rugar, Bettman and Fitzimmons (2007) unconsciously primed participants to experience an emotion, and observed that the primed emotion affected behavior. This shows that non-conscious emotions can have an effect on behavior.

2.1.4. Individual Differences in Emotional Response. Individual differences in the psychophysiological makeup of people can influence measures of emotion. Feldman (1995) identified two constructs key to understanding individual differences: valence focus and arousal focus. Valence focus is “the degree to which individuals attend to the hedonic component of their affective experience” (p. 295), which means the degree to which an individual notices and reports pleasant or unpleasant feelings. Arousal focus, similarly, refers to “the degree to which individuals attend to the arousal component of their affective experience” (p. 295), specifically physiological signals of arousal. Feldman conducted a longitudinal study of personality and mood, and found that participants’ reporting of their emotional state varied along the dimensions of valence and arousal, but also that these tendencies were not fixed, i.e.,, a participant who seemed to be low-valence-focus at one point in time later produced ratings more in line with a high-valence-focus because of temporal changes.

Individual differences may manifest themselves in various forms of emotion suppression and regulation. Gross and John (2003) separated emotion regulation processes into antecedent-based (before an emotion occurs) and response-based (after an emotion has occurred). They then looked at individual differences with respect to cognitive reappraisal (an antecedent-based strategy) and expressive suppression (a response-based strategy). Cognitive reappraisal is defined as when an individual reconstructs an emotional situation in a way that lessens its emotional impact (Lazarus & Alfert, 1964). Gross and John performed several experiments comparing individuals along racial, gender, and ethnic lines, and found that individuals who employ expressive suppression experience negative affect more often, perhaps related to feelings of inauthenticity. Expressive suppression is when an individual purposefully inhibits emotionally expressive behavior. Reappraisers, in contrast, experience more positive affect due to the "early intervention" of their emotion regulation strategy which shapes the way they feel/express emotions. Emotion regulation strategies can influence self-perception, self-reporting measures and behavior. Ohira et al. (2006) looked at the relationship between neural and physiological expressions of emotion during emotion suppression episodes and found that not only did certain brain regions such as the hippocampus and amygdala experience changes in blood flow, but skin conductance response was also enhanced during an emotional suppression episode. However, individual differences in emotional suppression resulted in varying responses, indicating that emotion modulation influences the expression of physiological signals differently for different kinds of people.

2.1.5. Emotion and the Stress Response. Four concepts are key to the understanding of stress and emotion: an agent that causes the stress (the stressor), the evaluation of the stressor, a coping process to deal with the stress, and the effects of that stressor on the mind and body. The coping processes employed in stressful situations involve both a psychological and physiological response. It is well understood that emotional regulation mechanisms can inform the experience of stress. Lok and Bishop (1998) looked at emotion regulation strategies and perceived stress, and found that benign impulse control and mental rumination were related to lower perceived stress. They also found that emotion inhibition was negatively correlated with stress. Gohm, Carser, and Darsky (2005) looked at emotional intelligence and stress among freshmen, and found that individual differences played a key role in whether emotional intelligence reduced stress; individuals' confidence in their emotional intelligence was a mitigating factor in reduced stress. Brosschot, Gerin and Thayer (2006) found that perseverative cognition (meaning prolonged worry or rumination) could prolong the physiological and immunological activation that happens during stressful episodes, resulting in greater susceptibility to illness. Studies have also shown that emotion can have the opposite effect on stress. Brownlow (2009) measured cortisol levels while participants were asked to recall humiliating experiences, and found that cortisol levels were high pre-experiment, but low during the actual recall of the events, indicating that the thought of recalling humiliating events may have been more stressful than the actual experience of recalling them. Brownlow further suggests that recalling these events may actually relieve them of their stressful aspects. It is clear that emotion is a fluid concept and that the relationship between emotion and stress is bidirectional, which makes untangling emotion and the stress response difficult.

2.2. Modeling Emotion and Information Behavior

Emotion is important to information behavior, as evidenced by the presence of emotional components in information-seeking models. However, affect in these models deviates from the psychological understanding of affect. “Affect” and “affective” are used in the information-seeking literature to represent any model or process with an emotional component. Affective will be used in this section where researchers make use of the term.

The Information-Seeking Process model (ISP) developed by Kuhlthau (1993) describes both the cognitive and affective states of students engaged in information-seeking to complete a class assignment. This model consists of six stages: initiation, selection, exploration, formulation, collection, and presentation. Each of these stages has an affective component: the person has apprehension at the beginning of the task, feels a sense of optimism followed by confusion and uncertainty, which culminates in a feeling of preparedness once the task has been completed and new information has been acquired.

Kracker (2002) used Kuhlthau’s model to explain library and research anxiety. Library anxiety refers to anxiety experienced as a result of needing to use library services to complete an assignment. This anxiety may at times affect the patron’s ability to effectively use the library. Research anxiety is similar to library anxiety except that research frequently occurs outside of the library. Kracker asked participants to rate their anxiety and cognitive and affective awareness levels over eight weeks after listening to a 30-minute presentation of Kuhlthau’s Information-seeking Process Model. Kracker found that participants who listened to the presentation had less anxiety than participants who had listened to a generic presentation, but there were no changes in cognitive or affective awareness levels between groups.

A different, perhaps better application of Kuhlthau's model was done by Hyldegård (2009), who wanted to explore whether the model was applicable within the context of academic group work. Hyldegård makes it clear that emotion, not affect, is being measured and asks participants to indicate their emotional state by rating their emotional experience on a scale of 0-5 in relation to six positive feelings (confidence, satisfaction, optimism, relief, motivation, and clarity) as well as seven negative feelings (confusion, doubt, stress, frustration, uncertainty, and worry). While the items on this questionnaire are problematic because many of them deviate from Ekman's six basic emotions, the emotions measured are closely aligned with the stages outlined in Kuhlthau's model. Hyldegård also explores a key component in the measurement of emotion: valence, which refers to the degree a person experiences feelings as well as how aware they are of what they feel. This is a slightly more effective application of emotion measurement, because of the specificity in construct definition as well as in emotion measurement.

Another model proposed in information science that has emotional components is Savolainen's (1995) model of Everyday Life Information-Seeking. Savolainen breaks everyday life into three components: (1) way of life, referring to a person's time budget, hobbies, and models of consumption; (2) mastery of life, referring to the kinds of approaches people take to solve problems; and (3) problem solving behavior. The model also takes into account variables like social, cultural, and material capital, values, and health. The emotional component of this model is in the mastery of life portion, which contains four types: optimistic-cognitive, pessimistic-cognitive, defensive-affective, and pessimistic-affective. Savolainen draws a line between cognitive and affective components, saying that cognitive orientation means "an analytic and systematic approach to problems," and affective orientation means “emotionally laden and rather unpredictable reaction to issues at hand” (p. 265). With these descriptions, Savolainen remains closer to the definition of affect as emotion without cognitive appraisal. He also introduces attitudinal components (optimism and pessimism) that offer deeper insight into problem-solving. In the world of emotion literature, psychologists typically refer to these attitudinal components as “dispositional” optimism or pessimism, and have long known that disposition can have lasting effects on affect or mood (Frijda, 1988).

Though the model is in keeping with affect as understood in psychology, Savolainen does not explore the validity of the affective components through traditional psychological tests. Instead, he pursues an indirect method of observing problem-solving behavior in order to understand emotional experience. Using critical incident technique and asking participants to extensively document a non-work information-seeking context, Savolainen found the participants (teachers and workers) were mainly located in the pessimistic-cognitive mastery of life. This orientation is described as systematic problem solving with less than optimal expected outcome. Fewer participants fell into the optimistic-cognitive category, which refers to systematic problem solving with the expectation of positive outcomes. One participant was placed in the pessimistic-affective category, which Savolainen calls “learned helplessness,” or avoiding systematic problem-solving. This classification was related to how affect was elicited through the critical incident technique. The outcome of the incident affects the post-incident emotion and the participant’s feelings are directly related to the type of incident used to anchor the discussion. For example, the participant who lost her job was in the midst of an unsuccessful job search, and her description of the incident was likely to be negative given that she had not found a job yet. Savolainen acknowledges that if any of the participants had chosen different incidents, including one that might have been less “affectively sensitive,” they might have been placed in different orientations. Savolainen’s model offers a more nuanced way of looking at how emotional differences may impact information-seeking behavior.

Nahl (2004) moves closest to the traditional definition and measurement of affect by defining specific affective components of the information-seeking process. Nahl sought to understand the emotional environment of searchers by measuring motivation, information-seeking, self-efficacy, time pressure, search optimism feeling, coping skills, effort, acceptance of search environment, and affective load. These constructs were measured using a 26-item questionnaire. Students who completed research projects for a writing class were asked to fill out these items related to searches they did for their research projects. Nahl found high ratings of affective variables such as felt effort, satisfaction, and affective load.

One challenge in interpreting Nahl’s findings is due to the way in which affective concepts were defined and measured. Nahl describes "affective load" as being composed of uncertainty and time pressure, but the definition involves the person employing "coping skills to avoid giving up on the task" (p. 3), which seems more indicative of cognitive load than affective experience. The components of this cognitive load, uncertainty and time pressure, both involve distinct cognitions, further reinforcing the idea that this is a purely cognitive rather than affective state. Nahl's definition of search environment acceptance, which focuses on how supported a person feels during search, is also unusual. Nahl's questionnaire asks the participant if he/she feels supportive of the search engine and how easy it was to use the search engine, but these items seem to deal more with perceptions of search engine performance rather than affective experience. Lastly, there is a conflation of “affective” with "emotional," but the use of a self-report questionnaire helps paint a more complete picture of emotional state as tied to an information-seeking event.

The affective sub-components Nahl defines outline a key part of affective experience: valence focus (Feldman, 1995). The sub-components that measure positive or negative feelings such as search optimism facilitate the participant's reporting of valence focus. However, Nahl's sub-components do not include another important component of affective experience, arousal focus. The modeling of affective information behavior in information science has provided us with clues as to what role emotional state can play in information-seeking behavior and information processing. However, these models could be expanded to include a deeper understanding of the distinction between emotion and affect and applied in such a way as to measure emotion more broadly, as well as taking into account physiological factors.

The discussion of affect within the information-seeking community has focused on how a person feels as they engage in information search. Though this approach borrows heavily from psychology's understanding of cognition as a process of self-reflection, it is unique because of the focus on individual differences in how a person responds to and interacts with their information environment. Nahl and Bilal (2007) in their book, *Information and Emotion*, outline a theoretical framework specifically for understanding information-seeking and affective principles. Nahl and Bilal place their definitions of affective behaviors in an information-seeking context within the same groups as psychological analysis of affect: cognitive, affective, and sensorimotor (similar to cognitive, physiological, and behavioral).

However, for Nahl and Bilal, these groups act in sequence: first, the sensorimotor action occurs, where the person takes note of their environment, then the cognitive, where the person interprets the information around them, and then affective, where the person evaluates information emotionally. This model is quite different from the way cognition, affect, and behavior are understood in psychology; behavior is seen as directly influenced by cognition and affect rather than cognition and affect following behavior.

This work, combined with studies of emotion done using web search tasks, shows that a person's mood influences their search efficacy as well as their information-seeking strategies. This body of work also shows that search can engender different valences of emotional response (Lopatovska & Arapakis, 2011).

2.2.1. Emotion in Human Computer Interaction. Affective computing in human computer interaction research has focused on measuring frustration and other negative emotions experienced while interacting with interfaces, as well as how to decrease negative emotions. For example, Fogg and Nass (1997) found that participants responded to negative and positive feedback from computers similarly to the way they responded to the same types of feedback from humans. This study found that participants reported more positive affect and spent longer on the task when they received help. Conversely, participants who received little help made significantly more mistakes, spent less time on the task, and reported less positive affect. This demonstrates that the way a user perceives the sympathetic response of the computer affects emotional state and performance on the task.

These ideas from this initial work were derived from those put forward in *The Media Equation* (Reeves & Nass, 1996), which posits that humans respond to media, specifically computers, as if they were responding to another human. The authors argue that responding this way is an automatic response and is as natural as interactions between humans. Other work done by Reeves and Nass (1996) found that humans respond to computers socially. In their study, a computer congratulated itself on its performance, and participants were asked to rate its performance electronically or via a paper questionnaire. Participants rated the computer more positively on the electronic evaluations, almost as though participants did not feel comfortable being honest with the computer “to its face” versus “behind its back.” This study found that when computers initiated a social situation, participants would respond in kind and even consider the “feelings” of the computer when responding to it. Other researchers have explored these ideas within the context of frustration (Kapoor, Burleson, & Picard, 2007; Grafsgaard et al., 2013). Frustration will be covered elsewhere in this review, in section 2.5.

Within the area of affective computing and search, behavioral signals have been combined with subjective response to show that affect detection can be used to improve the Web search experience (Wang, Chignell, & Ishizuka, 2006; Klein, Moon, & Picard, 2002). Lopatovska (2009) explored the relationship between mood and search task, using measurement of stress via the Positive and Negative Affect Schedule (PANAS). The PANAS is used to measure the degree to which a participant feels positive affect, characterized by attributes such as alertness, enthusiasm, and negative affect, or the degree to which a participant feels distress. This study found that search task difficulty, topic, and complexity did not have an impact on pre-task or post-task mood.

Though Lopatovska found that mood was not influenced by search task properties, Gwizdka and Lopatovska (2009) found that while higher happiness levels indicated better feelings during the search process, happier participants reported lower levels of satisfaction and poorer search outcomes. They also found that participants judged a task to be more difficult if it took longer to complete, and participants were unable to accurately predict difficulty before engaging in the task. In terms of behavioral signals, they found participants who viewed more pages reported feeling lost during the search less often but were overall less satisfied with their searches. Moving from detection and classification, to prediction, Lopatovska (2011) found correlations between emotions and search actions such as mouse clicks and scrolling, and suggested that these behaviors could potentially be used to detect a user's affect and trigger interventions to improve the user's emotional state.

2.3. Stress

2.3.1. Theories of Stress. Theories of stress, similarly to theories of emotion, have focused on people's perception of stress and their physiological response to it. Selye (1976), an endocrinologist, developed a theory of stress called General Adaptation Syndrome (GAS), which consists of three stages: alarm reaction, resistance, and exhaustion. Selye recognized that acute and chronic stress were processed by different parts of the body. Acute stress is processed by the sympathetic adrenal-medullary system (SAM), and chronic stress is processed by the hypothalamic pituitary adreno-cortical system (HPA). In the alarm stage, an acute stressor is detected, a fight or flight response is triggered, and the sympathetic branch of the autonomic nervous system sends electrical signals to the brain which trigger an adrenaline response. In the resistance stage, the stressor persists and the endocrine system sends signals to the pituitary gland to release cortisol, which becomes detectable in the saliva and the blood. During exhaustion, the last stage, the body becomes very susceptible to illness as the adrenal glands lose functionality. Chemicals such as adrenaline and cortisol are responsible for producing reactions such as increased heart rate and increased skin conductance. Selye’s GAS theory ties stress to the physiological response, as the constructivist theories of emotion tie emotion to physiological responses.

Other researchers have pointed out several weaknesses in Selye’s characterization of stress (Hobfoll, 1989). The most prominent one is that in his GAS model, responses such as anxiety or fear, which may precipitate stress but have different origins, become indistinguishable from the stress Selye identifies. Lazarus and Folkman (1984) characterized stress as a process of coping and appraisal, rather than a series of physiological responses. Though the concept of appraisal has already been covered earlier in this review, Lazarus states that primary appraisal (which is key to stress), is made up of goal relevance, goal congruence, and ego-involvement. Goal relevance refers to whether a situation is meaningful to the person experiencing it. Goal congruence refers to the extent to which an experience is in line with the person’s overall goals. Ego involvement refers to whether the person experiences ego-related feelings such as self-esteem. These cognitive appraisals allow an individual to determine whether the stimulus is a stressor. Specifically, Lazarus and Folkman identify the three types of appraisals that are involved in stress: harm, which refers to psychological damage or stress that has already happened, threat, which is the anticipation of harm, and challenge, which is stress someone feels from a demand they feel they can overcome. This cognitive understanding of stress can also help us understand coping behaviors that occur in the presence of a stressful situation.

2.3.2. Definitions and Types of Stress. Lazarus and Folkman define psychological stress as a “relationship between the person and the environment that is appraised by the person as taxing or exceeding his or her resources and endangering his or her well-being” (p. 40). Specifically, they divide stress into two types: either stimulus-response-defined or relationally-defined. The relational definition of stress states that a stress response occurs as the result of a combination of individual differences and environmental variables. Stimulus-response stress involves stimuli that engender a stress response; these kinds of events are outside of an individual’s control and can be global (natural disasters, war), or part of a local environment (chronic illness, divorce). People respond to these stressors by employing coping mechanisms that are highly individualized. However, there are also smaller events outside of one’s control (referred to as daily hassles), which Lazarus and Folkman consider even more impactful because they help establish long-term patterns of coping.

Selye (1976) and other researchers also recognized the importance of differentiating between environmental eustress and distress. Eustress is defined as the “good” kind of stress, primarily occurring as a result of positive motivation. Distress, the “bad” kind of stress, is the result of negative feelings. The important distinction between these two kinds of stress is the perception of the stressor by the person. If the stressor is perceived in a positive light, then it will be considered eustress, and if it is perceived negatively, then it is considered distress. There is evidence to support eustress’ ability to enhance immune systems while distress impairs it (Selye, 1975).

2.3.3. Physiological and Behavioral Expression of Stress. As stated earlier, stress affects the activation of certain areas in the brain that promote arousal, awareness, and survival mechanisms. The relationship between stress and behavior involves a complex interaction between the nervous system response and cognitive appraisals, in addition to difficult-to-predict individual coping mechanisms. Work has suggested there may be a relationship between cortisol and behavioral distress (Anders, Sachar, Kream, Roffwarg, & Hellman, 1970; Tennes & Carter, 1973), though there is no consensus on the strength of the association. Other studies have suggested that increased heart rate and cortisol are linked to increased anger and sadness respectively (Lewis, Ramsay, & Sullivan, 2006), confirming that different physiological responses to stress are controlled by different mechanisms (the cortisol by the adreno-cortical system, the heart by the autonomic nervous system). Lazarus and Launier (1978) posited that coping behavior in stressful situations depended on the person’s ability to identify resources for dealing with the stressor, and if their resources were minimal, they could attempt to regulate their distress but very little could be done to ameliorate it. Krantz (1983) placed participants within the stressful condition of preparing for an exam, and measured their stress responses, cognitive appraisals, and coping mechanisms. Krantz found that identification of a large and diverse number of options prior to a stressful event was related to more efficient coping behavior after the stressful event, confirming the complexities of the behavioral and physiological response to stress.

2.3.4. Individual Differences in Stress Response. Lazarus (2007) explains that one way of understanding stress is as a combination of individual differences in motivational and cognitive styles. Coping mechanisms vary greatly from person to person given that some people tend to ascribe stressful events to themselves, while others tend to ascribe them to outside events beyond their control. Individuals who attribute stressful events to themselves will experience higher stress than those who attribute stressful events to outside factors. Vollrath (2001) reviewed the literature on individual and stress as mediated by personality, specifically understood through the Big Five personality traits. The Big Five (McCrae & Costa, 1999) consist of openness, conscientiousness, extraversion, agreeableness, and neuroticism. These personality traits are important because they influence the way a person construes a stressful situation as well as how they appraise potential stressors. Parkes (1994) looked at the role of personality and environmental factors with regards to stress, and identified several kinds of interactions that produce stress. One kind is person-environment interactions, which consist of neuroticism and work demand, where individuals who are more neurotic are less likely to be adaptive to stressful events. Parkes found that while all variables were important individually, the only two that had significant interactive effects were person and situation, which, when combined, had a direct effect on coping strategies.

Given that the psychological community has explored the relationship between emotion, stress, and their physiobehavioral expression, we can learn a lot about how these relationships may be observed in the fields of affective computing and information science. Possible areas for exploring the relationship between stress and emotion in the field of information science include exploring whether participants' emotional experiences of an information retrieval session are related to greater reports of subjective stress, given that we already know participants can identify and report frustrating search experiences (Feild et al., 2010).

2.3.5. The Inverted-U Theory. The inverted-U theory, based on work done by Yerkes and Dodson (1908), states that there is a middle ground at which more arousal stimulates a positive response, and too much or too little results in negative response. Yerkes and Dodson experimented with stimulation of mice and maze navigation and found that "as the difficultness of discrimination is increased the strength of that stimulus which is most favorable to habit-formation approaches the threshold" (p. 22). The researchers conclude that there is a point above and beyond which the strength of the stimulus does not encourage the formation of habits. This basic idea has been applied in many different fields, in particular in the human computer interaction and ergonomics fields. In our discussion of activation and appetitive motivation (see section 2.5), the inverted-U theory lends itself to understanding where appetitive motivation becomes aversive and leads to frustration.

Näätänen (1973) saw the relationship between the inverted-U, activation, and performance as when a participant experiences a moderate level of activation during which they experience an optimum level of performance. Once they move past this, they experience an "increasing level of activation" past the optimal point and a subsequent "deteriorating level of performance" (p. 160). Näätänen suggests that five reactions can occur when a participant performs a task: anxiety, a "try harder" reaction, a "task-fatigue" reaction, self-consciousness (awareness of one's actions), and free association (thinking aloud to make sense of the task attributes or requirements). These reactions act as stimuli to spur the participant into activation. Note that while this effort will show traditional signals of "activation" which may manifest physiologically, the underlying emotions will be negative, behavior will change, and performance will decrease. This fits well into our understanding of frustration mirroring engagement because while the participant may appear to be engaged, the effort they expend during a frustrating session is markedly different in terms of the rate of expenditure of cognitive resources as well as affective state.

Work using the inverted-U model has also investigated whether adding more workload improves some aspects of participant performance. Wiener, Curry, and Faustina (2003) gave participants vigilance tasks, increased the mental workload of these tasks, and then measured the number of errors committed by each participant as an indicator of performance. They found that participants who were given tasks in which the load increased committed fewer errors than participants who were given tasks that required lower mental effort. The researchers surmise that this may be because tasks that require higher mental workload are already requiring the participant to be more vigilant, so they will be more aware of potential pitfalls and avoid committing errors. Participants who feel less stimulated are not as vigilant and thus may be more prone to making mistakes. This idea helps us make the case that a small amount of stress has the potential to increase performance, but an abundance of stress can cause errors in performance. Muse, Harris and Feild (2003) suggest that the literature on the inverted-U theory and the relationship between stress and performance can be improved and expanded by employing objective measures of stress instead of relying solely on self-report data. This study seeks to do this by using physiological stress as an objective measure.

2.4. Engagement

Engagement in pyschology has been characterized by a focus on creating a positive subjective experience to encourage people to continue performing a particular activity (Schaufeli & Salanova, 2008). Engagement in information search research has been anchored by these same ideas: discovering what makes a system engaging, and creating search experiences that promote engagement. While some researchers have created and evaluated psychometric scales for measuring engagement (c.f., O’Brien & Toms, 2008), engagement has become increasingly characterized by definitions that rely heavily on extrapolation of behavioral signals, such as number of clicks and dwell time.

In addition to this, the term engagement has been used with increasing frequency by researchers to describe search actions, even if no conceptual definition is offered. The use of behavioral signals to measure engagement is potentially problematic because these signals can be noisy and often difficult to interpret.

2.4.1. Definitions of Engagement. Engagement has been defined as a persistent and pervasive state containing both affective and cognitive attributes, but not focused on any particular object or event (Schaufeli & Salanova, 2008). It has been understood as related to, and perhaps a subset of, flow, which is defined as a state of complete cognitive absorption and focus on an activity (Csikszentmihalyi, 1991). Csikszentmihalyi stated that flow occurs when a proper balance between challenge and skill is reached. He conducted a study examining flow states and participants described their experiences of flow as balancing their need for information with search challenges and the limits of their own search skills. Some of the attributes of flow identified in this study were mental alertness, a sense of control and a reduced awareness of irrelevant factors. Schaufeli and Salanova (2008) extended this by stating that engagement is the peak of flow, suggesting that engagement causes flow, or at least creates the psychological state through which flow may be experienced.

Some of the work done on flow can help us understand and situate the work done on engagement and web search. Pace (2004) used grounded theory to investigate participants' experience of flow and information-seeking behavior during both directed searching and exploratory episodes. Pace wanted to observe the goal-orientation integral to flow. In directed search the user has a well-defined goal, while in exploratory search the goal is more amorphous and ill-defined. Pace found that though most participants reported feeling very confident in their search skills, they experienced challenges in the query reformulation stage as well as determining relevance. All participants reported feeling flow when they made progress towards their information goals, and feeling blocked from experiencing flow when they were unable to find information. Pace's work highlights the role of successful goal-orientation in information-seeking. In his concept map, the goal of the task positively influenced the user’s experience of feeling challenged, which is integral to flow attributes such as mental alertness and engagement attributes such as focused attention.

2.4.2. Defining and Measuring User Engagement in Search. Researchers in information retrieval have defined engagement (and related attributes such as interest) by an increase in search actions. Though engagement is often defined as “interaction with a system,” many researchers have parsed out engagement at the behavioral level to include specific behaviors. Bian, Dong, He, Reddy and Chang (2013) defined engagement as when a user examines a piece of recommended content, but also state that a click is indicative of engagement because it shows that the user looked at a link. Lehmann and colleagues have conducted a number of studies that provide a good illustration of how engagement has been studied in the context of large-scale search logs (Lehmann et al., 2012; Lehmann et al., 2013; Lehmann et al., 2013). In one of their first studies, Lehmann, Lalmas, Yom-Tov and Dupret (2012) proposed and evaluated three interaction-based models of engagement: a general model, a time-based model and a user-based model. Using search log data from millions of people, three measures of engagement were defined and examined in the context of each model: popularity, activity, and loyalty. *Popularity* was defined as the number of users that visit a site (including number of clicks). *Loyalty* was defined as the frequency with which a person returns to a site and how often they dwell on the site. *Activity* was defined as total dwell time on the site and number of page views per visit. Lehmann, et al.’s general model of engagement focused primarily on popularity and clicks on a site, 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. (2013) continued this work by proposing the concept of *networked user engagement*, which refers to engagement within a network of websites. This work focused on user clicks among different websites and posited users with high network engagement would make clicks among the websites within the network. They found that users performed more goal-oriented behaviors on a weekday (Wednesday), while they performed more browsing activities on the weekend. They also found that some users who were more active with regards to search behavior (referred to as VIP users) navigated more frequently between sites and had higher rates of return to previously visited sites than users who were less active. This conceptualization differed from the previous one in that it focused on activity within a collection of websites as opposed to activity at an individual website.

Lehmann, Lalmas, Dupret and Baeza-Yates (2013) furthered their work on engagement by focusing on user engagement with many tasks simultaneously, and analyzed online multitasking and engagement using two behavioral signals: dwell time and page views. Transforming these signals into metrics like attention shift, attention range, cumulative actions, visits, and sessions, Lehmann et al. grouped different kinds of sites based on levels of engagement and proposed a model in which dwell time and page views were conceptualized as tree-streams, or paths through which participants click at the session level. Shopping and mail sites were found to have high activity per visit and also short times between visits, indicating that participants progressively became more focused on their tasks. Search sites, front pages, and auction sites had lower dwell time overall but higher dwell time per session, and had high cumulative activity numbers, indicating that participants spent more time completing more activities. The most engaging set of sites had high ranges of attention shift and attention range, indicating that when participants did return to the site, they spent more time than before.

Though behavioral signals are useful in that they can serve as an indicator of variance in what the user is viewing and clicking on, using behavioral signals alone to define engagement is problematic because of their noisiness; some of these signals could also indicate confusion about the task or uncertainty about where to find information. They also do not effectively capture the cognitive or affective parts of engagement (Schaufeli & Salanova, 2008). Nonetheless, these signals still offer scalable and useful measurements of user behavior at relatively low cost, and so do offer their merits.

2.4.2.1. Redefining Engagement in Information Retrieval. O'Brien and Toms have made significant advancements in shifting the perception of engagement as completely interaction-based to include the cognitive and affective components of engagement. In their paper defining a conceptual framework of engagement, O'Brien and Toms (2008) state that engagement is 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). With this definition O'Brien and Toms capture how system properties and user behavior combine to create engaging experiences. Through analysis of semi-structured interviews, the authors isolate elements within the process of engagement, which is made up of four stages: point of engagement, period of engagement, disengagement, and reengagement. In this process, engagement begins when something captures the user's attention, and this attention is sustained, assisted by the challenging nature of the task. The user becomes disengaged when they choose to stop doing the activity, or when an external factor intervenes. Re-engagement occurs when the user decides to return to the task. Non-engagement, or not becoming engaged at all, occurs when users feel time pressure or are multitasking. This process-oriented model is useful not only because it can be used to track how engagement changes throughout an information-seeking task, but it also provides some context for interpreting interaction signals and engagement.

Building upon this work, O'Brien and Toms (2010) created a 31-item scale to measure engagement, called the User Engagement Scale (UES). Six attributes of engagement were identified using factor analysis: *perceived usability, aesthetics, focused attention, felt involvement, novelty,* and *endurability*. These factors capture the cognitive, affective, and usability-related attributes of user experience. A second study reported in this paper explored the relationships between these factors, and found, among other things, moderate correlations between *novelty* and *felt involvement,* and *focused attention* and *felt involvement*. They also found that *focused attention* predicted *felt involvement*, and *felt involvement* predicted *endurability*.

The UES has been used to measure engagement in interactive information retrieval studies. O'Brien and Lebow (2013) used the UES in conjunction with the Cognitive Absorption and System Usability Scales to examine what attributes were important during information-seeking experiences within an online news context. They found that participants who rated their level of interest in an article higher were also more engaged, solidifying the relationship between engagement and interest. They also found that participants who were less engaged spent more time browsing and visited more web pages, but participants who reported the highest levels of engagement spent the least amount of time browsing, visited the least amount of web pages, and spent the least time reading. Arapakis, Lalmas, Cambazoglu, Marcos, and Jose (2014) used the *focused attention* subscale of the UES in conjunction with other measures to observe what attributes of news articles and comments were important. They found that participants who read articles they labeled as interesting exhibited higher levels of focused attention. They also found that interest in the article and enjoyment experienced from reading it were higher when the participant had a strong sentiment and negative connotations.

2.4.3. Search Tasks. Search tasks are an important part of information retrieval, and engagement with search tasks has been explored in the literature. Though behavioral signals can consist of many types of interaction data, the behavioral signals used in engagement work are generally very focused on frequency - frequency of visits, frequency of re-use of search engines, and frequency of clicks. This approach is based on the belief that more interaction indicates more effort, more cognition and perhaps more investment in completing the task. Though measuring interest instead of engagement, Jiang, He and Allan (2014) illustrated this when they measured how different kinds of tasks over the course of a long session (10 minutes) affected search behavior, relevance judgments, and interest in the task over time. Participants were given tasks that varied across two axes: factual (information gathering) or intellectual (enhancing the participant's understanding of the topic) and specific (with a specific goal) or amorphous (without a specific goal). These two dimensions created four sets of tasks: known item (factual and well-defined), known subject (factual and amorphous), interpretive (intellectual and well-defined) and exploratory (intellectual and amorphous).

Jiang et al. found varying patterns of activity depending on the type of task. Participants searched more frequently but were less active in searching for exploratory and known item tasks, while they searched less frequently but were more active in their searches for known subject and interpretive tasks. In terms of "activeness," participants spent more time on the SERP for the KS and IN conditions and clicked on more results. Overall, participants in the exploratory search session were the most active in terms of frequency of search and SERP views. Though Jiang et al. were measuring interest, this kind of activity gives us important clues about engagement. Given that this task was the most open-ended of all the tasks, it may have encouraged more active search actions and thus been more engaging. Based on gaze fixations, Jiang et al. also note that participants spent more effort examining result abstracts for known item and exploratory tasks. In general, Jiang et al. found that more complex tasks led to more complex browsing and searching behavior. This reminds us that different task attributes can foster different indicators of engagement.

Engagement and cognitively complex tasks has also been examined. Arguello, Wu, Kelly and Edwards (2012) investigated whether blending vertical results into web results or allowing the verticals to be accessed indirectly had an effect on search behavior with complex tasks. Part of this study design was to measure how participants assessed the system, and this was done using the *focused attention, felt involvement, perceived usability, endurability*, subscales taken from the UES. Though they did not find any significant differences in interface ratings based on the subscales, they did compare interface preferences, and found that system evaluations were largely dependent on the participant rather than the interface. They found that people who preferred the non-blended interface rated it higher across the board, specifically for attributes such as endurability and perceived usability. These findings illustrate that one kind of search interface may not suit everyone, and part of understanding engagement is recognizing that different kinds of interfaces can be more appealing and seem more engaging to different kinds of people.

Kelly et al. (2015) created cognitively complex tasks with five levels: analyze, create, evaluate, understand, and remember. These tasks were also developed within four domains: health, entertainment, science and technology, and commerce. Forty-eight participants evaluated these tasks for difficulty, engagement, and their search behavior was recorded. There was a significant difference in engagement ratings between tasks, such that participants rated evaluate and create tasks as more engaging than remember tasks. Even though there were no significant differences in pre-search ratings of interest between tasks, post-search, participants rated evaluate and create tasks as significantly more interesting than remember tasks. Also, levels of interest increased significantly for create tasks over the course of the search.

2.4.4. Simulated Work Tasks. This review will discuss researcher-defined and participant-defined search tasks, focusing particularly on one type of researcher-defined task, simulated tasks. A type of search task that is especially useful for potentially engendering engagement is simulated work tasks. Simulated work tasks are designed to offer the benefits of both genuine and assigned tasks in that they simulate a potentially natural information need and inspire motivation but can be controlled in an experimental context. There is support in the literature for the belief motivation is critical to task completion. Matsumoto and Sanders (1988) looked at differences in engagement with the task when participants chose tasks that were intrinsically motivated versus extrinsically motivated. Participants were given definitions of extrinsic and intrinsic motivation (mainly that extrinsic motivation was a motivation to solve a task by an external goal or reward, and intrinsic motivation was defined as motivation to solve a task for the activity of solving the task itself), and then asked to describe an instance of each of these types of tasks, and how engaged they felt during it.

Matsumoto and Sanders found that participants reported higher happiness during intrinsically motivated tasks than during extrinsically motivated tasks. They also found that participants felt more happiness or pleasure when they were close to finishing extrinsically motivated tasks. Lastly, they found that interest in the task stayed high between tasks for intrinsically motivated tasks. In the context of simulated tasks, this means that engaging participants in a topic they find pleasing or useful to solve could result in higher motivation and differences in search behavior.

There is also support in information retrieval literature for the motivating properties of simulated work tasks. Ingwersen (2011) explains that while the simulated work task has elements that the traditional assigned information-seeking task does not, it incorporates many of the same elements. According to Ingwersen, socio-cognitive attributes are important for simulated work tasks. This means that work tasks must offer some sort of utility to the person completing them, which affects their retrieval behaviors and relevance judgments. In addition to this, the attributes of an information-seeking context such as the seeking process, queries, and documents, are also important as they pertain to the larger work task that the user is trying to solve. This model fully grounds simulated work tasks as a useful tool in IR evaluation, and highlight some of the potential differences in behavior.

Some researchers have used the word “natural” to describe participant-defined tasks (Vakkari, 2003; Capra, Sams & Seligson, 2011) while others have used the word “genuine” (Russell & Grimes, 2007). “Assigned” has been used frequently when referring to tasks created by the researcher (Bilal, 2002) as well as “imposed” (Byström & Hansen, 2005). Vakkari (2003) offers a helpful delineation of the two task types by conceptualizing them as search goals. Natural search goals refer to tasks that were originated by the user, simulated search goals refer to tasks that try to imitate a natural search goal, and assigned search goals refer to tasks that do not attempt to model natural search goals. For the purposes of this review, natural will be used to refer to tasks that were created by the user, assigned will refer to tasks created by the researcher that do not simulate a search need, and simulated will refer to researcher-defined tasks that model a natural search task.

2.4.4.1. Natural Tasks. Natural search tasks have been used in information retrieval evaluation to observe naturalistic search behavior, and have assisted in understanding behavior observed during assigned search tasks. Academic search tasks have received special treatment in the area of natural tasks because of their usefulness in studying motivated search behavior. Work has shown for academic tasks, people employ search refinement strategies as they progress through their search. Vakkari refined Kuthlthau's Information Seeking Process model (ISP) by investigating the information processes of students who were completing a master's thesis. Vakkari found that the students narrowed their search terms as the task progressed and used more search tactics and search operators. Specifically, the students' mental models became more refined as they gained more topic knowledge. They also used more synonyms and parallel search tactics as the task progressed. This work formed the basis of an information search process theory of the task-based information retrieval process. In Vakkari's theory, the task process begins with the construction of the mental model of search, which determines the search tactics and the specificity of the information search and ends with relevance judgments. This work parsed out many cognitive aspects of natural search goals.

Lee, Paik and Joo (2012) used a diary study to examine how undergraduates selected resources they needed for academic tasks. Participants were allowed to select individual or group assignments. Lee et al. found that participants selected more online resources than print or human resources, and used Google most frequently, followed by individual web pages and scholarly databases. The diaries revealed important attributes involved in resource selection. Credibility, accessibility, ease of understanding, and coverage of related material were major factors in choosing a particular resource. However, the most physically accessible resources (online databases) were considered less accessible, and more familiar resources (such as colleagues or friends) were also considered less credible. This demonstrates that there may be a wider breadth of accessible resources for people completing natural search tasks.

Work has also shown that natural tasks can be a useful evaluation tool. Capra, Sams and Seligson (2011) examined engagement with natural tasks during collaborative search. Participants searched in pairs, and each pair was given four imposed tasks and one natural task. Engagement was measured using a seven-point bipolar scale. They found that participants felt significantly more engaged with natural tasks than assigned tasks. Also, participants felt more engaged with exploratory tasks that required a decision as well as assembling different types of information. This study identified some attributes of natural search tasks that create motivation and engagement and also explored search behavior during natural tasks.

2.4.4.2. Assigned Tasks. While there has been some treatment of natural tasks in the literature, there have been more studies examining search behavior and assigned tasks. These studies have followed the traditional IR evaluation model which involves assigning search tasks in order to observe differences in performance. Wang, Hawk and Tenopir (2000) investigated the search behavior of students completing assigned tasks, paying particular attention to cognitive style and affective behaviors. Participants exhibited different search strategies; generally, they began with a query and narrowed focus as they gained more topical knowledge, but there were occasional changes in search engine between tasks. Participants also exhibited coping strategies when they experienced problems during the search. If they could not find information on a page, some participants backtracked to find results while others used a library homepage to find information. Lastly, Wang et al. found differences in participants' mental models of search. One participant assumed that everything listed on the Web was current, and another believed that typing a query and then clicking on a facet would search for a combination of those two things. This study showed differences in participants' conceptions of search and search strategies, which were more observable because the tasks were assigned and thus held constant during the experiment.

2.4.4.3. Natural Versus Assigned Tasks. There is a well-established body of literature on differences in search behavior between natural and assigned tasks. The advantage of using both task types in one study is not only to observe the similarities and differences in behavior between the two, but to observe the utility of natural tasks in traditional IR evaluation. Russell and Grimes (2007) also looked at the differences in search behavior between assigned and natural tasks in naturalistic settings. Participants in this study were given a list of 45 tasks to complete in the same environment they were did natural search tasks in (i.e., their home). Participants were also encouraged to make their work on these tasks "as near to their normal search activity as possible" (p. 3). These assigned tasks were simulated work tasks and were grouped into five categories: general search, local information, product information, image search, and news search. Twenty percent of the tasks participants completed were their own tasks. Participants were directed to "search for something on the web you would like to search for" that was "something you are genuinely interested in finding" (p. 3). Russell and Grimes found that participants spent more time on their own tasks than assigned tasks, had fewer unique queries for their own tasks, and returned to the SERP much more often for natural tasks than assigned tasks. This work seems to indicate that participants had a smaller range of behavior for their own tasks, which may be because of prior knowledge of the topic as well as knowledge of the scope of the task. This suggests that assigned tasks likely elicit more exploratory behavior because there is more uncertainty. However, the length of time spent on natural tasks seems to indicate a level of interest and engagement with the task that is not present in assigned tasks.

This is confirmed by Xie (2009), who found that participants were able to shift their search goals more easily for natural tasks. Participants were able to either broaden or narrow the scope of natural task as they progressed throughout the search, allowing for a more dynamic search strategy. In tasks that were assigned, participants felt less freedom to change their search goals or search strategy, because the scope of the task had been previously outlined. Xie also found that individual knowledge played a larger role in decisions regarding search strategies for natural tasks than assigned tasks. In practical terms, this means that the query terms and information resources selected by participants were dictated by the wording of the assigned tasks, while for natural tasks, participants had to rely on their own knowledge to generate query terms and information resources.

2.4.4.4. Defining Simulated Work Tasks. Research in the area of defining and understanding simulated work tasks has focused on dissecting the components of the task and how search behavior differs from assigned tasks. Byström and Hansen (2005) highlight the notion of task as process, stating tasks are "manifested through their performance," i.e., that tasks are conceptualized as centering on a particular item of work. This means that there is one unchangeable goal of a task, though there may be many ways of accomplishing that goal. This view is important to understanding the difference between natural and simulated tasks because it focuses on understanding the information behavior required to complete the task, which may vary depending on how familiar or comfortable the user feels with the task. Byström and Hansen also highlight the authenticity of tasks as an important part of understanding information-seeking. For Byström and Hansen, authenticity affects performance because the environment in which a task is completed is connected to what resources a person can access when completing a task. In addition to the presence of environmental variables, Byström and Hansen believe that natural tasks are subject to changes in performance because they present an authentic information need (which again commands a certain level of resources) with consequences that can affect the behavior of the user. In fact, they state that real-life tasks with consequences will likely result in an "authentic engagement in the task performance" (p. 1052). Thus, task contexts contain three types of attributes: contextual, situational, and individual. Contextual attributes remain stable over the course of a task, situational attributes are less permanent, and include things such as prior knowledge and available information sources, while individual attributes such as motivation may shift more readily during task completion. These attributes of natural tasks, therefore, should be reflected in simulated work tasks, in order to foster this authentic engagement.

2.4.4.5. Early Work in Simulated Work Tasks. Borlund and Ingwersen (1997) were two of the first researchers to perform IR evaluation with simulated information needs. Their study looked at both system performance and the relevance assessments of participants with both simulated and real work tasks. Their simulated tasks consisted of three parts: an indicative request, a definition, and a simulated work task situation. The indicative request was a directive that stated the information need (i.e., “find for instance something about critical success factors” (p. 5)). The definition stated what the object of the information need was (in this case, critical success factors). The simulated work task situation then contextualized the information need by providing a backstory (i.e., your boss has told you to prepare a report on critical success factors). Interestingly enough, this model of simulated work task is closely related to the format of TREC topics, which also contain sections such as domain, topic, narrative, and concept. The TREC framework was useful because it provided clear communication of information need, which is necessary for a simulated work task. However, the TREC topics very clearly outline qualifications for relevant documents, while the simulated work task allows for “user interpretations of the situation, leading to cognitively individual information needs” (p. 6). Thus Borlund and Ingwersen identify a crucial component of simulated work tasks: interpretation, which allows motivation to ensue. With regards to search behavior, Borlund and Ingwersen found that participants drew their queries from the text of the simulated work tasks. They also found that there were no differences in the way participants modified queries for simulated work tasks or their own tasks, leading them to posit that the personal information needs (or natural tasks) can be as useful as assigned tasks in IR evaluation.

Borlund (2000) oriented the simulated work task within IR evaluation further by describing how it bridges the gap between the traditionally system-driven evaluation approach and cognitive-centered evaluation approach. According to Borlund, simulated work tasks address the issues of “experimental control and realism” in three ways. Borlund defines a simulated work task as a short "'cover story' that describes an IR requiring situation" (p. 76). This description of the cover story generates an information need that the user is motivated to resolve. In addition to this, the user develops subjective and individual information needs based on the work task. The work task thus allows researchers to observe the effect of the system on realistic information needs, and prompts users to "dynamically assess the relevance of retrieved information objects" (p. 76) in ways that are different to assigned tasks.

Borlund also offers several recommendations for creating simulated work tasks. She advises that the level of "semantic openness" (p. 77) of the task dictates how users interpret the task and their subsequent search actions. In order for the user to perceive the task as realistically as possible, a sharing of a "universe" in the cognitive sense among participants is required. This means that the simulated task situation should present some measure of realism to all participants in that it is relatable and understandable. Where "real" information needs consist of a need that is of "personal interest and importance to the user" (p. 77), a simulated work task should create an external situation that stimulates an internal cognition - in other words, it creates a cognitive process similar to a personal information need.

This dynamic assessment of information needs is related to another crucial component of situated work tasks - situational relevance. Borlund defines situational relevance as "an assessment which points to the relationship between an information object presented to the user and the cognitive situation underlying the user's information need."(p. 77) This means that situational relevance is constantly being assessed during the search session. Simulated work tasks invoke situational relevance because they incite information needs that develop and mature over the course of the task, where traditional IR tasks have a concrete information need with predefined relevance criteria that does not encourage this continuous assessment.

2.4.4.6. Creating Simulated Work Tasks. Researchers have used this understanding of simulated work tasks to construct their own for use in IR evaluation studies. Li and Belkin used a faceted classification developed by Li and Belkin (2008) consisting of task attributes such as complexity level, subtasks of the task, and the goal of the task. To ensure that this was a simulated work-task, the tasks were revised versions of real work tasks collected in another study. Interaction signals were measured. Pre- and post-task questionnaires were given to evaluate participant perception of the task and the search process. Participants were asked to think-aloud during their searches, and exit interviews were conducted. Li and Belkin found that participants used more library sources when conducting schoolwork-related work tasks and for non-schoolwork work tasks, participants used more resources overall, in particular more web resources. Li and Belkin found that participants exerted more effort to find relevant information using library resources; this may be because more people considered themselves experts in using search engines versus using OPACs. They did not find significant differences in query behavior among work tasks, which Li and Belkin state may indicate that search tasks affect querying behavior more readily than work tasks.

Svarre and Lykke (2014) looked at how simulated work tasks could serve as a tool for IR evaluation in specific work contexts. The authors began by creating their tasks in a systematic fashion. First, they carried out a domain study with their population (e-government employees) to both understand the domain and glean ideas for work tasks. They used this information to design ten tasks that each had three search concepts or less. The work tasks were then pilot tested alongside a genuine search task to verify their clarity and realism. Though it is normally difficult to find time for multiple studies, this empirical evaluation of the tasks is useful because it addresses potential reliability issues during task construction.

Once these steps were completed, Svarre and Lykke had participants complete the simulated tasks, assessing both system performance and user experience of via think-aloud, questionnaires, and interviews. They found that participants had a “medium” level of knowledge about the topic of the search task, but did not experience great difficulty in completing the tasks. Participants surprisingly also did not find the tasks to be overwhelmingly similar to their own daily tasks. In completing these tasks, participants made use of structural, topical, and common knowledge. Participants used topical knowledge when they were working on topics that they had little experience with, and related knowledge was also used to supplement a lack of topical knowledge. Structural knowledge (or knowledge about the structure of the information need) was used when they completed the work task. Svarre and Lykke found that knowledge of the task was more important to task success than similarity to a genuine work task. These findings demonstrate that while the aim of simulated tasks is realism, creating an information need that the user is able to satisfy is most important. In experimental contexts, this means that simulated tasks should address the knowledge structures of the people they are designed for, rather than emulating tasks they should be able to complete.

2.4.4.7. Measuring Performance with Simulated Work Tasks. Studies that look at simulated work-tasks and search behavior often have a common methodological approach. These studies often incorporate objective measures such as search interaction measures, and subjective self-report measures. There is also usually some measure of cognitive strategy, most often think-aloud protocol, but interviews may also be employed for the same purpose. The effect of this methodological triad is that these studies can present a multifaceted picture of search behavior: search behavior, the motivations behind it, and evidence of a particular search strategy.

Li and Hu (2013) used both simulated and real work tasks to evaluate the usefulness digital library. Li and Hu do not describe how they created their simulated work tasks. They also gave participants criteria for the real tasks they were required to bring in, specifying that the task should have been part of a recent class assignment. Participants completed pre and post-task questionnaires, as well as an evaluation questionnaire which assessed items such as search skills and performance. Li and Hu found significant differences in topic familiarity and search experience between simulated and natural tasks, in that participants were more familiar and experienced with their own tasks. However, there were no significant differences in other aspects of the task, such as complexity, urgency, and difficulty in making relevance judgments. This indicates that participants did not feel more urgency in their own tasks, though Li and Hu claim that it is "obvious" that the real task would seem less complex and more urgent than the simulated task. One important significant difference was the difference in knowledge of task procedure; participants felt that they had less knowledge of the procedure to complete the simulated task than the real task. This is related to Svarre and Lykke's (2014) identification of structural knowledge as important to how people perceive and complete simulated tasks. Participants reported low ability to predict the difficulty of the real task, and found it harder after searching. They also felt low ability to predict the difficulty of the simulated task, but they found it easier after searching. Participants submitted more queries for real tasks, but viewed more search results pages, downloaded more documents, and had slightly longer queries for simulated tasks than real tasks. Participants also felt more success, frustration, and satisfaction with real tasks than simulated tasks, but these differences were not significant. However, Li and Hu found that feelings of success were significantly correlated with confidence and perceptions of task complexity and topic familiarity for the real task, i.e., participants felt more success if they were familiar with the topic. For the simulated task, satisfaction was significantly correlated with task difficulty and knowledge of task procedure, again illustrating the effect of structural knowledge on the ability to complete the task. Though the findings from this work are mixed, and many results are not significant, it does present some evidence for differences in perception and search behavior during simulated tasks. This challenges the idea that they represent a perfect compromise between effective system evaluation and realistic information needs.

Poddar and Ruthven (2010) investigated how natural and assigned tasks affect the emotional aspects of the search experience. Participants were given three different assigned: factual, complex, and exploratory. Participants were also asked to bring in their own task, which was not restricted by any criteria. Verbal utterances, observed actions, and questionnaire data were used in this evaluation. Generally, participants felt that the simulated tasks were less interesting than their own task, but there were no significant differences between the factual task and the participants' own tasks, suggesting that, at least in structure and content, the genuine tasks may have been most similar to the factual task. There were more positive emotions present before and after the natural task than the assigned tasks. Also, participants tended to bring in tasks similar to ones they had completed before, and ones they had topical knowledge on. This could explain the tendency to estimate lower task difficulty for natural tasks. Participants also used more search strategies, and expressed more positive body language as well as greater confidence in their search for natural tasks. Participants had similar levels of interest in all of the assigned task types, but struggled to form queries with the exploratory task, and also struggled with deciding the steps to completing the complex task. This study showed that task source can contribute to the emotional state of the participant, which in turn affects their search behavior. Simulated work tasks should seek to emulate natural search tasks and produce similar emotions.

Borlund, Dreier and Byström (2012) conducted two studies comparing perceptions of time spent searching between simulated work tasks and natural tasks. In the first study, the researchers created three simulated work task situations, which were pilot tested, and then evaluated by means of questionnaires and relevance assessments as well as post-search interviews. Borlund et al. asked participants to bring in their own tasks and advised them that their information needs should be either verificative (checking a specific piece of information), conscious topical (finding information about a familiar topic) or muddled topical (exploring an unknown topic). This framing illustrates one issue with eliciting natural tasks in IR evaluation: shaping. For comparison purposes, even natural information needs must be categorized and refined.

Borlund et al. found that most participants in the first study said that time spent searching for their own topics (and one simulated topic) was due to its interestingness, while participants in the second study said that interestingness contributed to time spent searching more for ‘conscious topical’ needs, followed by muddled topical and verificative. 55% of the participants in the first study said that time spent searching was an indicator of the simulated task involving a lot of information, while 67% of participants said they felt time spent searching on the verificative information need was a result of the topic having a little information. Lastly, 86% of participants said time spent searching was an indicator of a simulated task being too easy, while 67% of participants in study 2 said time spent searching was due to the verificative information need being easy, versus 56% who said the muddled topical need was difficult. Overall, Borlund et al. found that interest was a main indicator of time spent searching, but given that a variety of reasons were explored in this study, interest cannot completely explain time spent searching.

In addition to the findings, this study offered the opportunity for comment on the methodological implications of simulated work tasks. In comparing the two studies, Borlund states that participants tended to search longer during simulated tasks, which could be an indication of "over-performance" in an attempt to please the researcher. Therefore, though simulated tasks are designed to mimic real information needs, there may necessarily always be a distinct difference in the behavior of simulated tasks because of the experimental context - naturalistic search tasks in an experimental setting still did not yield similar task completion times.

2.4.5. Search User Interfaces. The search user interface aids users "in the expression of their information needs, in the formulation of their queries, in the understanding of their search results, and in keeping track of the progress of their information-seeking efforts" (Hearst, p.1). While researchers in information search have been interested in designing usable interfaces for quite some time, they have only recently moved beyond a focus on functional requirements and adopted the position that search interfaces should also be engaging, and that search experiences should be pleasurable (Belkin, 2008). The work reviewed in this section has either used the UES to evaluate search interfaces, or used terms like engagement when describing the goals and outcome of the work.

One of the first studies to use the UES to understand user experience in the context of search interfaces evaluated the display of vertical search results (Arguello, Wu, Kelly and Edwards, 2012). Arguello et al. (2012) examined differences between an interface that blended vertical results into web search result pages and an interface that displayed vertical results separately on individual search result pages that could be accessed via tabs. Arguello et al. (2012) did not use the complete UES in their study and also made several modifications to the items they did use. This limited their ability to make strong claims about the validity of the modified set of items, which had been established in previous work (O'Brien and Toms, 2010). However, previous testing of the UES was done in an ecommerce setting and the authors argued that the changes were needed to make the scales more suitable for the evaluation of search interactions. Most of the changes consisted of replacing words like "shopping" with "searching." In addition, the researchers dropped the aesthetics sub-scale as the basic elements of the interface remained constant throughout. Finally, the researchers indicated they deleted one item from each of the attention and endurability sub-scales after pilot participants reacted unexpectedly to them. Ultimately, Arguello et al. (2012) used the following UES sub-scales: Focused Attention, Felt Involvement, Perceived Usability and Endurability, and added a sub-scale about search effectiveness. Reliability analysis of responses to these modified sub-scales demonstrated these items had good reliability.

Arguello et al. (2012) did not find any significant differences between responses to these items according to interface. They went on to compare participants' interface preferences with their post-task questionnaire ratings on these sub-scales, and found that people who preferred one interface rated it higher along all aspects, specifically for attributes such as endurability and perceived usability. Participants who preferred one of the interfaces said they found it more visually appealing and felt the information was better organized and easier to understand, reinforcing the importance of usability. These findings are interesting because they show that engagement is related to a person's preferences and without knowing this preference, aggregate engagement scores for two or more interfaces might appear similar even when they produce different user experiences.

Moshfeghi, Matthews, Blanco and Jose (2013) evaluated whether adding a timeline and a named-entity component to a news search system would improve engagement, and whether engagement could be predicted based on interaction data. They created a search interface where a participant clicked on search results that were presented on a timeline in order to access content. In addition to the timeline, they added a list of entities for a given search result. For example, for a given entry such as (US) republican debates, the named entity list contained items such as "Newt Gingrich," "Herman Cain," and "Rick Perry."

Participants were recruited from Mechanical Turk and given explicit instructions about the assignment and how much time they would have to complete it (120 minutes). Engagement was measured using the UES. Similar to Arguello et al. (2012), Moshfeghi et al. (2013) modified the UES by changing the wording of the items for a news context, and each question was structured to ensure forced-choice instead of a range of values. They found that participants who used the enhanced interface rated felt involvement, endurability, novelty, and aesthetics (sub-scales of the UES) higher, which demonstrates the importance of moving beyond a purely functional assessment to more completely understand the user's search experience.

Bateman, Teevan and White (2012) created an interface where participants interacted with their previous search data and were able to compare themselves to three archetypes: the typical participant, search experts, and topical experts. Search experts were defined as frequent users of search operators, and topical experts were defined as having visited ten search results within the category. One version of the interface allowed participants to compare themselves to these archetypes, and the other did not. Engagement was derived from interactions found in participants' log data, specifically attributes such as time spent examining search results, likelihood of returning to the dashboard, and an affective learning dimension. The researchers (without mentioning engagement directly) also referenced engagement when they discussed participants' interest in learning and insights when using the interface. Participants were most interested in, and felt they gained more insight about themselves from the data about characteristics of search engine use and data on special search engine features and advanced query operators they viewed. Participants rated the comparison interface much higher than one that did not allow comparison, and were also more likely to report that the comparison interface would alter their search behavior later. Unlike the studies described above, this study focused on people's interactions with personalized content.

2.4.5.1. Perceived Usability and Control. Another key attribute of engagement is the perception of control (Csikszentmihalyi, 1991), which contributes to the perceived usability component of engagement. In O'Brien and Toms' UES, there is an item that states "I felt in control of my shopping experience," which loaded on the perceived usability factor. Perceived usability is related to control because poor interface design leaves users feeling lost and disoriented (Teevan et al., 2012), which contributes to feelings of lack of control.

Work done by Chen, Wigand and Nilan (2000) expanded upon this notion of control as integral to keeping participants engaged by asking participants about what made their experiences enjoyable via an open-ended questionnaire. Participants reported experiencing flow most frequently when “surfing,” or browsing. Most participants reported feeling that they were “always in control,” which seems to indicate that they felt totally certain at first. However, analysis of their descriptions reveals contradictions. One participant said that they did not know where to look for information when they initially began searching, but “an hour later or so, [they] hit the information and that gives me a good feeling of power” (p. 273). Csikszentmihalyi (1991) stated that neither total certainty or total uncertainty, but rather a balance of the two, facilitates the experience of flow, and creating this delicate balance of uncertainty and certainty as well as how users perceive it, can help us as researchers better facilitate engagement.

2.4.5.2. Content and Architecture. Research has also shown that the content of the information sources with which a user interacts play an important role in engagement. Arapakis, Lalmas, Cambazoglu, Marcos, and Jose (2014) used the focused attention subscale of the UES in conjunction with other measures to observe what attributes of news articles and comments were important to participants. They examined several attributes: genre, sentimentality of the article (the richness of the emotional tone of the article), polarity (positivity or negativity) and time of publication. Articles were then selected from three categories: crime, entertainment, and science. Participants indicated their interest before and after the task. Arapakis et al. found that participants who read articles they labeled as interesting exhibited higher levels of focused attention. They also found that interest in the article and enjoyment experienced from reading it were higher when the topic of the article had a strong sentiment and negative connotations.

Linking content focus and attention, Rohklenko, Golbandi, Lempel, and Leibovich (2013) looked at how interest in peripheral content, such as advertisements, varied based on interest in the primary content on the page. Participants (Mechanical Turk workers) were asked to read news articles until they felt they had discovered the purpose of an article, and then were instructed to answer questions based on the text. Results showed most participants missed the ads entirely; only a quarter of participants paid any attention to the ad image surrogates. Rohklenko et al. found that participants who spent a lot of time reading the content on a web page had higher recall for the advertisement images than participants who read less. If interest can serve as an indicator of engagement, then this study showed that engagement with content could lead to higher recall for peripheral images. This study also helps confirm that when participants are engaged they tend to display deeper information processing behaviors such as reading and absorbing more content. If engaged participants are able to recall many different types of information, then it is possible that engagement could lend itself to expanding attentional resources.

Song, Shi and Fu (2013) examined whether degraded search relevance had an effect on engagement. The researchers defined engagement both in terms of frequency of search engine reuse and behavioral signals. Participants in this study were given a search algorithm that provided low quality search results or received the normal search engine algorithm. Song et al. analyzed the session data of search logs from 2.2 million users. Query attributes such as queries issued per session, length, success, click-through rate, type, and session length as well as frequency of search engine usage were used to measure engagement. Song et al. found that though engagement decreased overall, there was some indication that participants might have been engaged. Participants in the treatment group issued more queries overall, issued more navigational queries, reformulated their queries more, and clicked on more results. They surmise that this search behavior could reflect increased effort, a consequence of struggling to complete the search with poor search results. This means that, for the engagement metrics defined in this study, engagement was initially negatively correlated with relevance. Song, Shi and Fu then tried to predict engagement using search actions, and found that number of clicks was the highest correlated feature with engagement. This study established a link between behavioral signals and engagement as induced by effort. In particular, effort invokes the factors of felt involvement and focused attention, which, as this study showed, can be induced by negative influences rather than positive ones.

Perhaps the most revealing studies are those that combine changes in both content and navigational structure. Chen, Lin, Yen, and Linn (2011) examined the effect of disorientation on engagement with a website given the breadth, familiarity, and media richness of the site. Two websites were created with different structures: the “broad” structure contained two levels, while the “deep” structure contained four levels. Familiar sites contained stationery products, while unfamiliar sites contained industrial products. Media richness was also manipulated; “media rich” sites contained images and videos, while “lean media” sites contained only text. Chen et al. found that participants preferred websites that had a deeper structure, and were more engaged with a site that had unfamiliar structure and lean media richness in addition to deeper structure. Higher disorientation was linked to less engagement and lower intentions to use the website in future. This study shows that engagement does not always occur when a participant is completely comfortable and familiar with a web interface. Rather, a combination of novelty and familiarity can foster engagement.

The structure and layout of a website and search interface is referred to as a representational context. Representational context includes the designer's decisions about how to represent actions that can be performed (e.g., search box, search button), the placement of elements and icons on a page and even the icons themselves. Subsequently, representational stability refers to the extent to which this representational context is maintained over the course of the entire search experience. Representational stability can be examined both within a single system and also across systems that are used to perform a similar function. For example, most major search systems employ interfaces that use a single box for query entry, and a rank ordering of search results. Duin and Archee (1997) posited that representational context must remain stable in order for the participant to become engaged.

Webster and Ahuja's research (2006) supports the relationship between engagement and representational stability by developing a model of disorientation and engagement in web systems. This model states that navigation systems affect perceived disorientation, which affects engagement, which affects both performance and future use. Engagement was operationalized as when a system "holds [a subject's] attention and they are attracted to it for their intrinsic rewards" (Jacques, Preece, & Carey, 1995, p. 58), and was measured with a seven-item questionnaire that contained items such as "the site kept me totally absorbed in the browsing" or "the site held my attention." To evaluate their model, Webster and Ahuja (2006) tested a simple navigation system against a global navigation system and an enhanced navigation system. The simple navigation system contained only hyperlinks, and these hyperlinks disappeared while the participant scrolled. The global navigation system contained a site map, a search form, and nested navigation bars (i.e., a parent topic contained child topics), but the navigational features also disappeared while scrolling. The enhanced global navigation system had the same features as the global one, but kept the features visible while scrolling. Their findings supported the model in that participants in both the global and enhanced global navigation conditions reported less disorientation, and participants in the enhanced global navigation system had the best performance. This group's high performance was also positively related to engagement, showing that navigational aspects of a search interface can affect engagement. Perceptions of navigation and orientation are shown to help maintain representational integrity, providing a link between engagement and usability.

This notion of stability is also supported by work done by Sundar, Xu, Bellur, Oh and Jia (2011), who looked at the effect of different user interface interaction modalities on engagement. Interaction modality refers to mouse-based interaction patterns, specifically zoom, drag, slide, mouse-over, cover flow, and click to download. Sundar et al. investigated these modalities on six artificial websites. Layout, page content, and color were kept constant between interaction modalities. These modalities allowed participants to access "hotspots" or links to information embedded in the website. Sundar et al. defined engagement as a combination of participant attitudes, actions, skill and behavior towards the content. They hypothesized different types of interaction modalities would lead to different levels of perceptual bandwidth, or the "range of sensory and preliminary attentional resources available to individuals" (p. 1478), referring to the resources a person has for understanding and perceiving interactivity in an interface; Sundar et al. defined this conceptually as "users memory for interface content" (p.1478). Reeves and Nass (2000) stated that perceptual bandwidth is increased by perceptual interfaces, which offer people "more and different sensory channels" (p.65) than traditional interfaces. This suggests that perceptual interfaces or increases in perceptual bandwidth can change interest in the content of an interface.

Perceptual bandwidth was measured in terms of recall and recognition, perceived interactivity, actions, behavioral intention towards content and the website, and attitudes towards content. Sundar et al. found significant differences between modalities, specifically, the slide modality showed higher recall than the zoom in/out modality. Participants who used the cover flow and mouse-over actions performed more actions overall than the other modality types. Some participants preferred modality types that gave them more control over their content, while others preferred modality types that allowed them to perform more actions. Sundar et al. remind us that interaction modalities can make content more absorbing and generate positive feelings, which are closely related to the interest and cognitive absorption that occurs during engagement. The distinct preference for modality among participants indicates that users want to maintain representational stability, though that representational stability may be subject to variation across individuals.

Sundar et al. collected attitude data and found that certain actions such as the mouse-over led to more positive attitudes than cover-flow, which led to more negative attitudes. This also shows that some interaction types are generally more preferable than others. Some users, referred to as "power users" (who were identified based on a questionnaire containing items about liking, skill, and dependence on technology) preferred modality types that gave them more control over their content, while other users who were not "power users" preferred modality types that allowed them to perform more actions, demonstrating the importance of individual differences. Other research has suggested that control is important in engagement (Webster & Ahuja, 2006), and this work showed that control might be more critical to engaging some users than others.

Teevan, Collins-Thompson, White, Dumais, and Kim’s (2013) findings challenge the notion of representational stability as necessary for engagement. In this study, Teevan, et al. studied one important structural element of search systems: latency. Latency refers to the interval between an action and the response. High latency can be thought of as disruptive to representational stability because it disrupts a person’s ability to maintain representational context. The purpose of this study was to examine how participants interacted with a search system that prioritized high quality results over speed. Specifically, Teevan et al. looked at querying behavior with navigational queries (those that “targeted specific web pages” (p.2)) and informational query types (those that are “intended to find information about a topic”(p.2)). The researchers also examined two post-query behaviors: abandonment rate and time to first click. Engagement was examined was defined as engagement with the search results in the form of more search interaction behaviors. Teevan, et al. (2013) found that click frequency decreased as page load times increased, which the authors claimed signaled a loss of interest. However, the results showed no increase in search abandonment (which is also posited as evidence of disengagement) as load times increased. They explain this by stating there is a point beyond which load times can increase without causing higher search abandonment rates. It is also possible that the clicking was more deliberate, as participants anticipated the page load times and wanted to be sure they clicked on the most fruitful result. Participants were asked how long they would be willing to wait if they knew search engines would give them the best response,versus an acceptable response, and most said they were willing to wait much longer for the best response. This indicated that participants may be able to tolerate shifts in their representational context and adapt to them if they receive some benefit.

Arapakis, Bai and Cambazoglu (2014) investigated the impact of response latency on the click behavior of participants, and the point at which response latency becomes noticeable in two studies. The first study looked at participants' sensitivity to latency, and used two manipulations: response latency and site speed. Response latency refers to the time between a user's action and the perception of the response. Site speed was operationalized as either slow (a search site with a slow response) or fast (a search site with a fast response). They found that participants were more likely to notice the response latency if it climbed above 1000ms. In the second study, they measured the effect of response latency on user engagement using the *focused attention* subscale of the UES (modified for a search context), satisfaction, and click behavior. They found a small effect for focused attention in participants in the fast condition, which they state suggests that these participants felt more deeply involved in the search task. They also found that though there were no significant differences in frustration, participants' positive search engine bias (the belief that the search system was helpful) was correlated with focused attention and perceived usability in both speed conditions. This suggests that search engine bias affects the way that participants interpret system response. Lastly, they found that participants were more likely to click on a result from a SERP that had been returned with low latency. This paper showed that conditions we may see as unfavorable to engagement (such as low latency) could encourage positive behaviors such as more examination of search results.

Work on engagement and search interfaces has shown that the interface can be crucial in fostering and maintaining engagement throughout the search session, and that altering the traditional search interface to include elements that allow users to reflect on their own behaviors, and compare them to others, can potentially improve user engagement. This body of research also shows that representational stability, while important to engagement, may be one facet where individual differences are important. The literature reviewed here shows that users can tolerate shifts in their representational contexts and that users can express preferences for different kinds of interaction.

2.4.6. Individual Differences and Engagement. Researchers studying engagement must be aware of the role of individual differences when considering the body of work on engagement and the potential for generalizing across users. Heinstrom (2006) looked at engagement through the lens of individual differences in information-seeking behavior due to personality traits, learning approaches, and discipline differences. Master’s students writing a thesis were chosen to participate, as it was assumed that they would be committed to completing an information-seeking task. Three questionnaires were used to discern the factors listed earlier: the NEO Five-Factor Inventory, which is a psychological measure that contains items about personality, the Approaches and Study Skills Inventory for Students, and a questionnaire about information-seeking behaviors. Three information-seeking patterns were discovered in this study: fast surfing, broad scanning, and deep diving. The behavior that emerged most closely related to engagement was deep diving. Heinstrom found that participants who exhibited “deep diving” (p. 1446) behaviors spent more time information-seeking and indicated that they preferred high quality documents. Heinstrom noted that they seemed “focused and structured” (p. 1446) in their searches and seemed to be searching to gain a thorough understanding of the topic rather than just scanning for information. However, some of the other information-seeking behaviors were linked to engagement with different kinds of materials: broad scanners seemed more engaged with documents that gave them new information, while fast surfers were more interested in documents that were easy to read and were less academically challenging. Heinstrom also noted that personality types may inspire engagement either through an “open eagerness to try new things” (p. 1448) or a “conscientious urge to strive” (p. 1448). Lastly, the study revealed that topical engagement seemed more likely to occur in relaxed settings because of their ability to provide low time pressure in tandem with high motivation.

What this work offers to our understanding of engagement and web search is that it is potentially highly context and topic dependent. Since the students in this study were completing master’s theses, there was an inherent interest in the topic that may have lent natural motivation to searches, no matter what information-seeking style they had. Topics or search tasks that are less inherently interesting will have an effect on how deeply engaged a user is with the task or topic at hand. What this also lends to the study of engagement and search is that personality and individual differences may play a bigger role than previously believed, and that information-seeking styles vary in their ability to promote engagement.

Hwang and Thorn (1995) stated that engagement is integral to system success, and the work reviewed in this chapter supports this idea. System success is a complex mix of attributes such as system response time, performance capabilities, and user experience, and the work done in this area has showed that engagement is also a complex interaction. Engagement is dependent on the structure and ease of use of the system, the performance of the system, the content within it, the complexity and difficulty of search tasks, and how users perceive all of these variables. Our role as researchers is to meet the needs of users by addressing all of these areas, and one way we can do this is by making sure to examine both the objective behavior of the user as well as their experience.

2.5. Frustration

2.5.1. Theories of Frustration. The study of frustration has followed similar patterns to the study of emotion and stress, characterized by both a psychological and physiological point of view. The benefit of these two approaches is that they provide a link between psychological and physical processes. From the psychological perspective, frustration can potentially be defined as a response to unfavorable outcomes. One example of this is the frustration-aggression hypothesis (Dollard et al., 1939), which states that frustration occurs when there is interference with expected attainment of a goal. However, there are other frustration theories, known in the psychological community as “integrated” theories, which orient frustration as a moderating variable in a larger set of emotional occurrences.

Amsel’s (1958) “frustrative nonreward theory” is an example of an integrated frustration theory. This theory states that frustration occurs when there is a psychological expectancy of reward that is not satisfied. There are four properties of frustration. The first two are primary frustration, which is an unconditioned response to a frustrating event, and the primary frustration drive stimulus, which is a stimulus that guides behavior. These two take place when the stimulus has not been encountered before. The second two are anticipatory frustration, which is frustration that occurs before a frustrating event occurs, and feedback stimulation from anticipatory frustration, which also occur before the frustrating event. When anticipatory and primary frustration occur frequently, a person develops a conditioned frustration response.

2.5.2. Frustration Responses. Amsel states that primary frustration is a "temporary state that results when a response is non-rewarded" (p. 1). This reward system is part of a paradigm crucial to frustration called reward-schedule effects, which states that learned behavior develops when a person experiences a series of rewards and non-rewards. The reward-schedule effects outlined by Amsel demonstrates that a frustrator should be unpredictable in order to be effective; participants need to experience periods of reward and nonreward in order to become frustrated instead of demotivated. A secondary (more learned) form of frustration result in four types of behaviors: invigoration, suppression, persistence, and regression. Persistence occurs when a person continues work on the task even though it is frustrating, and regression occurs when a person returns to an earlier "more successful" mode of behavior. Persistence is useful in search because it keeps participants engaged with the search process. However, there is a point at which the participant continues to persist even though they are experiencing frustration, and though the behavioral signals may appear the same, the underlying emotion is different. Invigoration is when a person renews their efforts during a frustrating task, while suppression is when a person attempts to suppress their frustration response. Invigoration is important because it is related to the initial engagement stage of O'Brien and Toms' (2008) model where increased general activity and goal-directed behavior occur. Invigoration is also responsible for what is known as the "frustration effect." The frustration effect, identified by Amsel and Roussel (1952) is characterized by increased response speed following a previously reinforced response. This means that when a person is not rewarded for one response, they experience increased vigor in their next response. In a search system where participants experience intermittent frustration, they may feel an increased need to perform more search actions in order to experience the “reward,” or the absence of a frustrator. Suppression and regression may manifest themselves in search as early abandonment or few search actions, as these states indicate the person is unable to cope with the “non-rewarded” response.

Understanding these response can help us identify what kinds of frustration responses participants experience and how this is reflected in their search behavior. Another response to frustration is demotivation (Seward, Pereboom, Butler & Jones, 1957), which is the opposite of the frustration effect. Demotivation is characterized by a temporary state of satiation or demotivation during an episode of nonreward. Amsel reconciles this with the frustration effect by stating that demotivation occurs after a person has not experienced appropriate levels of intermittent reward. In other work (Amsel & Roussel, 1952), the frustration effect was shown to appear and disappear once the participant had reached appropriate levels of frustration response. This demonstrates that inducing frustration must be done carefully to avoid unwanted effects and to properly qualify the measured response.

2.5.3. Methodological Implications of Frustration Induction. Research suggests that frustration and frustrating experiences have methodological value as impetuses for directed behavior. Though not operationalizing frustration specifically, researchers have created interfaces designed to make the participant work harder to achieve their goals, resulting in negative emotions. Morris, Morris and Venolia (2008) created an interface called SearchBar that presented search results hierarchically based on the search queries entered, and found that this interface was more beneficial over standard search with respect to allowing the user to re-find information. However, participants in the test condition often found it difficult to remember how to create tasks within the system, suggesting that there were some attributes of the system that were challenging. Still, participants in the experimental condition made more study entries when completing tasks, and found SearchBar helpful and easy to use, indicating that transforming user interaction with a system can create improvements in specific types of behavior.

Riche, Riche, Isenberg and Bezerianos (2010) posit that hard-to-use interfaces can be beneficial if the experienced frustration and effort is channeled effectively, citing a study examining frustration in collaborative spaces where individual frustration with the system led to increased collaboration and more complex group problem-solving dynamics. Cockburn, Kristensson, Alexander, and Zhai (2007) designed an effortful "frost-brushing" interface that resulted in increased spatial memory for visual location. However, participants in the test condition also experienced higher levels of frustration and mental demand and overall effort. The authors also found that subjective ratings of the effortful system were more positive than the standard interface, indicating that participants found some aspects of the more effortful interface enjoyable. This body of work shows us that while reducing frustration on the part of the user is a useful goal, frustrating experiences can be used as a tool to spur changes in behavior.

2.5.4. Physiological Indicators of Frustration. Prior work has shown patterns of physiological signals unique to frustrating experiences. Curiously, frustration and violence or aggression have been frequently paired when observing physiological signals. In an experiment that predates the establishment of the institutional review board for ethical treatment of human subjects, Freeman (1940) conditioned participants to experience frustration by delivering electric shocks whenever they failed to correctly indicate the presence of visual stimuli. He measured what he terms "palmar skin resistance," which we now know as skin conductance, and found increased palmar secretions during frustrating episodes. However, there were also increased palmar secretions during more difficult tasks (i.e., when the visual stimuli was more difficult to discriminate), so it is possible that participants were reacting to the difficulty of the task as well as frustration experienced. Gentry (1970) induced frustration in the form of interrupting their completion of a test, as well as personal insults by the experimenter, and found that both systolic and diastolic blood pressure increased overall. Doob and Kirshenbaum (1973) measured blood pressure and task performance under both frustrative and aggressive conditions. Participants were either frustrated or not frustrated via a Stroop-like task and were either shown or not shown a violent film. They found that participants who were both frustrated and shown a violent film had the highest readings of systolic blood pressure. This finding showed that aggression can compound frustration and increase the resulting physiological response.

Other studies have examined frustration and non-reward in more benign settings. Otis and Ley (1993) conducted an experiment where participants received a small monetary award if they pressed one lever with a certain amount of force, but no monetary reward if they pressed a second lever. Otis and Ley found that when they discontinued the reward response on the first lever, participants' level of skin conductance increased. In fact, there was a correlation between the magnitude of force the participant exerted on the lever and the level of skin conductance observed, suggesting that the effort involved in frustration can manifest itself in physiological response.

Still other work has used physiological signals to understand and predict frustration. Scheirer, Fernandez, Klein, and Picard (2002) linked physiological signals to frustration by measuring galvanic skin response (GSR) and blood volume pressure (BVP) while the participant experienced delayed mouse clicks. Participants were asked to complete a series of puzzles in a computer game as quickly as possible, creating a frustrating situation when the mouse clicks were delayed. Scheirer et al. found that they were able to discriminate between frustrated and non-frustrated states. They also found that some participants waited for the delay to finish and subsequently slowed their click behavior, while others experienced an increase in clicks. Kapoor, Burleson and Picard (2007) used skin conductance in addition to posture sensitivity, facial movement tracking, and mouse pressure signals to automatically predict frustration. They found that fidgets (meaning movements in the seat), velocity of head movements, and changes in posture were the features that were most predictive of frustration, with an accuracy of 79%. McLaughlin, Park, Chen, Zhu and Hoon (2004) used a pressure-sensitive touchpad to gather haptic (touch) data from which to predict frustration. In a pilot experiment, they used nine participants who completed a visual task. Though they failed to accurately predict frustration, they suggest that effective signal isolation techniques as well as combining physiological signals could lead to greater accuracy. Grafsgaard et al. (2013) found that facial dimpling and brow lowering were positively correlated with frustration and learning during an online learning interaction, and that frustration could often be detected in the first five minutes of the session.

2.5.4.1. Reduction of Physiological and Emotional Symptoms of Frustration. There has been some treatment in the literature of ways to mitigate frustrating episodes. Klein, Moon and Picard (1999) created a system to frustrate the participant as well as an agent that responded affectively to this frustration as a means of preventing any negative emotions. They found that participants who were in the frustrating condition and received support from the agent spent more time using the system. They also found participants who were allowed to vent reported less frustration than those who were completely ignored. This led the researchers to conclude that designing a system that appears interested in participants' emotions without actually responding to them could play a role in diminishing negative emotions.

Amershi and Morris (2008) created a system for collaborative searching, CoSearch, that was designed to improve the user experience and reduce frustration. Specifically, they wanted to reduce frustrations inherent to collaborative searching such as poor division of labor and ability to control aspects of the interaction. Participants completed tasks in three conditions: the shared (where participants used a single computer), parallel (where participants each used their own computer) and the CoSearch condition (where one person used a computer with CoSearch, and two people used CoSearch on their mobile phones). CoSearch offered the option of searching on the same page while using distinguishing features such as different colored cursors and different colored tabbed areas where the collaborator could share details about their search results. While most people expressed a (non-significant) preference for parallel search, participants felt most frustrated in the shared condition, and felt least frustrated in the parallel search condition, though there were no differences between participants who used mobile and computer CoSearch. They also found that participants collaborated more effectively in the Shared and CoSearch conditions. The most important contribution of this work is its demonstration that participants prefer more control in their search interaction, and that control can help reduce markers of frustration.

One concept has been of particular interest: non-conscious reappraisal. Non-conscious reappraisal, in contrast to cognitive appraisal (covered earlier in this review) is a re-assessment of emotional state that occurs unconsciously. Yuan, Ding, Liu and Yang (2014) conducted a study to look at the effects of non-conscious reappraisal on physiological signals during frustrating episodes. Participants were divided into three groups: the conscious reappraisal group, the non-conscious reappraisal group, and the control group. Participants in each group were given a complex and difficult arithmetic task with correct and incorrect feedback in order to induce frustration. Those in the conscious reappraisal group were given instructions to regulate any potential negative emotions, and those in the non-conscious reappraisal group were primed with reappraisal words (i.e., analyze, and think). They found that conscious reappraisal lowered ratings of frustration (as measured by the PANAS), while non-conscious reappraisal had no effect. However, they also found that both conscious and non-conscious reappraisal significantly lowered heart rate during frustrating episodes, indicating that either type is useful for reducing the physiological effects of frustration. One of the most interesting contributions of this study is its reinforcement of the idea that subjective experience alone can often be an unreliable measure of emotional state, and that there can be large discrepancies between reported emotion and actual experience. It also showed that while cognition is a crucial component of regulating emotional response, unconscious regulation is also possible and effective as well.

2.5.5. Frustration in Interactive Information Retrieval. Frustration in information retrieval is generally defined as the impediment of search progress. In contrast to engagement, it has incorporated much more study of the subjective experience of the user, perhaps because researchers tend to categorize frustration as an exclusively emotional state. This characterization is reflected in the way authors operationalize, study, and describe frustration. Hertzum (2010) investigated differences in the severity of frustration as well as how much time is lost to frustration, and whether it is likely to occur again later given prior frustrating experiences. Hertzum operationalized frustration similarly to frustration theorists: as a symptom of “unattained goals” (p. 1). Participants were asked to perform a search session using their own computer, and to report if they had a frustrating experience. Hertzum found that 27% of the search session was “lost” to or taken up by frustrating experiences. However, level of frustration was not correlated with length of frustrating experience. Instead, there was a correlation between importance of task and length of the frustrating experience, indicating that longer frustrating experiences happened during more important tasks. This study shows that frustration and goal-orientation and motivation are closely tied together as they are in engagement.

Hoppmann (2009) used think-aloud protocols to explore frustrating events leading up to an abandoned search. Participants were asked to search a website for items useful to the places they worked, and to stop once they felt they came to a satisfying result. Hoppmann found that participants experienced frustration when they were given search result lists that were long and not sorted in order of importance. Hoppmann also found that frustration was more likely to occur after the participant had identified an obstacle or if they found results or links unsatisfying in that they did not lead to a successful result. Participants who expressed positive feelings towards information found online also reported positive feelings during the search process, and negative feelings towards information found online were correlated with greater frustration. This study showed that attitudes and individual interaction with a search system are good indicators of frustration.

Behavioral signals are also used to predict frustration and search engine switching. Feild, Allan and Jones (2010) found that users generally reported frustration for half of their queries, and the top five reasons for frustration were: off-topic results, more effort than expected, results that were too general, un-corroborated answers, and seemingly non-existent answers. Feild et al. found that task duration and average task URL count were the most important features for predicting frustration, specifically that a “lengthy session with few URLs visited” is ideal for predicting frustration. This work is consistent with O’Brien and Lebow’s (2013) finding that engaged participants perform fewer behaviors, but overall this study suggests that while behavioral signals have an important place in determining frustration, they can be misinterpreted without contextualization. White and Dumais (2009) found that people switch search engines because they are frustrated roughly ten percent of the time, and for sessions with three or more queries, time spent in the search session, actions performed in session, and number of pages visited in a session can be used to predict search engine switching.

Though not always an indicator of frustration, search failures often precede a frustrating experience. The literature on search failures can offer some insight into the kinds of search behavior participants may exhibit during frustrating episodes as they employ coping strategies. For example, Mansourian (2008) looked at how users cope with information-seeking failure on the web through examination of users “coping” strategies. Mansourian separates coping strategies into passive and active strategies, where passive strategies include less search interaction and active strategies include more search interaction. Some passive strategies included giving up or goal modification, where participants failed to find what they were looking for and instead changed their initial goal to match the results they have retrieved. Active strategies included revising search queries, shifting search tools, narrowing the domain, or switching mediums (from web to print, for example). Participants also sought help in the form of asking someone for advice or asking someone else to carry out the search, or postponed the search as a way of overcoming failure. These coping behaviors are directly tied to the demotivation process and frustration effect outlined by Amsel and Roussel (1952); participants feel thwarted in their goals, and so either pursue them with renewed vigor or abandon them entirely.

Sun and Spears (2011) offer some insight into the attribution of frustration during e-commerce search behavior. Sun and Spears conducted a critical incident survey where they asked participants about frustrating search experiences (particularly their keywords) and how they responded to less than satisfying search experiences. They then applied frustration theory to the results of this survey. They found that when participants primarily wanted to find relevant search results (i.e., they were not primarily trying to save time), they attributed their frustration to poor keywords. When they were trying to save time and experienced a frustrating search, they attributed their frustration the search engine itself. Sun and Spears also found that when participants had a primary goal of relevance, they addressed their frustration by altering their thoughts and behaviors about the search. When participants had a primary goal of saving time, they addressed their frustration by either abandoning their search or blaming of the search system. This study is interesting because it recalls many theories of emotion that state that emotional response is governed by attribution of emotion and cognitions surrounding that emotion. This has strong implications for my work because it again demonstrates the prominent role of goal orientation in frustration and in the subsequent behavioral response.

2.6. Disambiguating Engagement and Frustration

Physiological signals seem to hold promise for disambiguating engagement and frustration. Grafsgaard, Wiggins, Boyer, Wiebe, and Lester (2013) investigated the usefulness of facial expression analysis in understanding the affective states of engagement and frustration within a learning context. Students were given a modified version of the *endurability* scale of the UES, as well as questions about temporal demand, performance, and frustration portions from the NASA-TLX. They found that *endurability* was predicted by inner brow raising, while temporal demand was predicted by outer brow raising. They also found that performance was predicted by mouth dimpling, and frustration was predicted by brow lowering. The major contribution of this paper is its connection of facial expressions to reliable measures of engagement and frustration. However, they measured frustration using one item on the NASA-TLX, which (on its own) is not a strong indicator of frustration.

Vail, Grafsgaard, Wiggins, Boyer, Wiebe and Lester (2013) examined the utility of one of the Big Five personality traits (introversion and extraversion) in conjunction with facial and postural gestures as predictors of engagement and frustration during an online tutoring session. They found that predictive models relied heavily on dialogue (speaking with the tutor) for extraverted participants, and in particular, engagement and learning gains were positively and negatively affected by dialogue. However, frustration was more often correlated with changes in posture and seat movement. For introverts, engagement was correlated with forward postural movements, while frustration was correlated with backward postural movements. This indicates that introverts may express their feelings behaviorally rather than with dialogue. This study adds an interesting multimodal analysis of behavioral data and affective state, but they measure frustration using only one frustration item, while a more comprehensive scale could have offered more depth.

Müller and Fritz (2015) used different physiological signals to determine whether software engineers were “stuck in flow” or “frustrated and happy” (p. 1). Specifically, they wanted to observe the range of developers’ emotions and how their tasks affected their emotions. They collected electrodermal, heart rate, fixations and EEG data, and interrupted participants when they had been working for five minutes. This interruption was used to ask the participant how they felt at that moment and how far they felt they had progressed on the task. Lastly, they conducted an interview at the end of the experiment. They found a significant effect of arousal and valence (meaning the valence of the emotion) on estimates of progress, though the correlation between arousal and progress was very weak. This means that participants may not have necessarily felt a strong physiological reaction when they reported feeling blocked on their progress, but were able to report a strong valence of that emotion (usually negative) when they felt blocked. Participants reported feeling negative emotions when their progress was blocked on a task, or when they were unsure of how to do something. Müller and Fritz were also able to create a model that was able to predict emotions using physiological data with 71% accuracy. This study demonstrated that combinations of physiological signals can be useful in predicting emotional states, particularly in the scope of task completion. However, it is important to consider the methodological implications of interrupting a participant to gather information on their emotional state, which may be altered based on that interruption.

2.7. Physiological Signals and their Role in IR Evaluation

2.7.1. Electrodermal Activity. Electrodermal activity refers to different types of electrical activity from the skin. The description of the processes of the skin and glands below come from Boucsein (2012)’s guide to understanding electrodermal activity. Skin is made up of several layers, but the most important layers for electrical activity are the epidermis, dermis and the hypodermis. The epidermis is made up of five layers and contains the stratum corneum, the outermost layer where most electrodermal activity is measured. The dermis is thicker, made up of dense collagen fibers, and contains the part of the sweat glands through which sweat is actually expressed. The hypodermis, made up of connective tissue, is below the dermis and contains blood vessels, nerve endings as well as the part of sweat glands that control secretion. Eccrine sweat glands are located in various places in the body and function primarily as a method of thermoregulation, or cooling the body if it overheats. The nerve endings around these sweat glands are connected to the same autonomic nervous system processes responsible for expressions of emotion. Emotional response is directed through a neural activity loop called the Papez circuit (Papez, 1937) where emotions are generated, checked against sensory information and against the hypothalamus as a means of deciding whether to inhibit or allow behavioral response. Thus, different kinds of emotions may produce the same levels of general arousal and need to be differentiated at the skin level. Glands on the palms and soles of feet produce sweat mainly as a result of emotional activity, while the forehead and other sites produce sweat due to regulating body temperature as well as emotional episodes (Boucsein, 2012). However, there is some dispute about this; two studies have found no difference in skin conductance levels between the forehead, palm, and sole regions (Conklin, 1951; Rickles & Day, 1968).

Electrical skin activity measurement consists of two methods: exosomatic and endosomatic (Boucsein, 2012). Endosomatic activity refers to differences within the skin itself, while exosomatic activity refers to differences measured when current is passed through the skin for measurement. Exosomatic measurement can refer to skin conductance, resistance, admittance, and skin impedance levels and responses. As sweat levels change, skin conductance (the ability of the electrical signal to pass through the electrode site) levels change accordingly. These changes can be grouped into either short-term changes, which are called phasic changes, or long-term changes, which are called tonic changes.

There are several important steps in the measurement of electrodermal activity. First, there is electrode placement, which is largely dependent on the kind of skin activity one is trying to measure as well as study design. Boucsein recommends the placement of electrodes on the palms or fingers, specifically on the distal phalanx of the fingers (fingertips) rather than the medial or proximal phalanx of the fingers (middle or lower sections) as studies have found higher levels of skin conductance response on the distal regions. Boucsein recommends the sides of the palm closest to the thumb or the pinky fingers (the thenar and hypothenar eminences, respectively), instead of the center of the palm because of the potential for unplanned and disruptive movement. Roth, Dawson and Filion (2012) also recommend abrading either site to eliminate undue skin potential from the surface of the skin. However, they caution against washing the site with soap or abrading the surface of the skin with any solutions containing more than 70 percent alcohol because of the potential for drying out the skin and thus reducing conductivity. Electrodermal activity is extremely susceptible to changes independent of the intended stimulus, leading to potentially noisy data if the environment has not been properly insulated from noise. For this reason, many researchers limit EDA measurement to short periods of activity in controlled environments with well-defined stimuli.

2.7.2. Electrocardiography. Similarly to many physiological signals, the autonomic nervous system (ANS) is crucial to regulation of the heartbeat. The sympathetic branch of the ANS, which is responsible for supporting the fight-or-flight response, increases heart rate, while the parasympathetic branch, slows down heart rate. These two systems constantly work in tandem to regulate heart rate. Measuring changes in heartbeat as an indicator of emotion usually takes one of three forms: cardiac output, heart rate variability, and electrocardiography. Cardiac output is merely a count of the measure of heartbeats, which can be compared over a measure of time. Heart rate variability is measured as a change in the interval between beats of the heart. This interval is referred to as the interbeat interval, and is measured in R waves, expressed in milliseconds. Levels of heart rate variability vary both between individuals and within individuals; illness, age, and regular exercise levels all affect a person's heartbeats. These beats are regulated by the sinoatrial and atrioventricular nodes. The sinoatrial node is composed of specialized cells that have high "intrinsic frequency" (p. 7), control autorhythmic response and are responsible for sending an electrical signal to begin the pumping action of the heart. This signal is then passed to the atrioventricular node which sends the current to the left and right ventricles of the heart. The cells in the atrioventricular node delay the current that controls each ventricular contraction by 100ms, which allows the atria to contract before the ventricle. Electrocardiography (ECG) refers to the measurement of this electrical activity. Electrodes placed on the chest measure the electrical activity generated during the contraction of the atria and left and right ventricles (Barber, Brown, & Smallwood, 1984).

Heart rate variability can be calculated in a number of ways. Malik, Bigger, Kleiger, Malliani, Moss, and Schwartz (1996) summarize different methods of measuring heart rate, and divide them into time domain measures, frequency domain measures, and rhythm pattern analysis. Time domain measures refer to measuring intervals between heartbeats (referred to as normal-to-normal (NN) intervals) or measuring the heart rate at a point in time. A common calculation to perform using NN intervals is the standard deviation of the NN interval (SDNN), which gives us the variability in NN intervals over a period of recording. However, the SDNN is only useful when the recording period is less than or equal to five minutes, given that the longer the recording lasts the more variability it will produce. To address the issue of variability, the standard deviation of the average NN interval (SDANN), can be used on five-minute segments of the total recording to calculate the change in variability of intervals that last longer than five minutes. There is also the heart rate variability (HRV) triangular index (considered a geometric measure versus a time measure), which measures overall variability in heart rate. Within the frequency domain, there are different types of spectral analyses that can be performed to create a better picture of the changes in NN intervals. Rhythm pattern analysis involves looking at the cycle length of beats, mainly the "oscillation" between increasing and decreasing heart rate.

The measurement of electrocardiography comes with recommendations as well (Kligfield et al., 2007). The American College of Cardiology (ACC) recommends a 12-point lead setup consisting of three leads, three augmented limb leads, and six precordial leads. In most cardiology studies, or studies in which the researchers are examining ECG for heart abnormalities, six leads are attached to the wrists and ankles, while six leads are placed on the chest in a specific pattern. Participants are usually asked to lie on their backs as the measurement is recorded. Studies have shown that heart rate can be greatly affected by changes in emotion (Malik, Bigger, Camm, Kleiger, Malliani, Moss, & Schwartz, 1996; Sakuragi, Sugiyama, & Takeuchi, 2002). Appelhans and Luecken (2006) state that higher heart rate variability reflects a "greater capacity for regulated emotional responses" (p. 235), and that if emotional regulation is an ongoing process, then changes among within-subject recordings of resting heart rate may be a good indication of emotional regulation. Sakuragi, Sugiyama, and Takeuchi (2002) found that laughter caused strong but transient changes in autonomic nervous system response (which regulates heart rate), while crying had weaker but more prolonged effects.

2.7.3. Physiological Measures in Information Retrieval Studies. Researchers in information retrieval have used different types of physiological measurement to capture different kinds of data. Facial electromyography (measurement of the movement of the muscles in the face) and facial expressions have been used as an indicator of positive and negative affect (Partala & Surakka, 2004), dimensions of emotion (Gilroy, Cavazza, & Vervondel, 2011), and to improve systems in terms by inferring relevance (Arapakis, Konstas, & Jose, 2009; Arapakis, Konstas, Jose & Kompatsiaris, 2009; Arapakis, Athanasakos, & Jose, 2010) and recommendation (Arapakis, Moshfeghi, Joho, Ren, Hannah, & Jose, 2009). Myography has also been used to measure hand gestures (Saponas, Tan, Morris & Balakrishnan, 2008), along with mouse movements and pressure-sensitive keyboards as indicators of stress (Epp, Lippold & Mandryk, 2011; Sun, Paredes, & Canny, 2014; Hernandez, Paredes, Roseway & Czerwinski, 2014). Posture-sensitive chairs have been used to measure body posture as an indicator of emotional state (De Silva, Kleinsmith, & Bianchi-Berthouze, 2005). Pupil diameter and eye movement have been measured using eye-trackers or similar devices as an indicator of emotional state (Ren, Barreto, Gao, & Adjouadi, 2013; Cole, Gwizdka, & Belkin, 2011). Electroencephalography (EEG), the measurement of electrical brain activity, has been measured using sensors as an indicator of preference (Ellick, Mirza-Babei, Wood, Smith, & Nacke, 2013), emotional response (Moshfeghi & Jose, 2013) and engagement (Szafir & Mutlu, 2012).

Skin conductance has been used in many different studies as indicators of slightly different phenomena. Galvanic skin response has been used as an indicator of cognitive load (Nourbakhsh, Wang, Chen, & Calvo, 2012; Solovey, Zec, Garcia Perez, Reimer, & Mehler, 2014), responses to interface changes (Pan, Chang, Himmetoglu, Moon, Hazelton, MacLean, & Croft, 2011), and stress (Mooney, Scully, Jones, & Smeaton, 2006). Heart rate monitoring is usually done via monitors similar to those used in health studies. Some studies have monitored heart rate directly by monitoring heartbeats (Anttonnen & Surakka, 2005; Magielse & Markopoulous, 2009). Others have done this by extracting heartbeat from ECG signal (Cai, Liu, & Hao, 2009). Still other researchers have created their own wearable sensors to capture these same types of data (Fletcher et al., 2010) in addition to affectively intelligent interfaces (McDuff, Karlson, Kapoor, Roseway, & Czerwinski, 2012).

To better illuminate what relationship emotions have to these attributes, researchers in this area have combined physiological, behavioral, and other affective signals for feature extraction via machine learning, and self-reports of emotions (Arapakis et al., 2009). Affective signals have been used to improve systems by attempting to predict relevance as well as offer recommendations. Arapakis, Jose, and Gray (2008) linked facial expressions with emotions experienced during search tasks. Participants were asked to assess the difficulty, complexity, and ambiguity of three search tasks, as well as to report levels of difficulty, interest, and fatigue experienced. Using these subjective reports in combination with facial expression analysis, the researchers were able to link unpleasant emotions such as irritation and anxiety to perceptions of difficulty and feelings of fatigue experienced during a task. Facial expression analysis also proved useful for prediction.

Arapakis, Konstas, and Jose (2009) combined physiological measures and facial expression analysis with machine learning to predict which documents or video snippets would be relevant. They extracted facial expressions, heart rate, galvanic skin response, and temperature, and asked participants to view documents and videos for four search tasks and mark whether they were topically relevant or not. The authors found their models performed better for video content, suggesting that audio-visual stimuli is more emotionally-laden than text, but they acknowledged that stronger emotional reactions (based on facial and physiological data) could have been elicited by the content itself, rather than the fact the participant marked the item as relevant. Later, Arapakis, Athanasakos and Jose (2010) compared personalized versus general affective models in terms of their ability to successfully predict topical relevance, and found personalized affective models overall successfully predicted relevance significantly better than general models.

Arapakis, Konstas, and Jose (2009) identified a problem with physiological data; it is often noisy, as signals can be elicited easily but are often difficult to interpret. This is where self-reporting of emotions can be helpful; interpretation can be elicited from the user (though this too can be problematic at times). This study also demonstrated a problem with using a machine learning approach to build a model with physiological data: since the data is noisy, it is difficult to extract distinct features that perform well. Most of the features extracted using the two different machine learning approaches (K-Nearest Neighbor and Support Vector Machine) did not exceed the baseline. Moshfeghi and Jose (2013) successfully combined facial expression, heart rate, galvanic skin response, and EEG signals with dwell time (a behavioral signal) to predict relevance. Again, using a machine-learning approach for feature classification, they found that behavioral and physiological data could act as reliable signals for relevance.

Niu, Zhao, Zhu and Li (2013) took a novel approach to video recommendation by identifying the emotional state of videos. They used automated computing of the affective signal (meaning the overall emotion or emotional tone) of the video based on attributes such as frame rate change, audio pitch, and motion. These attributes were used to cluster videos with similar affective signal, and could be used effectively in a recommendation system. Essentially, if a system detected happiness in a participant, then it would recommend a video with a happy affective signal. The research in this area is progressing towards systems that are continually able to monitor a participant's affective state and tailor their experience accordingly. Given that the work done in this domain is heavily based on visual stimuli, it is possible that affective signals would not be very useful in text-based recommendation systems.

Identifying emotional states can help pinpoint moments of stress. Lazarus and Folkman (1984) define psychological stress in terms of a coping response to difficult situations that produce negative emotions. However, though emotions have received some treatment in the information retrieval literature, the stress response has not. To understand stress fully as a phenomenon, linking self-report data to physiological measurements can give us a complete picture both of what the person is able to self-report as well as the stress they are unable to self-report. This essentially serves as confirmation that a stressful stimulus was present. Physiological measurement is not without its problems, however. Though the signals and experimental constructions are different, what is clear from the work done in this area is that the particular signal must be chosen carefully and appropriately for the context in which it is studied. The data is still extremely susceptible to noise, as a cough or sneeze can produce false (or stimulus-independent) changes in both skin conductance and heart rate. The key to measurement of this kind effectively is testing and defining a stimulus and creating ideal conditions to measure it, and it may serve as a powerful indicator of stress. There is also the normalization factor; data that has not been normalized often shows unclear trends and is non-significant as compared with normalized data, which can show clear stimulus-linked patterns.

2.7.4. Problems and Recommendations for Measurement. Physiological measurement has been employed more readily within the physical science fields and psychology, but more recently in affective computing and human-computer interaction fields. The measurement of these data in the context of computing is difficult, and guidelines on measurement are borrowed from other fields. Fairclough (2009) developed a catalog of potential issues with measuring and generalizing physiological signals. Of particular interest to the field is validity with regards to linking physiological data to events. Fairclough recommends choosing physiological variables that have already been linked to psychological constructs to preserve content validity. Also, it is recommended that the experimenter should test the validity of the system being used to measure these signals, to ensure that when the study protocol is put into place stimuli are measured correctly. To preserve concurrent validity (meaning when the stimulus can be used to predict a certain response) the researcher should gather an appropriate range of test conditions and participants with wide variation in individual differences.

Fairclough correctly points out the problem of specificity even within appropriate and well-designed test conditions. The problem of specificity in physiological measurement is that though a physiological response may be expressed during a task, the response produced during the task may not have been directly caused by the task. Thus, it is difficult to make inferences about physiological responses without correctly isolating a task and stimulus pair within an experiment. Lastly, Fairclough raises the issue of the difficulty in linking physiological signals to specific, exclusionary emotional states, especially since emotions change frequently and can be difficult to characterize to the exclusion of other emotions. Fairclough’s example is systolic blood pressure rising with the presence of frustration and anger, but as emotion theorists have pointed out, frustrating feelings can also be present in positive experiences of stress (eustress), so it is difficult to tell whether this frustration is a negative emotion or not. Fairclough’s review of the difficulties in physiological computing is a helpful guide for researchers to check their assumptions and protocols when designing studies that incorporate physiological measurement.

Physiological measurement is almost always used to characterize and measure emotion and emotional shifts, but there are many contexts in HCI where emotion becomes useful. Dirican and Gokturk (2011) reviewed the literature on physiological measurements and psychological processes, as well as the sensitivity of each signal. Dirican and Gokturk state that even though galvanic skin response is linearly correlated with arousal and is less sensitive to noise than ECG data, it has poor temporal sensitivity. Conversely, heart rate data is very sensitive to cognitive demands and attention as well as mental workload, but it is difficult to interpret because it shifts frequently. A focus of physiological measurement studies in human-computer interaction has been evaluating the usability of web search, computer systems, and video games. The common theme among all of these studies is the ability to map physiological signals to specific events. Ward, Marsden, Cahill and Johnson

(2002) looked at physiological indicators of arousal to both poorly and well designed web interfaces. More specifically, they use blood volume pressure and skin conductance while participants used a “poor interface” which had “impoverished navigational cues,” lots of animation, and an excessive number of pull-down lists, which deliberately obscured links and directory structure. Ward et al. (2002) found that overall the differences between groups were not statistically significant. However, when they completed an event-by-event analysis the data showed that skin conductance changed during specific moments during the task. Specifically, they found that levels of skin conductance spiked during the appearance of advertisements in the poor interface condition, of which there were 93. This study is useful mainly because it shows the problems with generalizing physiological data (there were large group variances which meant that the means were difficult to compare) as well as its ability to be tied to event-specific stimuli.

CHAPTER III: METHODS

This chapter details the definitions of the constructs in this study as well as how they were measured. This chapter also details the experimental setup, experimental procedure and discussion of the participants in this experiment.

3.1. Experimental Setup

A laboratory experiment was conducted with an independent variable of task interest and a moderating variable of frustration. Task interest was defined as level of interest in the task as determined by rankings of interest in the task completed before the experiment began. Participants completed a total of four search tasks: the two tasks they ranked most interesting and the two tasks they ranked least interesting. Thus, task interest was operationalized as a within-subjects variable with two levels: yes and no. Frustration was also a within-subjects variable that had two levels: present and absent. The frustrator was operationalized as poor search result quality, which took the form of search results which began at the 500th rank. The frustrator was delivered in counterbalanced order such that it appeared in all task positions an equal number of times. An experimental cover story (see section 3.5) was employed in order to prevent participants from becoming suspicious of the frustrator.

3.1.2 Pilot Tests. The experimental protocol was solidified through pilot tests with eight participants. Though there were only four participants who received poor results (i.e., poorly ranked results), there were differences in questionnaire scores that I believed would be more pronounced in a larger sample size. Given these results, we conducted a power analysis using a larger expected effect size. Results from the pilot testing are included in their corresponding section

3.2. Search Task Development

The tasks used in this study were based on tasks created and evaluated in an earlier study carried out with my colleagues (Kelly et al., 2015). These tasks were modeled after the cognitive complexity dimensions of Anderson and Krathwohl (2001)'s taxonomy of learning. In this study, cognitively complex tasks with five levels were designed and evaluated: *remember, understand, analyze, evaluate*, and *create*. These tasks also spanned four domains: health, entertainment, science and technology, and commerce. Forty-eight participants evaluated these tasks for difficulty, engagement, and their search behavior was recorded. There was a significant difference in engagement ratings between tasks, such that participants rated *evaluate* and *create* tasks as more engaging than *remember* tasks. Even though there were no significant differences in pre-search ratings of interest between tasks, post-search, participants rated *evaluate* and *create* tasks as significantly more interesting than remember tasks. This paper showed that participants who completed the *create* and *evaluate* tasks had significantly higher ratings of engagement via subscales of the User Engagement Scale (UES) and also searched for longer periods of time. Levels of interest also increased significantly for *create* and *evaluate* tasks over the course of the search.

Therefore, the *evaluate* task offered a useful structure for creating new tasks as it has proven engaging to participants. Only one task type was selected so as not to introduce another experimental variable. The *evaluate* task type required participants to search for information in order to evaluate several options, make a selection based on this information, and then justify the selection. A template was created in order to capture this process (see

Figure 2). The four *evaluate* tasks from our initial study were incorporated and slightly modified, and then four additional tasks were created using the template for a total of eight tasks (see Appendix A).

> Task Template: Compare and contrast two concepts and arrive at a decision. What are current methods of x and how effective are they? Which type of x do you think is best? Why?

Figure 2. Task Template Used to Create Evaluate Tasks

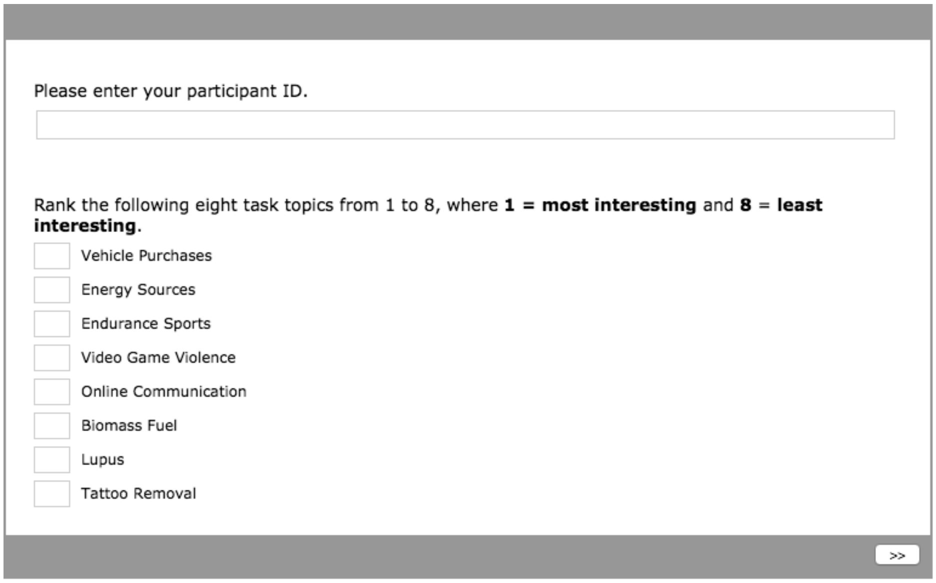
These tasks also incorporated four domains: health, science, technology, and entertainment (see Figure 3), as done in the previous study (Kelly et al., 2015).

> You're working on an assignment for an environmental science class about different kinds of energy sources and their efficacy. Your essay compares nuclear and solar energy. Which one is most cost-efficient to produce? How do different types of energies compare with regards to environmental impact? Which type of energy do you think is better? Why?

Figure 3. Example Task from the Science Domain

3.3. Task Interest

Other work has shown that interest is related to engagement (Csikszentmihalyi, 2014), and this experiment posited that participants would experience greater engagement with tasks they were interested in. The development of the search tasks was done with careful consideration of participant interest; the domains were diversified and topically relevant, though non-polarizing, topics were chosen to appeal to the desired population (undergraduates between the ages of 18-24). Participants were given a list of the eight task topics the day before their participation in the study (see Figure 4) and were asked to rank them. The rankings were organized such that 1 = most interesting and 8 = least interesting. They were then given their two most interesting and least interesting tasks to complete during the study. This was done to properly investigate the relationship between engagement, task interest, and frustration, which could be accomplished most effectively by observing behavior during two high-interest and two low-interest tasks. In pilot tests, participants reported higher ratings of interest for tasks they ranked more interesting, confirming the success of the experimental manipulation.



Please enter your participant ID.

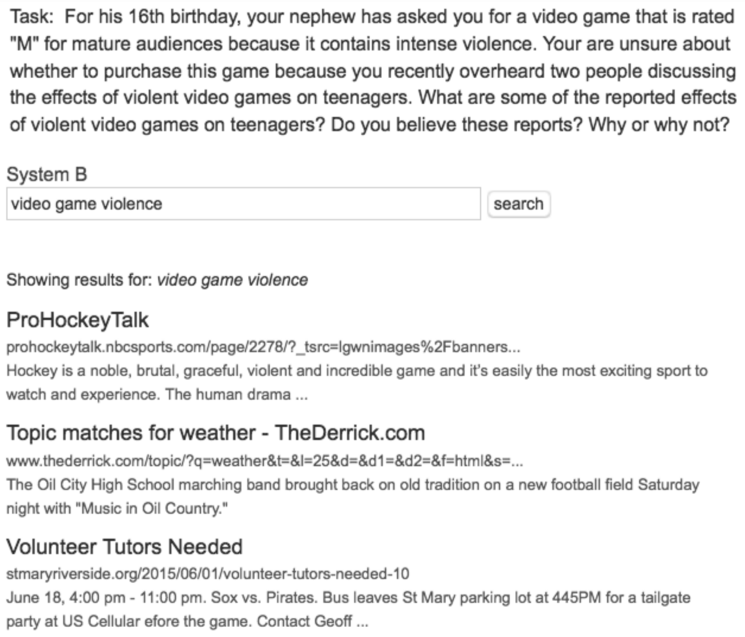
Rank the following eight task topics from 1 to 8, where **1 = most interesting** and **8 = least interesting**.

□ Vehicle Purchases
□ Energy Sources
□ Endurance Sports
□ Video Game Violence
□ Online Communication
□ Biomass Fuel
□ Lupus
□ Tattoo Removal

Figure 4. Ranking Questionnaire with Task Topics Presented to Participants

3.4. Frustrator

The method of inducing frustration in this experiment was operationalized as poor quality search results retrieval. This was accomplished by modifying the source code of the experimental search system to change the display of the search results. The search system was designed to provide results from the open web, and for tasks during which a frustrator was delivered, the results were reversed and presented from the 500th rank, i.e., the first result was rank 500, the second was rank 501, etc. This ensured that retrieved results were appropriately irrelevant (see Figure 5).



Task: For his 16th birthday, your nephew has asked you for a video game that is rated "M" for mature audiences because it contains intense violence. Your are unsure about whether to purchase this game because you recently overheard two people discussing the effects of violent video games on teenagers. What are some of the reported effects of violent video games on teenagers? Do you believe these reports? Why or why not?

System B

video game violence

Showing results for: *video game violence*

ProHockeyTalk
prohockeytalk.nbcsports.com/page/2278/?_tsrc=lgwnimages%2Fbanners...
Hockey is a noble, brutal, graceful, violent and incredible game and it's easily the most exciting sport to watch and experience. The human drama ...

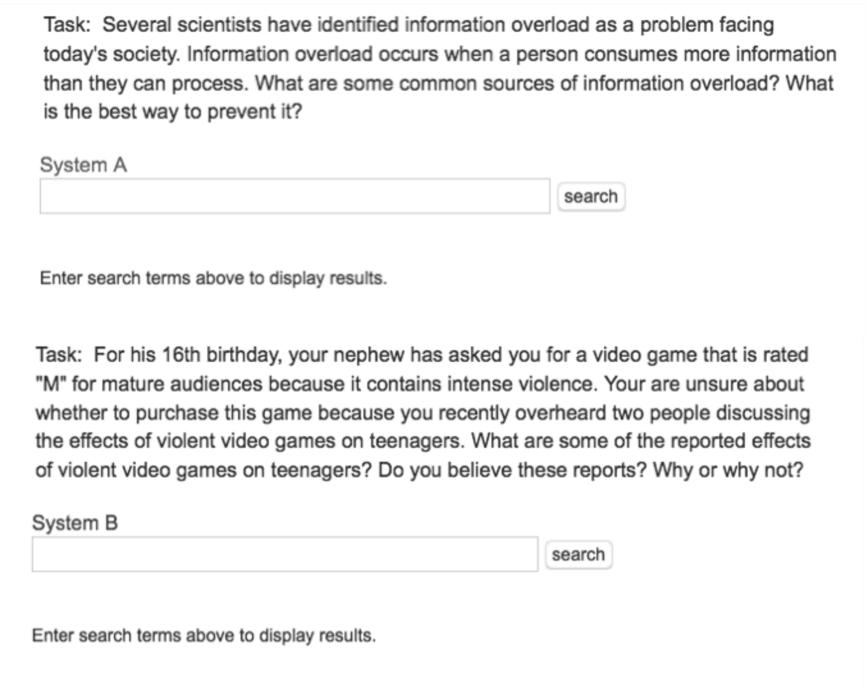
Topic matches for weather - TheDerrick.com
www.thederrick.com/topic/?q=weather&t=&l=25&d=&d1=&d2=&f=html&s=...
The Oil City High School marching band brought back on old tradition on a new football field Saturday night with "Music in Oil Country."

Volunteer Tutors Needed
stmaryriverside.org/2015/06/01/volunteer-tutors-needed-10
June 18, 4:00 pm - 11:00 pm. Sox vs. Pirates. Bus leaves St Mary parking lot at 445PM for a tailgate party at US Cellular efore the game. Contact Geoff ...

Figure 5. Example of Poor Quality Search Results

3.5. Cover Story

A cover story was necessary in order to keep participants from becoming suspicious of poor search results. The cover story used in the experiment involved telling participants that they were evaluating four different systems. This cover story was chosen because it was surmised that if participants thought they were evaluating different systems (rather than the same system with one modification) they would have greater expectations regarding search result quality, and would become more frustrated when they experienced poor result quality. Each “system” was labeled above the query box with either A, B, C, or D, in different colors, to connote difference to the participant (see Figure 6). Participants were debriefed at the end of the experiment regarding the experimental manipulation.



Task: Several scientists have identified information overload as a problem facing today's society. Information overload occurs when a person consumes more information than they can process. What are some common sources of information overload? What is the best way to prevent it?

System A

Enter search terms above to display results.

Task: For his 16th birthday, your nephew has asked you for a video game that is rated "M" for mature audiences because it contains intense violence. Your are unsure about whether to purchase this game because you recently overheard two people discussing the effects of violent video games on teenagers. What are some of the reported effects of violent video games on teenagers? Do you believe these reports? Why or why not?

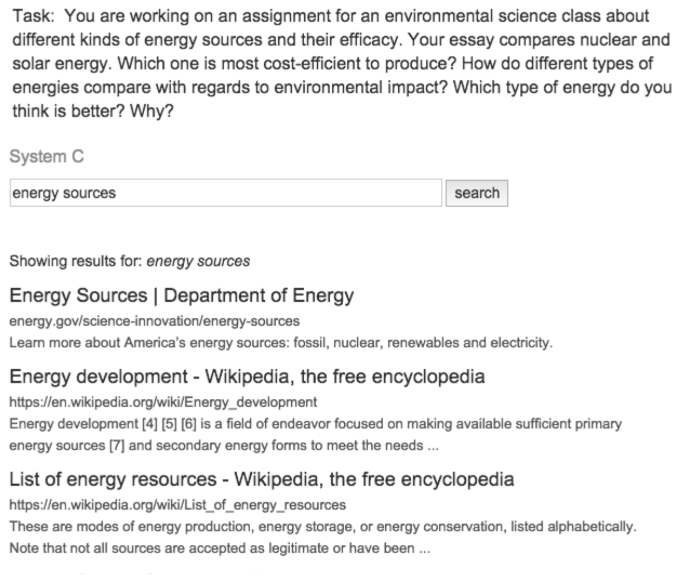
System B

Enter search terms above to display results.

Figure 6. Example of Cover Story Implementation

3.6. Search System

Search behavior was recorded via a search system developed for and used in other studies (Capra et al., 2015). This system used the Bing Web Search API to provide search results from the open web and appears similarly to a standard search system (Figure 7). This system also allowed participants to query, click and view results, and bookmark pages they found relevant. In addition to searching, participants were asked to provide a short description of why they bookmarked each document (Figure 8).



Task: You are working on an assignment for an environmental science class about different kinds of energy sources and their efficacy. Your essay compares nuclear and solar energy. Which one is most cost-efficient to produce? How do different types of energies compare with regards to environmental impact? Which type of energy do you think is better? Why?

System C

energy sources

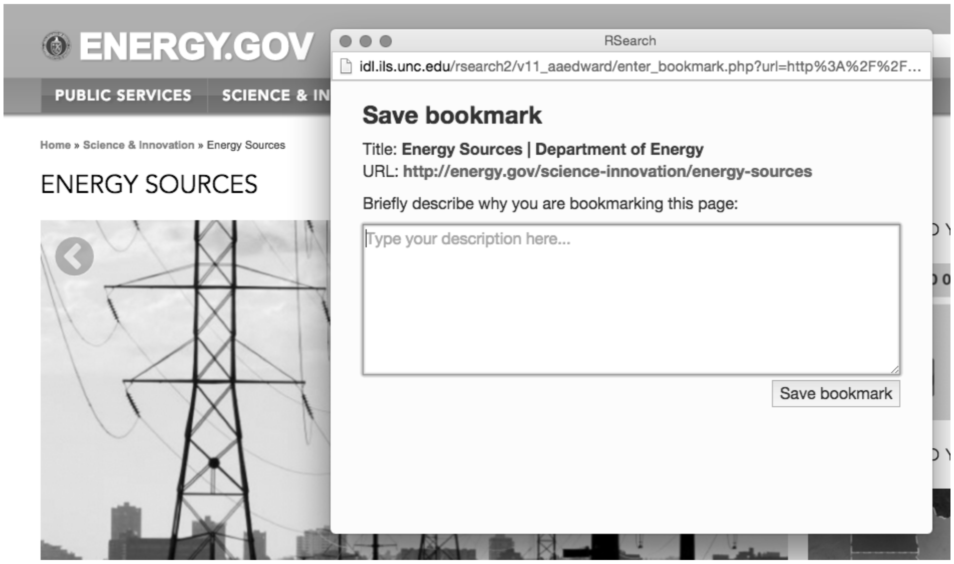
Showing results for: *energy sources*

Energy Sources | Department of Energy
energy.gov/science-innovation/energy-sources
Learn more about America's energy sources: fossil, nuclear, renewables and electricity.

Energy development - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Energy_development
Energy development [4] [5] [6] is a field of endeavor focused on making available sufficient primary energy sources [7] and secondary energy forms to meet the needs ...

List of energy resources - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/List_of_energy_resources
These are modes of energy production, energy storage, or energy conservation, listed alphabetically. Note that not all sources are accepted as legitimate or have been ...

Figure 7. Screenshot of Search System



idl.ils.unc.edu/rsearch2/v11_aaedward/enter_bookmark.php?url=http%3A%2F%2F...

Save bookmark

Title: **Energy Sources | Department of Energy**
URL: **http://energy.gov/science-innovation/energy-sources**
Briefly describe why you are bookmarking this page:

Type your description here...

Figure 8. Screenshot of Bookmarking Process

3.7. Pre – Search Questionnaire

Participants completed a pre – search questionnaire (Appendix B) before each search to better understand their overall search experience, interest in the topic, and appraised difficulty, among other items. The pre – search questionnaire assessed participant's prior knowledge and experience searching the topic, and contained the following five items: "How relevant is this topic to your life?", "How interested are you to learn more about this topic?", "Have you ever searched for information related to this topic?", "How much do you know about this topic?" and "How difficult do you think it will be to search for information related to this topic?" The responses on this questionnaire were recorded on a five-point Likert-type scale. This questionnaire was administered before the task, but after the participant read the full description of the task.

3.8. Post – Task Questionnaire

A post-task questionnaire (Appendix C) was given to participants after they completed each task. The post-task questionnaire assessed participants' experiences of difficulty and success with the search process, as well as perceptions of their own skill, and the ability of the system to retrieve documents. The questionnaire contained the following four items: "How difficult was it to find relevant documents?", "How would you rate your ability at retrieving documents?", "How successful was your search?", and "How would you rate the system's ability at retrieving relevant documents?".

3.9. Measurement of Engagement

Engagement was defined in this study using O'Brien and Toms' (2008) definition of engagement, which is "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). O'Brien and Toms' (2008) User Engagement Scale (recently refactored in a study on the UES and exploratory search (O'Brien & Toms, 2013)) was used to measure engagement. The scale contains items that load on six components of engagement: *focused attention*, *perceived usability*, *aesthetics*, *endurability*, *novelty*, and *felt involvement*. People respond to these items using a five-point

Likert-type scale where 1 = strongly disagree and 5 = strongly agree. Participants in this study only responded to the *focused attention*, *endurability*, *novelty*, *felt-involvement*, *and perceived usability* portions of the UES after each task. They did not respond the *aesthetics* items because those were interface-related, and the interface was constant throughout, so there was no need to measure these items after each task. These were also excluded because of concerns regarding the number of items participants were required to respond to. Questionnaire items were changed from the possessive "my" to "this" (Appendix D). Participants completed the UES as part of a larger post-task questionnaire at the end of each task. The *perceived usability* subscale was presented before the other subscales, and the ordering of the questions in this subscale were rotated for each participant. The other subscales were interleaved with questions from the frustration and SSSQ scales, and the order of these questions was counterbalanced as well.

3.10. Measurement of Frustration

Frustration was defined in this study using Amsel (1992)'s definition of primary frustration, or "a temporary state that results when a response is nonreinforced" in the presence of reward expectancy" (p. 2). Frustration is usually measured from three angles: a person's natural predilection for frustration, how they cope with frustrating experiences, or what level of frustration they feel in the moment. A review of the literature revealed no standardized, widely accepted instrument to measure frustration. Peters, O'Connor and Rudolf (1980)'s three-item questionnaire (Appendix E) was used to measure frustration, and the items were changed to reflect a search context. Frustration was also measured via one item in the *perceived usability* subscale of the UES. The items from this questionnaire were interleaved with items from the SSSQ questionnaire and items from the *focused attention, felt involvement, endurability*, and *novelty* subscales of the UES.

3.11. Measurement of Stress

Stress was defined using Lazarus and Folkman’s (1984) definition of stress as “a particular relationship between the person and environment that is appraised by the person as taxing or exceeding his or her resources” (p. 40). This definition was especially helpful because it highlights the cognitive appraisal of resources available to cope with the task, which is key to engendering feelings of engagement and frustration. Stress was measured physiologically using skin conductance and heart rate collected via the BioPac. Stress was also measured using the Short State Stress Questionnaire (Helton, 2004). Both objective and subjective measures were collected to allow for comparisons between perceived stress and the physiological experience of stress.

3.11.1. Measurement of Physiological Stress. This study measured electrodermal activity and electrocardiography as indicators of physiological stress. Electrical skin activity measurement consists of two methods: exosomatic and endosomatic (Boucsein, 2012). Exosomatic measurement can be defined in terms of skin conductance, skin resistance, skin admittance, and skin impedance levels and responses. This study measured skin conductance. Exosomatic measurement can be measured with direct or alternating current, with the idea that as sweat levels change, skin conductance (the ability of the electrical signal to pass through the electrode site) levels will change accordingly. Changes in skin conductance can be grouped into either short-term changes, which are called phasic changes, or long-term changes, which are called tonic changes.

Measuring changes in heartbeat as an indicator of emotion usually takes one of three forms: cardiac output, heart rate variability, and electrocardiography. Heartbeats are regulated by two nodes: the sinoatrial and atrioventricular nodes. The sinoatrial node is responsible for sending an electrical signal to begin the pumping action of the heart. Electrocardiography (ECG) refers to the measurement of this electrical activity. Electrodes placed on the chest measure the electrical activity generated during the contraction of the atria and left and right ventricles (Barber, Brown, & Smallwood, 1984).

3.11.2. BioPac. BioPac Systems is the name for an entire family of devices that include wearables as well as stationary devices. The device used in this study was the MP35, which "can be used with BIOPAC amplifiers and accessories,"[1] thus making it customizable, and offers "high resolution (16 bit), variable sample rates for analog and calculation channels, 16 analog inputs and two analog outputs, digital I/O lines (automatically control other TTL level equipment), and 16 online calculation channels." This device, combined with the AcqKnowledge software and BSL 4.0 (the latest version of software), were used to collect and analyze physiological signals. In this experiment, electrodermal activity (skin conductance) and heart rate were measured. These two measures were chosen because studies have shown that these are the most prominent signals of stress aside from saliva samples and electroencephalography (EEG) signals (Boucsein, 2012).

Braithwaite, Watson, Jones, and Rowe (2013) offered recommendations specifically for using the BioPac and accompanying software to perform electrodermal measurement. Braithwaite et al. recommend a low sampling rate, or the rate at which the system takes measurements of electrodermal activity, for long-term studies that do not require "temporal precision:" one to five samples per second. For shorter studies, where there are specific

[1] http://www.biopac.com/products/

events or stimuli that are timed, they recommend a much higher sampling rate of 2000 samples per second. For channel sample rates, measured in Hz, Braithwaite et al. recommend that using 1000Hz - 2000Hz sample rate to maintain proper signal integrity. Gain must also be calibrated. Gain is a circuit's ability to increase the amplitude of a signal between input and output points, and is important to measuring electrodermal activity because of potential interference during measurement; i.e., if a signal is weak, it should be amplified in order to properly measure it. Braithwaite et al point out that setting gain is a matter of deciding between the tradeoffs in gain and dynamic range; i.e., the higher the gain, the lower the dynamic range, and so recommend a gain of x2000, though indicating that both x2000 and x5000 work well.

Recording electrodermal activity with the BioPac requires placing electrodes on the palm, checking to make sure the signals are being recorded correctly, and observing a baseline. This study can be thought of as continuous measurement of electrodermal signals because participants spent between 20-30 minutes searching. One thousand samples per second was used in this experiment and offered a more fine-grained look at electrodermal activity, which is better when recording over a long period of time as it allows more opportunity to observe nuances in recordings.

The BioPac has a set of prescribed methods of measurement for ECG and heart rate variability. These manuals recommend a three-point lead arrangement, on the right arm near the wrist, the right leg just above the ankle, and the left leg right above the ankle. The ankle electrodes should be placed on the skin, and not over the bone. They also recommend that the skin be lightly abraded before placing the electrodes on the area. In keeping with recommendations from the ACC, it is advised that the participant lay flat on their backs during the recording. Heart rate will be extracted from these ECG values. Recording ECG with the BioPac involves attaching the appropriate electrodes to the rib and collarbone of the participant, checking to make sure the values are being recorded correctly, and obtaining a baseline.

To measure skin conductance, two EL507 electrodes were attached to the thenar and hypothenar eminences of participants' palms (see Appendix F). Both tonic (overall change) and phasic skin conductance (change during the task) were measured. To measure heart rate, two EL503 electrodes were attached to the right collarbone and left rib of participants to gather electrocardiography (ECG) data from which heart rate was extracted (see Appendix F). A three-minute resting period was observed at the beginning of the experiment to obtain a baseline. Participants were instructed to remain seated and wait until they were instructed to proceed with the task; the investigator made light conversation during this time period to put the participant at ease. The beginning of a task was defined as when the participant sees the search task prompt and answers the pre-task questionnaires, and the end of the task was defined as when the participant clicks "end task."

In pilot tests, the physiological signals also showed encouraging changes between tasks – one participants' data followed a near sine-like curve that showed more arousal during more frustrating tasks.

3.11.3. Measurement of Self-Reported Stress. Stress was also measured using the Short Stress State Questionnaire (SSSQ) developed by Helton (2004) as a short version of the Dundee Stress State Questionnaire (DSSQ) (Matthews et al., 1999). Helton extracted 24 items from the DSSQ and consolidated the factor structure, revealing three factors: distress, worry, and engagement. The distress factor is a measure of negative affect as it contains items related to negative mood, i.e., “I feel depressed,” “I feel sad” or “I feel irritated.” The engagement factor is a measure of motivation, because the questions are exclusively about abilities and feelings about performance, i.e., “I am committed to attaining my performance goals,” and “I feel confident in my abilities.” The worry factor is a measure of cognition as well as self-perception: “I feel concerned about the impression I am making,” and “I thought about how I would feel if I were told how I performed.” Participants responded to these items using a five point Likert-type scale where 1=never and 5=very often (Appendix G). Questionnaire items were changed to reflect the past tense, and items were changed from “this task” to “this search task.”

3.12. Search Behavior

Search behavior was measured using behavioral signals in four categories: query-based measures, SERP-based measures, click-based measures and time-based measures. Table 1 details the measures and their definitions.

Table 1

Search Behavior Measures and Definitions

Type	Measure	Definition
Query	Number of Queries	The number of queries entered into the search results box during the search session
	Query Terms	The number of distinct query terms in each submitted query (not including stopwords)
	Clicks per Query	The number of clicks for a particular query
	Term Uniqueness	How unique a query term is to a corpus
	Queries without Clicks	Queries submitted without a result click
SERP	SERPs Displayed	Number of unique SERPs displayed
	Scrolls	The number of mouse scrolls on the SERP, where a mouse scroll is defined as movement of the wheel button in the middle of the mouse while the participant is on the SERP.
	Documents Bookmarked	The number of documents participants bookmarked in total
Clicks	Clicks on the SERP	Number of clicks on a SERP result
	Clicks on Documents	Number of clicks on a link in a document
Time	Time Spent on Task	The time between the first query the participant enters to when he/she indicates that they have completed the task.
	Time Spent on the SERP	Time spent on the SERP over the course of the task (in minutes)
	Time Spent on Documents	Time spent on a document (in minutes)
	Query Time Intervals	The time between queries for each task

These interaction signals were chosen because they have been used in other work examining engagement and search behavior (O’Brien & Lebow, 2013) and are among those that have been used to determine whether a participant is struggling or exploring (Hassan et al., 2014), which is closely related to the concepts of engagement and frustration. The query-based measures (query terms, time per query, and documents bookmarked per query) were chosen because they can serve as an indicator of search strategy (Hassan et al., 2014). Number of clicks was also chosen as an indicator of search strategy, given that they too were shown in Hassan et al. to be a reliable indicator of struggling or exploring. Morae Observer was also used to record video of the participant as they completed the search tasks as a secondary, visual measure of search behavior, as well as a measure of clicks on documents.

In pilot tests, participants appeared to click the most during tasks where they were engaged and did not receive a frustrator, followed by tasks during which they were engaged and received a frustrator, ending with tasks during which they were not engaged and received no frustrator. Therefore, the trend seemed to be that participants who were engaged clicked more, but this was dampened by the frustrator.

3.13. Procedure

The experiment was conducted in Room 12 in the Interactive Information Systems Lab in Manning Hall. Participants ranked the tasks the day before they participated in the experiment. When participants arrived to the lab, they were greeted, received a brief description of the experiment from a prepared script, and then were asked to sign the consent form. After signing the consent form, they completed the demographics questionnaire (see Appendix H). Then, the four required electrodes were attached to the participant, the BioPac program was started, and a three-minute baseline was recorded. Next, Morae Recorder was started, and participants completed four tasks. These four tasks were the two tasks the participant rated as most interesting (in order to foster interest) and the two tasks they rated as least interesting. The frustrator was delivered during the most interesting task (the task ranked 1) and the most uninteresting task (the task ranked 8). The tasks were counterbalanced such that all task rankings appeared in all positions, and the frustrator was evenly distributed throughout. Participants completed the pre-task questionnaire before each of the four tasks, and the post-task, UES, SSSQ, and frustration questionnaires at the end of each task. Once all tasks were complete, the electrodes were removed. The participant was then paid and thanked for their participation.

3.14. Participants

Sample size was determined using a power analysis. With an effect size of 0.25 (assuming a conservative effect size), a power of 0.95, an α of 0.05, with two independent variables (interest and frustration) and four measurements (four tasks) 40 participants were required. These 40 participants were selected via a convenience sample from the UNC undergraduate population, and were recruited via a recruitment email (see Appendix I) that was distributed to all majors and class levels. Information science majors were excluded because of their knowledge of search systems. Participants were required to be between the age of 18-24 and in reasonable physical health, meaning free from any obvious physical impairments. If participants engaged in strenuous physical activity up to 30 minutes before their participation in the study, they were asked to observe a ten-minute waiting period to ensure that they were relaxed and ready to participate in the experiment (Braithwaite et al., 2013). No participants had to observe this ten-minute waiting period.

Thirty-one participants were female, and eight participants were male. Participants’ average age was 20 (*SD*=1.77). There were two first-year students, 12 sophomores, six juniors, and 20 seniors. Thirteen participants were in the natural sciences, 11 were in the humanities, eight were in the social sciences, seven were in professional schools, and there was one undecided person. Participants reported having on average between seven and nine years of search experience (*SD*=0.87), and said that they conducted online searches for information on average more than seven times a day (*SD*=1.11).

CHAPTER IV: RESULTS

The results section is organized as follows: first, participants' pre-task perceptions of the search task and post-task evaluations of the search experience are summarized. Next, the findings from the self-report data (i.e., the UES, SSSQ, and frustration questionnaires) are summarized. Next, a section is devoted to summarizing the results of the search behavior analysis. Finally, the findings from the analysis of the physiological data are summarized. For each of these sections, means and standard deviations will be reported across three dimensions: interest, frustration, and the combinations of the two. Two-way repeated measures ANOVAs with two factors (interest and frustration) at two levels (yes/no and present/absent, respectively) were conducted throughout to examine the main and interaction effects of interest and frustration on several constructs. Post-hoc tests with Bonferroni correction were also used throughout.

4.1. Task Topic Ranking Questionnaire

Participants were asked to rank the task topics on a scale of 1-8, where 1 = most interesting and 8 = least interesting. Table 2 details how many people ranked each task topic as interesting or uninteresting. The task topics were presented via a 1-2-word description. The online communication task was the most popular. Twenty-five people rated this task as the task they thought was most interesting. The vehicle purchasing and tattoo removal tasks were the least popular; 15 people rated each of those tasks as uninteresting.

Table 2

Frequencies of Participant Rankings for Each Task Topic

Task Topic	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8
Tattoo Removal	3	4	4	4	5	7	6	7
Lupus	4	5	8	3	5	5	3	7
Energy Sources	4	7	7	6	8	2	5	1
Biomass Fuel	2	4	2	4	7	8	6	7
Online Communication	16	9	2	3	3	4	2	1
Vehicle Purchases	4	4	5	7	2	3	6	9
Video Game Violence	1	3	7	8	4	5	5	7
Endurance Sports	6	4	5	5	6	6	7	1

Note. Lighter color indicates most interesting tasks, darker color indicates least interesting tasks.

4.2. Pre-Search Experience Questionnaire

The pre-search experience questionnaire measured items such as participants' perceptions of difficulty and pre-search knowledge before the search based on their understanding of the full-text description of the task. The items were measured on a five-point Likert-type scale where 1=strongly disagree and 5=strongly agree, therefore higher scores indicate more positive values. Table 3 shows the means, standard deviations, and t-test results of these data. Paired samples t-tests were used to analyze these data. As shown in Table 3, participants rated their knowledge of the task significantly higher for tasks in which they were interested (i.e., the two tasks they ranked as most interesting) than tasks in which they were not interested (i.e., the two tasks they ranked as least interesting). They also rated relevance significantly higher for tasks in which they were interested than tasks in which they were not interested. Participants also rated interest significantly higher for tasks in which they were interested than tasks in which they were not interested, indicating that the manipulation of interest was successful. Lastly, participants said they searched significantly more frequently for information related to topics in which they were interested than topics in which they were not interested and indicated that they felt tasks in which they were interested would be significantly less difficult than tasks in which they were not interested.

Table 3

Pre-Search Questionnaire Item Means According to Interest

Questionnaire Item	Interesting Tasks	Uninteresting Tasks	*t*	*df*
Knowledge	3.00 (1.13)	2.20 (1.80)	5.27***	79
Relevance	3.61 (1.23)	2.24 (1.23)	7.06***	79
Interest	4.01 (0.96)	2.77 (1.15)	7.39***	79
Frequency	2.44 (1.20)	1.64 (0.94)	4.69***	79
Difficulty	3.35 (1.15)	3.81 (0.93)	-2.72***	79

Note. ****p*<0.001. Standard deviations appear in parentheses.

4.3. Post - Task Questionnaire

The five-item post-task experience questionnaire measured constructs such as perception of difficulty, skill, and system ability. These questionnaire items were scored on a five-point Likert-type scale, but this was inverted from the pre-search questionnaire, where 1=not good and 5=very good, for example. This means that lower scores represent more difficulty, poorer success, lower skill, and poorer system ability. Means and standard deviations of tasks by dimension and their combinations (i.e., interest, frustration, and the combination of the two) were computed. Table 4 details the means and standard deviations of each item on the post-task questionnaire by interest and frustration levels. Table 4 shows participants found interesting tasks less difficult, felt they had greater skill, and felt the system showed greater ability in retrieving relevant documents during interesting tasks. Participants also felt that they had greater success with interesting tasks than uninteresting tasks. Participants found frustrating tasks more difficult than not frustrating tasks. They also rated their skill lower for frustrating tasks than non-frustrating tasks. Participants rated the system's ability to retrieve documents lower for frustrating tasks than non-frustrating tasks, indicating that the cover story for the experiment was largely successful. Lastly, participants felt that they were less successful when completing frustrating tasks than not frustrating tasks.

Table 4

Post-Task Questionnaire Results by Interest and Frustration Levels

	Interest		Frustration		Combinations			
Item	Yes	No	Present	Absent	IyFy	IyFn	InFy	InFn
Difficulty	3.61 (0.83)	3.38 (0.89)	3.33 (1.26)	1.66 (0.91)	3.25 (1.35)	1.52 (0.88)	3.42 (1.34)	1.80 (1.04)
Skill	3.85 (0.70)	3.62 (0.85)	3.29 (1.02)	4.19 (0.84)	3.45 (1.15)	4.25 (0.78)	3.12 (1.26)	4.13 (0.91)
System Ability	3.49 (0.86)	3.35 (1.04)	2.49 (1.50)	4.26 (1.19)	2.48 (1.50)	4.32 (1.16)	2.50 (1.50)	4.20 (1.22)
Success	3.77 (0.69)	3.80 (0.77)	3.14 (1.28)	4.44 (0.90)	3.05 (1.28)	4.50 (0.88)	3.23 (1.29)	4.38 (0.92)

Note. Standard deviations appear in parentheses. Y indicates a variable is present, N indicates it is absent.

A two-way repeated measures ANOVA was completed with two factors (interest and frustration) at two levels (yes and no, and present and absent, respectively) to observe the main and interaction effects of interest and frustration on the post-search questionnaire items. Table 5 shows the results of this analysis. There were no significant main effects for interest for any of the post-task questionnaire items. However, there were significant main effects of frustration for each of the post-task items. This means that participants experienced significantly greater difficulty, felt significantly less success, felt significantly less skillful, and rated the system's ability to retrieve relevant documents as significantly poorer when the frustrator was present. There were no significant interaction effects.

Table 5

Results of ANOVA for Post-Task Questionnaire Items

		Source		
Post – Task Items	Critical Values	Interest	Frustration	Interest x Frustration
Difficulty	SS	2.02	112.22	0.10
	F	1.76	56.27***	0.20
	η2	0.04	0.59	0.00
	p	0.19	<0.001	0.66
Skill	SS	2.02	32.40	0.40
	F	2.47	23.57***	0.72
	η2	0.06	0.37	0.02
	p	0.12	<0.001	0.40
System Ability	SS	0.10	126.02	0.22
	F	0.07	43.31***	0.28
	η2	0.00	0.53	0.00
	p	0.79	<0.001	0.60
Success	*SS*	0.02	67.60	0.90
	F	0.03	32.99***	1.23
	η2	0.00	0.46	0.03
	p	0.86	<0.001	0.27

Note. ****p*<0.001. Degrees of freedom are (1, 39) for each item.

4.4. UES (User Engagement Scale)

The UES is constructed such that higher overall scores indicate higher engagement, and higher scores on each of the subscales (*perceived usability, felt involvement*, *focused attention*, *novelty*, and *endurability*) represent higher levels of each of these constructs. The UES data was analyzed to look for differences among subscales as well as in total overall engagement. Before analysis was completed, some items of the perceived usability and endurability subscales were reverse-scored, as these items were worded negatively and thus higher scores indicated greater dissatisfaction. This means that higher perceived usability scores indicate greater satisfaction. To aid in interpretation of these values, the range of scores for each UES item is described in Table 6.

Table 6

Range of Values for UES Subscales

Subscale	Number of Items	Minimum and Maximum Scores
Perceived Usability	7	7 – 35
Focused Attention	5	5 – 25
Felt Involvement	2	3 – 15
Endurability	6	6 – 30
Novelty	2	2 – 10
Total Engagement	22	23-115

The scores for the UES data were computed such that the values of for each question of each subscale were averaged, and then this average score was added to the other items in the subscale to produce the score for that subscale (Table 7). Overall engagement was higher for interesting tasks than for uninteresting tasks, and participants rated interesting tasks higher for each engagement subscale. Overall engagement was also higher for not frustrating tasks than frustrating tasks. Table 7 also details the means and standard deviations for each UES subscale by frustration. The frustration results are similar to the interest results in that participants reported higher *novelty, endurability*, *felt involvement,* and slightly higher *focused attention,* and higher *perceived usability* for not frustrating tasks. Examining the means and standard deviations by combinations of interest and frustration, we see that interesting and not frustrating tasks had the highest overall engagement, and uninteresting and frustrating tasks had the lowest overall engagement.

Table 7

UES Scores for Each Subscale by Interest and Frustration Levels

	Interest		Frustration		Combinations			
Item	Yes	No	Present	Absent	IyFy	IyFn	InFy	InFn
Perceived Usability	3.63 (0.60)	3.56 (0.79)	2.93 (1.05)	4.25 (0.82)	2.91 (1.02)	4.34 (0.75)	2.94 (1.09)	4.18 (0.89)
Focused Attention	2.87 (0.79)	2.64 (0.66)	2.73 (0.82)	2.78 (0.85)	2.87 (0.85)	2.87 (1.00)	2.59 (0.79)	2.68 (0.70)
Felt Involvement	3.61 (0.50)	3.27 (0.72)	3.16 (1.00)	3.72 (0.72)	3.36 (0.85)	3.85 (0.73)	2.96 (1.11)	3.58 (0.70)
Endurability	3.41 (0.58)	3.20 (0.76)	2.83 (1.02)	3.78 (0.79)	2.94 (0.99)	3.87 (0.71)	2.72 (1.05)	3.68 (0.86)
Novelty	3.72 (0.67)	3.09 (0.90)	3.15 (1.20)	3.66 (0.97)	3.52 (1.04)	3.91 (0.86)	2.77 (1.24)	3.41 (1.01)
Total Engagement	3.45 (0.42)	3.15 (0.64)	2.96 (0.82)	3.64 (0.64)	3.12 (0.74)	3.77 (0.58)	2.80 (0.90)	3.51 (0.67)

Note. Standard deviations in parentheses. Y indicates a variable is present, N indicates it is absent. Shaded values indicate significance.

A two-way repeated measures ANOVA was conducted to look for significant main and interaction effects of interest and frustration on the subscales of engagement (see Table 8). A significant main effect was found for frustration for *perceived usability, felt involvement*, and *endurability*. A significant main effect for interest was found for *felt involvement, focused attention*, *novelty*, and total engagement. No significant interaction effects were found for any of the subscales.

Table 8

Results of ANOVA of UES data by Subscale

		Source		
Items	Critical Values	Interest	Frustration	Interest x Frustration
Perceived	SS	0.16	70.60	0.37
Usability	*F*	0.24	56.17***	0.99
	η2	0.00	0.59	0.02
	p	0.62	<0.001	0.32
Felt	SS	4.22	12.66	0.16
Involvement	*F*	5.35*	19.87***	0.19
	η2	0.12	0.34	0.00
	p	0.03	<0.001	0.66
Focused	SS	2.21	0.08	0.08
Attention	*F*	7.21**	0.22	0.26
	η2	0.16	0.00	0.00
	p	0.01	0.64	0.61
Endurability	SS	2.06	37.22	0.00
	F	3.41	36.26***	0.00
	η2	0.80	0.48	0.00
	p	0.07	<0.001	0.99
Novelty	SS	15.62	10.51	0.70
	F	19.66***	8.65	0.62
	η2	0.33	0.18	0.02
	p	<0.001	<0.01	0.34
Total	SS	3.51	18.57	0.03
Engagement	*F*	10.02**	28.42***	0.10
	η2	0.20	0.42	0.00
	P	0.003	<0.001	0.75

Note. **p*<0.05, ***p*<0.01, ****p*<0.001. Degrees of freedom are (1, 39) for each item.

4.5. SSSQ (Short Stress State Questionnaire)

The data from the SSSQ was examined for differences in overall stress and differences in stress by the two SSSQ subscale, distress and worry. The range of scores for the SSSQ items for the worry subscale is 7 – 35, while the range of scores for the distress subscale is 8 – 40. This means that the total range of scores for overall stress is 15 – 75. The subscales were also weighted evenly, though the factor loadings for each question with regards to worry and distress were different. The means for overall stress and subscale were examined by dimension (interest, frustration, and the combination of the two). Table 9 details the means and standard deviations for distress, worry, and overall stress scores by interest and frustration level. Participants reported higher distress, but lower worry scores for interesting tasks. Overall, participants reported slightly lower stress for interesting tasks than uninteresting tasks. Participants reported higher distress and worry as well as higher overall stress for frustrating tasks.

Table 9

SSSQ Results by Distress, Worry and Overall Stress by Interest and Frustration Levels

	Interest		Frustration		Combinations			
Item	Yes	No	Present	Absent	IyFy	IyFn	InFy	InFn
Distress	1.56 (0.62)	1.60 (0.73)	1.58 (0.67)	1.27 (0.47)	1.88 (0.64)	1.24 (0.40)	1.92 (0.76)	1.29 (0.55)
Worry	1.52 (0.51)	1.50 (0.54)	1.56 (0.56)	1.46 (0.48)	1.57 (0.56)	1.50 (0.46)	1.55 (0.58)	1.44 (0.49)
Stress	1.55 (0.45)	1.56 (0.53)	1.75 (0.50)	1.35 (0.39)	1.74 (0.45)	1.35 (0.36)	1.76 (0.56)	1.35 (0.42)

Note. Standard deviations in parentheses. Y indicates a variable is present, N indicates it is absent.

Table 9 shows participants reported the highest overall stress during uninteresting and frustrating tasks, while tasks without a frustrator had lower overall stress. Examining the data by distress and worry, participants reported the highest levels of distress for uninteresting tasks with a frustrator, and reported the lowest distress for interesting tasks without a frustrator. By contrast, participants reported the lowest worry during uninteresting and not frustrating tasks, and the highest worry during interesting and frustrating tasks.

A two-way repeated measures’ ANOVA was conducted to examine the main and interaction effects of interest and frustration on stress (Table 10). As shown in Table 10, there were significant main effects for frustration for distress, but no significant main effects for frustration on worry, and no significant interest or interaction effect for either subscale. There was a significant main effect for frustration found on overall stress, but there were no significant main effects found for interest and no significant interaction effects.

Table 10

Results of ANOVA of SSSQ Data by Subscale

		Source		
SSSQ Items	Critical Values	Interest	Frustration	Interest x Frustration
Distress	SS	0.07	15.83	0.00
	F	0.26	34.54***	0.00
	η2	0.01	0.47	0.00
	p	0.61	<0.001	0.95
Worry	SS	0.03	0.40	0.00
	F	0.16	2.35	0.06
	η2	0.00	0.57	0.00
	p	0.76	0.13	0.80
Total Stress	SS	0.05	6.32	0.00
	F	0.03	30.18***	0.00
	η2	0.00	0.44	0.00
	p	0.86	<0.001	0.94

Note. ****p*<0.001. Degrees of freedom are (1, 39) for each item.

4.6. Post-Search Frustration Questionnaire

The frustration questionnaire contained three items. Each item asked about the experience of frustration during the task. Table 11 details the means and standard deviations for frustration for each frustration question by interest and frustration level. Interesting tasks were described as slightly more frustrating than uninteresting tasks. Frustrating tasks were described as more frustrating than not frustrating tasks, indicating a successful manipulation of frustration. Examining the tasks by both dimensions, uninteresting and frustrating tasks had the highest frustration, and interesting tasks without a frustrator had the lowest frustration.

Table 11

Frustration Questionnaire Results by Interest and Frustration Levels

	Interest		Frustration		Combinations			
Item	Yes	No	Present	Absent	IyFy	IyFn	InFy	InFn
Average Frustration	2.42 (1.04)	2.52 (0.97)	3.18 (1.10)	1.75 (0.94)	3.16 (1.04)	1.68 (0.80)	3.21 (1.17)	1.82 (1.08)

Note. Standard deviations appear in parentheses. Y indicates a variable is present, N indicates it is absent.

A two-way repeated measures ANOVA was conducted to look for main and interaction effects of interest and frustration on frustration. Table 12 summarizes the results of this analysis. A significant main effect was found for frustration, but there were no significant main or interaction effects for interest.

Table 12

Results of ANOVA of Frustration Questionnaire Data

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	0.00	2	0.02	0.01	0.89
Frustration	6.28	2	70.52**	0.97	0.01
Interest x Frustration	0.09	2	0.86	0.30	0.45

Note. ***p*<0.01.

4.6.1. Summary of UES, SSSQ, and Frustration Questionnaire Data. The results of the questionnaire data (see Table 13) show that participants experienced significantly higher engagement when searching during tasks that they were interested in (regardless of frustrator). Participants also felt significantly more frustrated when they completed tasks with a frustrator, indicating that the engagement and frustration manipulations were largely successful.

Participants also reported less stress for interesting tasks than uninteresting tasks, and more stress for frustrating tasks than not frustrating tasks.

Table 13

Summary of Responses for UES, SSSQ, and Frustration Questionnaires

	Interest		Frustration		Combinations			
Item	Yes	No	Present	Absent	IyFy	IyFn	InFy	InFn
Engagement	3.45 (0.42)	3.15 (0.64)	2.96 (0.82)	3.64 (0.64)	3.12 (0.74)	3.77 (0.58)	2.80 (0.90)	3.51 (0.67)
Frustration	2.42 (1.04)	2.52 (0.97)	3.18 (1.10)	1.75 (0.94)	3.16 (1.04)	1.68 (0.80)	3.21 (1.17)	1.82 (1.08)
Stress	1.55 (0.45)	1.56 (0.53)	1.75 (0.50)	1.35 (0.39)	1.74 (0.45)	1.35 (0.36)	1.76 (0.56)	1.35 (0.42)

Note. Standard deviations appear in parentheses. Gray selections indicate significance.

4.7. Search Behavior Data

Behavioral data was gathered from two sources: log data from the search system, and Morae recordings. The system log data included signals from actions on the SERP, such as clicks, queries, and mouse activity. The Morae recordings collected data about clicks, mouse activity and time on documents that were linked to from the SERP. The signals gathered can be separated into four categories: SERP (SERPs displayed, scrolls on the SERP), Queries (e.g., queries entered, query length, clicks per query, query term uniqueness), and Clicks (e.g., clicks on SERP results, as well as clicks on documents) and Time (time spent on the SERP and on documents, as well as query time intervals). Scrolls on the SERP are defined as when the mouse wheel (or the center button in the middle of a mouse) is scrolled on a SERP. SERP clicks are defined as when the participant clicks on a SERP result. Document clicks are defined as when a participant clicks on a link within a document, meaning a page that is not the SERP. In addition to these interaction signals, bookmarks were collected, as well as bookmark annotations, which were text participants entered as justification for bookmarking a particular webpage.

4.7.1. Queries. Participants encountered the query box and were able to enter queries after answering the pre-search questionnaire and viewing the full description of the task. Means and standard deviations for number of queries, words per query, and clicks per query were examined for the tasks by dimension (interest, frustration and the combinations of the two).

4.7.1.1 Number of Queries Submitted. Participants submitted slightly more queries for interesting tasks than uninteresting tasks. Participants submitted more queries for frustrating tasks than not frustrating tasks. Table 14 summarizes the means and standard deviations for queries submitted by both interest and frustration. Participants submitted the most queries for interesting and frustrating tasks and the least queries for uninteresting tasks without a frustrator.

Table 14

Queries Submitted by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	7.87 (4.73)	3.42 (2.73)	5.65 (3.73)
	No	7.15 (4.62)	2.97 (2.08)	5.06 (3.35)
	Totals (Frustration)	7.51 (4.67)	3.20 (2.40)	

Note. Standard deviations appear in parentheses.

A two-way repeated measures ANOVA was conducted to determine the main and interaction effects of interest and frustration on number of queries submitted; the results of this analysis are summarized in Table 15. There were no significant main effects for interest or significant interaction effects. However, a significant main effect for frustration was found.

Table 15

Results of ANOVA of Queries Submitted

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	13.81	39	0.96	0.02	0.33
Frustration	743.91	39	57.18***	0.60	<0.001
Interest x Frustration	0.76	39	0.06	0.00	0.81

Note. ****p*<0.001.

4.7.1.2. Words Per Query. Uninteresting and interesting tasks had virtually the same number of words per query. Frustrating tasks and not frustrating tasks also had a similar amount of words per query. Table 16 shows the means and standard deviations of query length in words by interest and frustration. The query lengths were very similar, but overall uninteresting and not frustrating tasks had the highest number of words per query, while interesting tasks and not frustrating tasks had the lowest number of words per query.

Table 16

Query Length by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	3.65 (1.88)	3.36 (1.47)	3.50 (1.67)
	No	3.40 (1.77)	3.91 (1.80)	3.65 (1.79)
	Totals (Frustration)	3.53 (1.83)	3.62 (1.64)	

Note. Standard deviations appear in parentheses.

A two-way repeated measures ANOVA was conducted to examine the main and interaction effect of interest and frustration on query length (as shown in Table 17). There were no significant main effects for interest or significant interaction effects. However, there was a significant main effect for frustration on query length.

Table 17

Results of ANOVA of Query Length Data

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	180.62	39	0.73	0.02	0.40
Frustration	9030.02	39	52.07***	0.57	<0.001
Interest x Frustration	184.90	39	1.15	0.03	0.29

Note. ****p*<0.001.

4.7.1.3. Clicks Per Query. Participants clicked more during interesting tasks than uninteresting tasks. Participants also clicked more during frustrating tasks than not frustrating tasks. Examining the data by combinations of interest and frustration (see Table 18), participants clicked the most per query for interesting and frustrating tasks and clicked the least per query for uninteresting and not frustrating tasks.

Table 18

Clicks Per Query by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	8.32 (5.32)	7.22 (4.16)	7.77 (4.77)
	No	8.22 (4.79)	5.35 (2.15)	6.79 (3.97)
	Totals (Frustration)	8.27 (5.03)	6.29 (3.42)	

Note. Standard deviations appear in parentheses.

A two-way repeated measures ANOVA was conducted to look at the main and interaction effects of interest and frustration on clicks per query. The results of this analysis are detailed in Table 19. Again, there were no significant main effects for interest or significant interaction effects. There was a significant main effect for frustration.

Table 19

Results of ANOVA of Clicks Per Query Data

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	39.01	39	2.09	0.05	0.16
Frustration	158.01	39	11.18**	0.22	<0.01
Interest x Frustration	31.51	39	3.24	0.08	0.08

Note. ***p*<0.01.

4.7.1.4. Term Uniqueness for Queries. The corpus of queries was analyzed to determine term uniqueness. This was done using a script that computed the minimum IDF (inverse document frequency, a measure of how important a term or word is to a corpus), maximum IDF, average IDF, sum IDF, and the standard deviation of IDF. This analysis was done by attaching task number labels to each of the queries, then computing the IDF for that task by group (i.e. were the queries for this task unique in comparison to the other tasks for this task type?). Thus, the results allowed us to see whether the queries for the tattoo removal task for interesting and frustrating tasks were more or less unique than the queries submitted for the tattoo removal task for uninteresting and not frustrating tasks. Stop words were removed and stemming was used. Table 20 shows the means and standard deviations of the IDF for all task types for each task; this table shows that the IDF was fairly similar for tasks between all task types.

A two-way repeated measures ANOVA was conducted to see if there was any effect of interest and frustration on mean IDF, and there were no significant main effects found either for interest (*F*(1,7)=1.20, *η2*=0.15, *p*=0.31), or frustration (*F*(1,7)=1.38, *η2*=0.16 *p*=0.28). There was also no significant interaction effect (*F*(1,7)=0.63, *η2*=0.08, *p*=0.45).

Table 20

IDF Scores for All Task Types

Task	IF	NIF	INF	NINF
Tattoo Removal	5.90 (0.94)	5.94 (0.96)	6.15 (1.00)	6.07 (0.94)
Lupus	5.93 (0.89)	5.84 (1.21)	5.87 (1.04)	6.35 (0.60)
Energy Sources	5.77 (1.06)	NA	5.26 (1.17)	5.88 (1.08)
Biomass Fuels	5.65 (0.65)	5.94 (1.02)	6.03 (0.96)	5.63 (1.00)
Online Communication	6.08 (1.00)	6.05 (0.78)	6.23 (0.80)	6.02 (0.52)
Vehicle Purchases	5.68 (1.01)	5.52 (1.24)	5.07 (0.89)	5.85 (0.70)
Video Game Violence	6.19 (0.73)	5.89 (0.95)	6.74 (0.00)	5.50 (0.85)
Endurance Sports	5.89 (1.00)	6.27 (0.90)	6.15 (0.94)	5.95 (0.99)
Average	5.95 (0.15)	5.99 (0.39)	6.10 (0.24)	6.09 (0.22)

Note. Standard deviations appear in parentheses. IF=interesting and frustrating, NIF=not interesting and frustrating, INF=interesting and not frustrating, NINF=not interesting and not frustrating.

4.7.2. SERP

4.7.2.1. SERPs Displayed. SERPs displayed means when a participant submitted a query, clicked submit, and a SERP was displayed. As with other signals, the means and standard deviations of the signal based on interest, frustration, and the combination of the two were computed. There were slightly more SERPs displayed for interesting tasks than uninteresting tasks. There were also more SERPs displayed for frustrating tasks than not frustrating tasks. Table 21 describes the means and standard deviations of SERPs displayed by interest and frustration. Similarly to other signals, the most SERPs were displayed for interesting and frustrating tasks and loaded the fewest SERPs for uninteresting and not frustrating tasks.

Table 21

SERPs Displayed by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	15.90 (10.09)	9.70 (6.95)	12.80 (8.52)
	No	15.15 (8.70)	7.42 (4.42)	11.29 (6.56)
	Totals (Frustration)	15.52 (9.87)	8.56 (5.90)	

Note. Standard deviations in parentheses.

A two-way repeated measures ANOVA was conducted to examine the main and interaction effects of interest and frustration on SERPs displayed (Table 22), and while there were no significant main or interaction effects for interest, there was a significant main effect for frustration.

Table 22

Results of ANOVA for SERPs Displayed

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	91.51	39	1.83	0.04	0.18
Frustration	1939.06	39	39.72***	0.50	<0.001
Interest x Frustration	23.26	39	0.56	0.01	0.46

Note. ****p*<0.001.

4.7.2.2. Scroll Behavior on the SERP. The scroll behavior observed during this experiment represents movement of the wheel button (located in the middle of the mouse) on the SERP. There were more scrolls performed for interesting tasks than uninteresting tasks. There were also more scrolls performed for frustrating tasks than not frustrating tasks. As shown in Table 23, participants scrolled most on the SERP for interesting tasks and scrolled the least during uninteresting and not frustrating tasks.

Table 23

Scrolls on SERP by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	47.80 (39.08)	15.77 (16.11)	31.78 (27.59)
	No	34.50 (24.95)	11.55 (12.93)	23.02 (18.94)
	Totals (Frustration)	41.15 (32.01)	13.66 (14.52)	

Note. Standard deviation appears in parentheses.

Results of the two-way repeated measures' ANOVA test for main and interaction effects of interest and frustration on scrolls indicate that there was a significant main effect for interest (see Table 24), and a significant main effect for frustration, but there was no significant interaction effect.

Table 24

Results of ANOVA for Scrolls on SERP

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	3071.26	39	5.91*	0.13	0.02
Frustration	30222.51	39	54.45***	0.58	<0.001
Interest x Frustration	823.56	39	1.38	0.03	0.25

Note. *p<0.05, ***p<0.001.

4.7.2.3. Bookmarks. Participants bookmarked more documents for interesting tasks. Participants also bookmarked more documents for non-frustrating tasks than frustrating tasks. Table 25 shows that participants bookmarked the most documents for interesting and not frustrating tasks, and bookmarked the least amount of documents for interesting and frustrating tasks.

Table 25

Documents Bookmarked by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	3.07 (1.47)	4.12 (2.42)	3.60 (2.06)
	No	3.21 (1.34)	3.65 (2.20)	3.42 (1.83)
	Totals (Frustration)	3.14 (1.40)	3.88 (2.31)	

Note. Standard deviations appear in parentheses.

A two-way repeated-measures ANOVA (see Table 26) revealed a significant main effect for frustration but no significant main effect for interest or interaction effect.

Table 26

Results of ANOVA of Documents Bookmarked

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	1.22	39	0.41	0.01	0.52
Frustration	22.50	39	7.53**	0.16	0.01
Interest x Frustration	3.60	39	1.16	0.03	0.29

Note. ***p*<0.01.

4.7.3. Clicks. ***4.7.3.1. SERP Clicks.*** Clicks were defined as when a participant clicked on a link on the SERP or clicked on a link within a document (i.e., a webpage). SERP clicks were collected from the search log. The means and standard deviations for SERP clicks show that there were more SERP clicks for interesting tasks than uninteresting tasks. Participants also clicked more on the SERP for frustrating tasks than not frustrating tasks. Table 27 shows that participants clicked on the SERP the most during interesting and frustrating tasks, and clicked the least on the SERP during uninteresting tasks without a frustrator.

Table 27

SERP Clicks by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	23.65 (15.31)	15.42 (12.05)	19.52 (14.30)
	No	22.45 (15.96)	12.95 (9.03)	17.70 (13.75)
	Totals (Frustration)	23.04 (15.57)	14.19 (10.66)	

Note. Standard deviations appear in parentheses.

A two-way repeated measures ANOVA was conducted to examine the main and interaction effects of interest and frustration on clicks on the SERP (Table 28). There were no significant main effects for interest, and no significant interaction effect for interest and frustration. However, there was a significant main effect for frustration.

Table 28

Results of ANOVA of SERP Clicks

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	133.22	39	1.39	0.03	0.24
Frustration	3132.90	39	21.15**	0.35	<0.001
Interest x Frustration	16.90	39	0.12	0.00	0.73

Note. ****p*<0.001.

4.7.3.2. Clicks on Documents. Clicks on documents were gathered from Morae. Table 29 shows that participants clicked more on documents for interesting tasks than uninteresting tasks. Participants clicked more on documents for not frustrating tasks than frustrating tasks. Analyzing the means and standard deviations of clicks on documents by dimension (Table 29), participants had the most clicks on documents for interesting and frustrating, and had the fewest clicks on documents for uninteresting and not frustrating tasks.

Table 29

Clicks on Documents by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	53.40 (35.08)	40.87 (29.57)	47.14 (32.85)
	No	52.30 (38.77)	37.40 (30.25)	44.85 (35.36)
	Totals (Frustration)	52.85 (36.74)	39.14 (29.77)	

Note. Standard deviations are in parentheses.

A two-way repeated measures ANOVA was conducted to examine the main and interaction effects of interest and frustration on clicks on documents (Table 30). As with clicks on the SERP, there were no significant main effects for interest on clicks on documents, and no significant interaction effects for interest and frustration. There was a significant main effect for frustration on clicks on documents.

Table 30

Results of ANOVA of Clicks on Documents

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	209.306	39	0.31	0.01	0.58
Frustration	7521.306	39	7.07**	0.15	0.01
Interest x Frustration	56.41	39	0.08	0.00	0.78

Note. ***p*<0.01.

4.7.4. Time. Participants spent, on average, more time (in minutes) completing interesting tasks than uninteresting tasks (Table 31). Participants also spent more time on frustrating tasks than not frustrating tasks. As shown in Table 31, participants spent the longest time completing tasks that were both interesting and frustrating, and spent the least amount of time on uninteresting tasks that were not frustrating.

Table 31

Time Spent on Task by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	7.42 (3.95)	6.30 (3.69)	6.86 (3.84)
	No	6.80 (3.37)	4.97 (2.43)	5.88 (3.06)
	Totals (Frustration)	7.11 (3.66)	5.63 (3.17)	

Note. Standard deviations appear in parentheses.

Task time was examined for main and interaction effects of interest and frustration (see Table 32), and a significant main effect was found for frustration on task time and a nearly significant effect for interest. However, there were no significant interaction effects.

Table 32

Results of ANOVA for Total Task Time

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	38.02	39	3.66	0.09	0.06
Frustration	87.02	39	7.26**	0.16	0.01
Interest x Frustration	4.90	39	0.64	0.02	0.43

Note. ***p*<0.01.

4.7.4.1. Time on SERP. Time on the SERP was defined as time spent on SERP in minutes over the course of the whole task. Table 33 details the means and standard deviations for time spent on SERP by interest and frustration levels. Participants spent more time on the SERP for interesting tasks than uninteresting tasks. Participants also spent more time on the SERP for frustrating tasks than not frustrating tasks. Examining time spent on SERP by multiple dimensions (see Table 33), participants spent the most time on the SERP for interesting and frustrating tasks, and spent the least time on the SERP for uninteresting and not frustrating tasks.

Table 33

Time Spent on SERP by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	2.92 (2.02)	1.17 (0.99)	2.04 (1.80)
	No	2.50 (1.35)	0.99 (0.66)	1.74 (1.30)
	Totals (Frustration)	2.71 (1.71)	1.08 (0.84)	

Note. Standard deviations appear in parentheses.

A two-way repeated measures ANOVA was completed to look for significant main and interaction effects for interest and frustration; the results of this analysis are summarized in Table 34. A significant main effect was found for frustration on time spent on the SERP, but no significant main effect for interest or significant interaction effect were found.

Table 34

Results of ANOVA of Time Spent on SERP

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	3.21	39	2.00	0.05	0.17
Frustration	95.29	39	56.13***	0.62	0.00
Interest x Frustration	0.56	39	0.33	0.01	0.57

Note. ****p*<0.001.

4.7.4.2. Time Spent on Documents. Participants spent slightly more time on documents for interesting tasks than uninteresting. Participants also spent slightly more time on documents for frustrating tasks than not frustrating tasks. Analysis of the means and standard deviations of time spent on documents by interest and frustration level (Table 35) shows that (like many other signals) participants spent the most time on documents during interesting and not frustrating tasks, and spent the least time on uninteresting and not frustrating tasks.

Table 35

Time Spent on Documents by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	2.64 (2.29)	2.81 (2.23)	2.72 (2.25)
	No	2.68 (2.04)	1.98 (1.45)	2.33 (1.79)
	Totals (Frustration)	2.66 (2.16)	2.40 (1.91)	

Note. Standard deviations appear in parentheses.

A two-way repeated measures' ANOVA was conducted to examine the main and interaction effects of interest and frustration on time spent on documents. As shown in Table 36, there were no significant main or interaction effects for time spent on documents.

Table 36

Results of ANOVA on Time Spent on Documents

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	5.72	39	2.07	0.06	0.16
Frustration	2.36	39	0.63	0.02	0.43
Interest x Frustration	6.80	39	2.69	0.07	0.11

4.7.4.3. Query Time Intervals. Query time intervals were defined as the time between queries for each task. This was done to investigate whether there were shorter times between queries for different types of tasks, which would indicate more rapid query reformulation. There were greater times (in minutes) between queries for interesting tasks than uninteresting tasks. There was also slightly more time spent between queries for frustrating tasks than not frustrating tasks. Table 37 details the means and standard deviations for time between queries for tasks by interest and frustration. Participants spent the longest time between queries for interesting and frustrating tasks, and spent the shortest time between queries for uninteresting and not frustrating tasks.

Table 37

Query Time Intervals (in Minutes) by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	5.65 (3.94)	4.40 (3.36)	5.02 (3.70)
	No	3.06 (3.57)	2.84 (3.57)	2.96 (3.55)
	Totals (Frustration)	4.36 (3.95)	3.62 (3.53)	

Note. Standard deviations appear in parentheses.

A two-way repeated measures ANOVA was conducted to look for main and interaction effects of interest and frustration on time spent between queries (Table 38). The results of this ANOVA show a significant main effect for interest, but no significant main or interaction effects for frustration.

Table 38

Results of ANOVA of Query Time Intervals

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	169.62	39	19.30***	0.33	<0.001
Frustration	22.44	39	2.91	0.07	0.10
Interest x Frustration	10.14	39	0.89	0.02	0.35

Note. ****p*<0.001.

4.7.5. Queries Without Clicks

Queries without clicks were isolated via extraction from the database. Any query that was not associated with a click of any kind was identified. There were a total of 860 queries submitted, and of these queries 257 were submitted without a click. Participants submitted more queries without a click for interesting tasks (*M*=1.71, *SD*=2.51) than uninteresting tasks (*M*=1.47, *SD*=2.51). Participants also submitted more queries without a click for frustrating tasks (*M*=2.89, *SD*=2.96) than not frustrating tasks (*M*=0.30, *SD*=0.70). Participants submitted the most queries without a click for interesting and frustrating tasks (*M*=3.15, *SD*=2.83), followed by uninteresting and frustrating tasks (*M*=2.62, *SD*=3.10). Participants submitted fewer clicks for uninteresting and not frustrating tasks (*M*=0.32, *SD*=0.69) and interesting and not frustrating tasks (*M*=0.27, *SD*=0.71). A repeated measures' ANOVA was conducted on these data, and a significant main effect was found for frustration ($F(1,39)=57.09$, $\eta^2=0.59$, $p<0.001$). There were no significant main effects for interest ($F(1,39)=0.44$, $\eta^2=0.01$, $p=0.51$) and no significant interaction effects (F(1,39)=0.71, $\eta^2=0.02$, $p=0.40$).

4.7.6. Multiple Regression Analysis of Search Behavior

A linear regression approach was taken to further investigate which behavioral signals were most predictive of engagement and frustration, using the lm command in R to fit a model.

First, engagement and search behavior signals were investigated. The signals included in the regression were chosen because they are commonly accepted measures of search behavior, and have also been used in other studies that seek to understand whether participants are struggling or exploring (Hassan et al., 2014). The analysis was set up as follows: engagement scores were averaged for each participant for each task (i.e., they were not divided by subscale). Then, the scores for each variable for each task were also added to the data file. The regression began with queries and progressively added each signal in the temporal order the participant completed them, as other work has shown that differences in the temporal order of queries can be indicative of struggling or exploring during a search session (Hassan et al., 2014). Therefore, the signals were included as follows: queries (a participant queries first), SERPs displayed (a SERP is then loaded), scrolls (a participant scrolls on the page with the SERP), clicks (a participant clicks a SERP result), followed by documents (a participant clicks within a document). Bookmarks were excluded from this analysis because they are an artifact of this experiment and not likely to occur in a naturalistic setting, and thus would not useful for automating prediction of engagement or frustration. The coefficients for each predictor and multiple R-squared of each model are reported in Table 39.

The coefficients represent the change in one variable in response to another, i.e., the change in the predictor variable in relationship to the dependent variable. The multiple R-squared value describes the amount of variation in the response that is explained by the least squares line, and so represents the strength of linear fit of the model. Therefore, a higher R-squared represents a model with better fit. An ANOVA was conducted to compare each regression to the one before it, to see if the added predictor made the model significantly different than the one before it. The ANOVA showed that the only model that showed any significant improvement was Model 5 (the model with queries, SERPS, scrolls, and SERP clicks). This model significantly improved (*F*(1, 155)=6.17, *p*<0.01) over Model 4 (the model with only queries, SERPs, and scrolls). In Model 5, SERP clicks were the significant predictor (*t*(159)=2.48, *p*<0.05).

Frustration and search actions were also modeled using the same analysis. Frustration scores were averaged for each participant for each task. The coefficients for each predictor and multiple R-squared of each model are reported in Table 40. An ANOVA was also run to compare these regression models. The model improvements for this regression were mixed. Model 3 (the model with queries, SERPs, and scrolls) was a significant improvement (*F*(1, 156)=10.60, *p*<0.001) over Model 2 (with only queries and SERPs). In Model 3, queries were a significant predictor of frustration (*t*(159)=3.44, *p*<0.001), as well as scrolls (*t*(159)=3.18, *p*<0.01). Model 5 (with queries, SERPs, scrolls, SERP clicks and clicks on documents) was a significant improvement (*F*(1,154)=6.84, *p*<0.01) over Model 4 (which had queries, SERPs, scrolls, and SERP clicks). In Model 5, there were several significant predictors of frustration: queries (*t*(159)=0.05, *p*<0.001), scrolls (*t*(159)=0.004, *p*<0.01), SERP clicks (*t*(159)=0.01, *p*<0.01), and clicks on documents (*t*(159)=0.004, *p*<0.01). The evolution of these models indicates that queries, scrolls, and clicks (of each type) may be the most helpful predictors of frustration, and given that SERP clicks were also significant predictors of engagement, it may be useful to note the directionality of this signal when disambiguating engagement and frustration.

Table 39

Summary of Multiple Regression Analysis for Behavioral Signals Predicting Engagement

	Model 1		Model 2		Model 3		Model 4		Model 5	
Variable	β	R^2	β	R^2	β	R^2	β	R^2	β	R^2
Queries	0.28	0.02	-0.06	0.03	-0.12	0.03	-0.16	0.07	-0.19	0.08
SERPs			0.21		0.18		0.11		0.14	
Scrolls					0.03		0.02		0.01	
SERP clicks							0.14		0.05*	
Clicks on Docs									0.04	

Note. *p<0.05.

Table 40

Summary of Multiple Regression Analysis for Behavioral Signals Predicting Frustration

	Model 1		Model 2		Model 3		Model 4		Model 5	
Variable	β	R^2	β	R^2	β	R^2	β	R^2	β	R^2
Queries	0.05***	0.25	0.05***	0.25	0.05***	0.30	0.05***	0.31	0.05***	0.36
SERPs			-0.00		-0.00		-0.00		-0.01	
Scrolls					0.00**		0.00**		0.00**	
SERP clicks							0.00		0.01**	
Clicks on Docs									-0.07**	

Note. **p<0.01, ***p<0.001.

4.7.7. Logistic Regression of Search Behavior. A logistic regression was performed to see if, when forced into a binary outcome, engagement and frustration could be predicted by behavioral signals. Since engagement and frustration were not already binary, a threshold needed to be established by which to indicate whether a participant was engaged or not, or frustrated or not. Both scales were constructed as a 5-point Likert-type scale, so a cutoff point was established at 3 to indicate engagement or frustration. This cutoff point represented a "middle ground" of sorts, i.e. a score of 3 indicated that participants did not feel strongly regarding engagement or frustration. Participants were given a score of 0, meaning "not engaged" or "not frustrated" if their engagement or frustration score was below 3, and were given a score of "1" if their engagement or frustration score was above 3. Participants who had a score of 3 were discarded; five participants had their engagement scores discarded, while eight participants had their frustration scores discarded. This resulted in engagement scores for 155 participants and frustration scores for 152 participants. This setup created a bivariate logistic regression with two dichotomous dependent variables, which was computed using the Zelig package (Imai, King & Lau, 2015). Two separate logistic regressions were run using engagement and frustration as two dichotomous dependent variables, and the five search behavior signals used in the earlier regressions were included as predictors. Therefore, the results showed us the strength of the predictor for each state.

The results of this regression confirm that several signals are more predictive of frustration than engagement. Table 41 shows that queries, scrolls, and clicks on the SERP were more predictive of frustration than engagement. This analysis seems to confirm that it is difficult to distinguish frustration and engagement, as there are no significant predictors for engagement, and the significant predictors for frustration all have very low coefficients. This analysis is also hampered by the forced creation of a binary of both variables (i.e., forcing engagement and frustration into “yes” or “no” conditions); though in this experiment engagement and frustrated are situated as opposites, it’s possible to exist in a middle ground, or to be neither engaged or frustrated or both, and so by constructing these as binary variables some of the nuance in both (that could help strengthen their relationship with predictor variables) is lost.

Table 41

Results of Logistic Regression of Engagement, Frustration, and Search Actions

Dependent Variable	Predictor	Coefficient	*z*	*p*
Engagement	Queries	0.03	0.42	0.67
	SERPs	0.01	0.35	0.65
	Scrolls	0.00	0.37	0.71
	SERP Clicks	0.04	1.67	0.09
	Clicks on Docs	-0.01	-0.73	0.46
Frustration	Queries	0.18	2.00*	0.05
	SERPs	-0.04	-0.88	0.38
	Scrolls	0.06	3.95***	<0.001
	SERP clicks	0.06	2.19*	0.05
	Clicks on Docs	-0.02	-1.37	0.17

Note. **p*<0.05, ****p*<0.001.

4.7.8. Summary of Search Behavior. Table 42 details a summary of selected search actions. In general, participants completing frustrating tasks performed more search actions (such as submitting more queries and clicks) than participants who completed not frustrating tasks. Also, participants generally completed more search actions for tasks that were interesting than tasks that were not interesting. However, many more significant main effects were found for frustration, which suggests that frustration has a greater effect on search behavior than interest. This fact was confirmed both in the multiple and logistic regressions of engagement and frustration. There were more significant predictors for frustration than engagement in both of these analyses, indicating that it may be easier to identify frustration than engagement with regards to search behavior.

Table 42

Summary of Selected Search Behavior Results by Interest and Frustration Levels

	Interest		Frustration		Combinations			
Item	Yes	No	Present	Absent	IyFy	IyFn	InFy	InFn
Queries	5.65 (4.45)	5.06 (4.13)	7.51 (4.66)	3.20 (2.42)	7.87 (4.73)	3.42 (2.73)	7.15 (4.62)	2.97 (2.08)
Clicks Per Query	7.77 (4.77)	6.79 (3.97)	8.27 (5.03)	6.29 (3.42)	8.32 (5.32)	7.22 (4.16)	8.22 (4.79)	5.35 (2.15)
SERPs displayed	13.80 (9.16)	12.29 (7.88)	15.52 (9.37)	8.56 (5.90)	15.90 (10.09)	9.70 (6.95)	15.15 (8.70)	7.42 (4.42)
SERP clicks	19.52 (14.30)	17.70 (13.75)	23.04 (15.57)	14.19 (10.66)	23.65 (15.31)	15.42 (12.05)	22.45 (15.96)	12.95 (9.03)
Clicks on Docs	47.14 (32.85)	44.85 (35.36)	39.14 (29.77)	52.85 (36.74)	53.40 (35.08)	52.30 (38.77)	40.87 (29.57)	37.40 (30.25)
SERP Time	2.04 (1.80)	1.74 (1.30)	2.71 (1.71)	1.08 (0.84)	2.92 (2.02)	1.17 (0.99)	2.50 (1.35)	0.99 (0.66)
Time on Docs	2.72 (2.25)	2.33 (1.79)	2.66 (2.16)	2.40 (1.91)	2.64 (2.29)	2.81 (2.23)	2.68 (2.04)	1.98 (1.45)
Scrolls	31.78 (27.59)	23.02 (18.94)	41.15 (32.01)	13.66 (14.52)	47.80 (39.08)	15.77 (16.11)	34.50 (24.95)	11.55 (12.93)
Query Interval (in mins)	5.02 (3.70)	2.96 (3.57)	4.36 (3.95)	3.63 (3.53)	5.65 (3.94)	4.40 (3.36)	3.06 (3.57)	2.84 (3.57)

Note. Standard deviations appear in parentheses. Y indicates a variable is present, N indicates it is absent. Gray selections indicate significance.

4.8. Physiological Data

4.8.1. Overview. Physiological data from only 39 participants is reported because of a logging failure. Data was gathered from the Biopac at a rate of 1sample/msec. There were four channels programmed for data collection: EDA, ECG, Rate (which extracted heart rate from the raw ECG signal) and Switch (used for task start and stop annotation). During the experiment, the following parameters were set for each signal: 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. The skin conductance data is 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 35 Hz, with a gain of 1000Hz. Heart rate was extracted from the raw ECG signal and is reported in bpm (beats per minute). A three-minute baseline was recorded for each participant (EDA: *M*=6.45 μS, *SD*=4.31, HR: *M*=85.70 bpm, *SD*=16.10).

Heart rate and skin conductance data were collected from each participant for each task. The beginning and end of each task were manually annotated using the BioPac switch during the experiment. The beginning of a task was defined as when the participant clicked “start task” and the end of a task was defined as when the participant clicked “end task.” After the experiment, these markers were double-checked for accuracy by comparing them to the timestamps logged in the search system. Through this, length of task, and accuracy of task endpoints was confirmed for each task for each participant. Both the skin conductance and heart rate data were visually inspected for any errors or potential aberrant artifacts. In addition to this, a Shannon entropy analysis was conducted to look for aberrations or potential artifacts in the data as was done in Barreda-Angeles et al., (2015). 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. It is difficult to say what high entropy values in physiological data are, but these values and examination of the data seem to indicate that there is very little aberrant fluctuation in either signal. Once these points were confirmed, the means and standard deviations for each physiological signal were computed.

4.8.2. Analysis of Physiological Data. The graphs below represent an overview of the data by task type over the course of the entire task. To aid in interpretation of the figures, the following key will be used: IF = Interesting tasks with a frustrator; NIF = uninteresting tasks with a frustrator; INF = Interesting tasks without a frustrator; NINF = uninteresting tasks without a frustrator. These graphs help us understand the shape of the data for each signal over the course of the entire task. Figure 9 shows the graph of the skin conductance data for all users over the entire task (in seconds) for each task type. Figure 10 shows a graph of the heart rate data (in beats per minute) for all participants for the entire task. These figures show what is confirmed in the data – while heart rate is quite stable (i.e., there is very little variability) there is much greater variability in the skin conductance signal. In fact, the skin conductance signal for interesting tasks with a frustrator seems to be the most highly variable, especially in the beginning of the task.

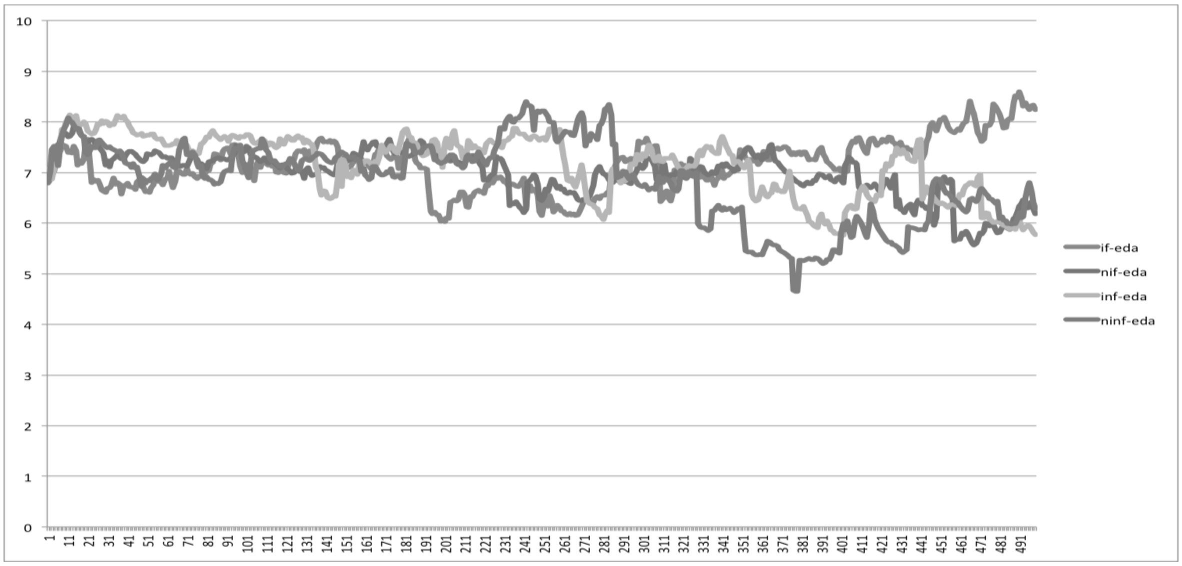


Figure 9. Skin conductance data (in seconds) for all participants for all tasks by task dimensions.

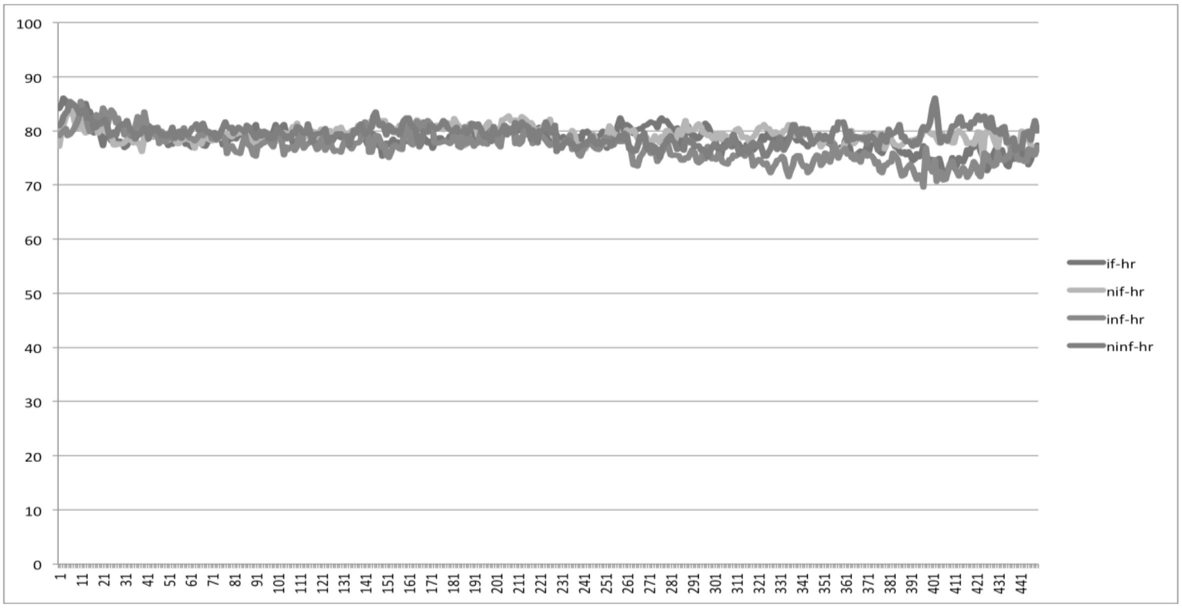


Figure 10. Heart rate data (in seconds) for all participants for all tasks by all task dimensions.

Fine-grained analysis of the data was performed, as these measures often yield more information about the data than coarser analyses. The fine-grained analysis was completed as follows: since the data was recorded at 1000 samples a second, each 1000-sample interval was averaged to get the electrodermal and heart rate values for that second. Though participants had varying task times, given that task times tended to range between five to seven minutes, 420 seconds' (seven minutes) worth of samples were collected for each participant for each task type to capture all possible signal values (skin conductance and heart rate) for the task. Where participants completed tasks early (meaning they did not have seven minutes worth of data), NA was entered as a placeholder. Means and standard deviations were computed for each signal for each dimension (interest and frustration) as well as the combinations of the two.

Table 43 summarizes the means and standard deviations of the skin conductance values by interest and frustration. Participants had similar skin conductance values for interesting tasks and uninteresting tasks. Slightly higher levels of skin conductance were experienced during frustrating tasks than not frustrating tasks. As shown in Table 43, participants experienced the highest skin conductance during interesting and frustrating tasks, and experienced the lowest levels of skin conductance during interesting and not frustrating tasks. Table 43 also shows that participants experienced the highest heart rate during uninteresting and not frustrating tasks, and the lowest during interesting and frustrating tasks.

Table 43

Skin Conductance Results by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	7.33 (4.37)	7.05 (4.81)	7.19 (4.57)
	No	7.12 (4.71)	7.18 (4.81)	7.15 (4.73)
	Totals (Frustration)	7.22 (4.52)	7.12 (4.78)	

Note. Standard deviations are in parentheses.

A two-way repeated measures' ANOVA was conducted to look for main and interaction effects of interest and frustration on skin conductance. Table 44 details the results of this ANOVA; there were no significant main or interaction effects for these data.

Table 44

Results of ANOVA of Skin Conductance Data

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	0.08	38	0.02	0.00	0.89
Frustration	0.48	38	0.12	0.00	0.73
Interest x Frustration	1.13	38	0.22	0.01	0.64

Table 45 summarizes the means and standard deviations of the heart rate value by interest and frustration. The heart rate data means and standard deviations show a similar trend to the skin conductance data. Participants had similar heart rate values for interesting and uninteresting tasks. Participants experienced slightly higher heart rate for frustrating tasks than not frustrating tasks. Participants had the greatest heart rate for uninteresting and not frustrating tasks, and the lowest heart rate for interesting and frustrating tasks.

Table 45

Heart Rate Results by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	79.10 (16.31)	80.03 (17.46)	79.57 (16.79)
	No	79.65 (16.55)	80.10 (15.72)	79.88 (16.03)
	Totals (Frustration)	79.38 (16.33)	80.07 (16.50)	

Note. Standard deviations appear in parentheses.

Another two-way repeated measures ANOVA was conducted to examine the main and interaction effects of interest and frustration for heart rate data. Table 46 summarizes the results of this ANOVA. As shown in Table 46, there were no significant main or interaction effects for heart rate data.

Table 46

Results of ANOVA of Heart Rate Data

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	3.72	38	0.24	0.01	0.63
Frustration	18.52	38	2.93	0.07	0.09
Interest x Frustration	2.20	38	0.24	0.01	0.63

4.8.3. Window Analysis. An additional window analysis was performed to examine the effect of interest and frustration on skin conductance and heart rate data during different windows of time. Window analyses are commonly performed on physiological data, which is time-sensitive and can fluctuate between periods of time. Other studies have used window analysis to observe changes in physiological signals (Feild et al., 2010; O'Brien & Lebow, 2013). The window analysis was done by splitting the data along ten-second intervals, and averaging that window for each participant for each task type - this served to smooth the data. This resulted in 42 ten-second-time windows for each task type (interesting and frustrating, not interesting and frustrating, etc.). Table 47 details the results of the window analysis of the skin conductance data. Similar to the overall analysis, the means show that participants had greater skin conductance during interesting tasks than uninteresting tasks, and greater skin conductance during not frustrating tasks than frustrating tasks.

Table 47

Window Analysis of Skin Conductance Results by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	7.04 (0.37)	7.26 (0.53)	7.15 (0.47)
	No	7.11 (0.28)	6.94 (0.80)	7.03 (0.60)
	Totals (Frustration)	7.08 (0.33)	7.10 (0.70)	

Note. Standard deviations appear in parentheses.

A two-way repeated measures ANOVA was conducted to look at the effect of interest and frustration on these windows of time for skin conductance data. As shown in Table 48, there were no significant main effects for interest and frustration, but there was a significant interaction effect. Post-hoc investigation of the interaction effect with Bonferroni correction showed significant differences between the two interesting tasks and the other tasks at *p*<0.05.

Table 48

Results of ANOVA of Window Analysis of Skin Conductance Data

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	0.61	41	3.68	0.82	0.62
Frustration	0.02	41	0.05	0.00	0.83
Interest x Frustration	1.70	41	14.45***	0.26	<0.001

Note. ****p*<0.001.

Table 49 shows the results of the window analysis for heart rate results by interest and frustration. Participants had greater heart rate for not interesting tasks, and had similar heart rate levels for frustrating and not frustrating tasks. Participants also had the highest levels of heart rate for uninteresting and not frustrating tasks, and had the lowest heart rate levels for interesting and not frustrating tasks.

Table 49

Window Analysis of Heart Rate Results by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	78.42 (1.64)	76.97 (2.48)	77.70 (2.21)
	No	79.34 (1.05)	79.37 (1.38)	79.36 (1.22)
Totals (Frustration)		78.88 (1.45)	78.17 (2.33)	

Note. Standard deviations appear in parentheses.

Another two-way repeated-measures ANOVA was conducted to examine the main and interaction effects of interest and frustration on window analysis of the heart rate data; this analysis is summarized in Table 50. As shown in Table 50, there were significant main effects for interest and frustration and a significant interaction effect. Further contrasts showed significant differences (*p*<0.01) in heart rate between all task types.

Table 50

Results of ANOVA of Window Analysis of Heart Rate Data

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	116.03	41	35.97***	0.47	<0.001
Frustration	21.12	41	9.48***	0.19	<0.001
Interest x Frustration	23.29	41	29.12***	0.41	<0.001

Note. ****p*<0.001.

4.8.4. Skin Conductance Responses. The skin conductance data alone was examined to determine the number of skin conductance responses for each participant for each task type. Skin conductance responses (SCRs) are characterized by a sharp increase in electrodermal response followed by a decrease in response to the stimulus, usually involving an increase of one or more microsiemens (Boucsein, 2012). Inspecting the data visually and identifying points that matched the required characteristics determined skin conductance responses. Each participant had between 0 (meaning no distinct change in skin conductance) to 15 SCRs per task. Computation of means and standard deviations for each of these dimensions (Table 51) shows that more skin conductance responses occurred during interesting tasks than uninteresting tasks. There were also more skin conductance responses during frustrating tasks than not frustrating tasks. Table 51 details the means and standard deviations for tasks by combination of interest and frustration dimension. We see that interesting and frustrating tasks had the highest number of skin conductance responses, and uninteresting tasks without a frustrator had the lowest amount of skin conductance responses.

Table 51

Skin Conductance Responses by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	5.54 (4.72)	2.97 (3.13)	4.50 (4.18)
	No	4.05 (3.53)	2.84 (2.20)	3.45 (2.99)
	Totals (Frustration)	4.79 (4.21)	2.91 (2.71)	

Note. Standard deviations are in parentheses.

A two-way repeated measures ANOVA was conducted to look at main and interaction effects for interest and frustration on skin conductance responses. The results of this analysis are shown in Table 52. Significant main effects were found for interest and frustration, but there were no significant interaction effects.

Table 52

Results of ANOVA of Skin Conductance Response Data

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	20.10	38	4.12*	0.10	0.05
Frustration	144.23	38	16.40***	0.30	<0.001
Interest x Frustration	23.08	38	2.73	0.07	0.11

Note. **p*<0.05, ****p*<0.001.

4.8.5. Changes in Physiological Data in Initial Task Stages. The data was also examined for changes within the first 60 seconds of the task, which is where the stimulus is likely to have the most dramatic effect (Boucsein, 2012). The first 60 seconds of the task was defined in this study as the first 60 seconds after the participant clicked “start task”. This time frame was examined for valence of change from the beginning of the task, which would indicate a skin conductance response. This data was gathered by averaging the data for each participant for each second of the first 60 seconds (for example, an average was computed for second 1, second 2, etc.). Examination of the means and standard deviations for these data showed that interesting and uninteresting tasks had similar skin conductance levels. Valence of change was also computed for these data (Table 53). Valence of change refers to whether the change in skin conductance (or heart rate) was positive or negative from the beginning of the task to the 60-second value. Therefore, valence was calculated by subtracting the value at 0 seconds from the value at 60 seconds. The valence then represents a measure of change from 0 to 60 seconds. Uninteresting tasks had a greater positive change than interesting tasks. Participants experienced higher skin conductance during not frustrating tasks than frustrating tasks. There was greater positive change for not frustrating tasks than frustrating tasks.

Table 53

Skin Conductance Values and Valence of Change for First 60 Seconds by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	6.96 (0.32)	7.80 (0.29)	7.38 (0.30)
		+0.07	*+0.53*	*+0.60*
	No	7.43 (0.23)	7.36 (0.36)	7.39 (0.31)
		+0.30	*+0.29*	*+0.59*
	Totals (Frustration)	7.20 (0.37) *+0.37*	7.58 (0.40) *+0.82*	

Note. Valence of change appears in italics. Standard deviations appear in parentheses.

Table 54 details the results of a two-way repeated measures ANOVA of the skin conductance data. No significant main effect for interest was found, but a significant main effect was found for frustration, and a significant interaction effect for interest and frustration was also found. Investigation of the interaction showed significant differences between the two frustrating tasks and all other tasks at p<0.001.

Table 54

Results of ANOVA of Skin Conductance Data for First 60 Seconds

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	0.01	59	0.59	0.01	0.45
Frustration	8.50	59	128.65***	0.69	<0.001
Interest x Frustration	12.26	59	156.41***	0.73	<0.001

Note. ****p*<0.001.

Table 55 summarizes the heart rate values by interest and frustration. Participants had similar heart rate during interesting and uninteresting tasks. There was greater negative change in heart rate for interesting tasks (-6.53 bpm) than uninteresting tasks (-0.28 bpm). There was slightly lower heart rate for frustrating tasks than not frustrating tasks. There was also greater negative change for frustrating tasks (-5.04 bpm) than not frustrating tasks (-1.77 bpm). This table also shows that interesting and not frustrating tasks had the greatest positive change for skin conductance data, while interesting and frustrating tasks had the lowest positive change in skin conductance. This table shows that interesting and frustrating tasks had the greatest negative change in heart rate, while uninteresting and frustrating tasks had the lowest negative change in heart rate. Uninteresting and frustrating tasks had the only positive change in heart rate out of all task types.

Table 55

Heart Rate Values and Valence of Change for First 60 Seconds by Interest and Frustration Levels

		Frustration		
		Present	Absent	Totals (Interest)
Interest	Yes	80.14 (2.66)	80.72 (1.76)	80.43 (2.26)
		-6.46	*-0.07*	*-6.53*
	No	79.42 (1.72)	80.92 (1.78)	80.17 (1.90)
		+1.41	*-1.70*	*-0.28*
	Totals (Frustration)	79.78 (2.26) *-5.04*	80.82 (1.77) *-1.77*	

Note. Valence of change appears in italics. Standard deviations appear in parentheses.

A two-way repeated measures ANOVA was conducted on the heart rate data; the results are reported in Table 56. A significant main effect for frustration was found, and a significant interaction effect was also found. Post-hoc investigation of the interaction effect with Bonferroni correction showed that all tasks were significantly different from one another at *p*<0.001.

Table 56

Results of ANOVA of Heart Rate Data for First 60 Seconds

Source	*SS*	*df*	*F*	*η2*	*p*
Interest	4.15	59	2.63	0.04	0.11
Frustration	64.46	59	27.56***	0.32	<0.001
Interest x Frustration	12.57	59	4.26*	0.07	0.04

Note. **p*<0.05, ****p*<0.001.

4.8.6. Mixed-Effects Multilevel Models of Physiological Data. A mixed-effects multilevel growth model was completed 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 (Barreda-Angeles et al., 2015). Thus, the data set was constructed as follows: skin conductance data was entered for each participant for 60 seconds, 120s, 180s, 240s, 300s, 360s, and 420s. This was to get an accurate representation of the skin conductance over the course of the task (or the first seven minutes, as most tasks were within the four to seven minute range). Where participants were missing data, (i.e., they had short task times) NA was entered as a placeholder. Interest was coded as 1 = interesting and 0 = uninteresting, and frustration was coded the same way. Task order (the task position each task type was delivered in for each participant) was also added as a column. Lastly, participant identification numbers and task time (in minutes) for each participant were entered as columns. These data were transformed such that the data were ordered by time, i.e., 60s was time 1, 120s was time 2, etc.

Mixed multilevel models are made of both fixed and random effects. Fixed effects refer to variables where the possible values are fixed, while random effects refer to those for which the set of possible values can vary (Starkweather, 2010). Thus, in this experimental setup, interest, frustration, and task order are fixed effects, while participant number and task time are random effects. The models were created with guidance from the procedures outlined in Bliese (2013).

There are also several other components of multilevel models. Bayesian Information Criterion (BIC) values represent an indication of the fit of model, and lower values represent a better fit. Phi values are a measure of autocorrelation estimates; more specifically, this is the measure of the similarity between observations as a function of the time lag between them (Bliese, 2013). Lastly, there is the log likelihood ratio, or L-ratio. 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 (Starkweather, 2010). 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.

First, a null model was created to examine the properties of the electrodermal data. The null model (Model 0 in Table 56) showed that approximately ten percent of the variance in electrodermal activity was due to the individuals. The second model (Model 1 in Table 56) placed task time as a random effect (as task time varied between participants). This model assumes that the relationship between task time and electrodermal activity is not constant for all participants. This model was compared with the null model (where task time was not placed as a random effect) and the random effect was found to significantly improve the null model. Fixed effects (interest, frustration, and task order) were added to the model progressively. These are considered fixed effects because these did not vary randomly among participants (Starkweather, 2010). The IntraClass Correlation (ICC) was computed as a measure of reliability as it compares the variability of different values of the same participant to the total variability across all values and participants.

Each model was compared to the model before it to see whether adding a fixed effect significantly improved the model. The models were compared using an ANOVA. Before comparison, the models were refitted from REML (Restricted Estimated Maximum Likelihood) to ML (Maximum Likelihood) in order to perform the ANOVA (Reid, 2015). As stated earlier, the result of this ANOVA is an l-ratio, which represents the test statistic produced when comparing the two models; the l-ratios listed in Table 57 represent the test statistic produced when an ANOVA compared a given model to the one before it. There is one exception to this;

the models with interest alone and frustration could not be compared because they have different fixed effects, where the other models have added fixed effects. When comparing models, we confirm that the model with task time as a random effect significantly improved the model, but as fixed effects are added, the l-ratio decreases, as does significance.

All models were significantly different from the null model. Overall, the best fitting model was Model 4, which only had interest and frustration as fixed effects. We see that the BIC was fairly high for this model, but this model fit the data significantly better than the models with only interest or frustration. Model 4, when examined, showed a significant effect for interest (*p*<0.05), while there was no significant effect for frustration (*p*=0.24). The model with interest alone had a nearly significant p-value (*p*=0.06), indicating that some combination of interest and frustration contributed to the skin conductance scores.

Table 57

Fixed and Random Effects, Correlation Coefficients, BIC values, and L-Ratios for Models of Skin Conductance Data

Model	Fixed Effects	Random Effects	Coefficients	BIC	ICC	L-ratio
0	Intercept			3673.61	0.00	
1	Intercept	Task Time		3600.05	0.04	36.45***
2	Interest	Task Time	-0.51	3604.10	0.04	3.49
3	Frustration	Task Time	-0.10	3607.89	0.04	
4	Interest + Frustration	Task Time	I: -0.63 F: -0.26	3610.70	0.04	4.59*
5	Interest + Frustration + Task Order	Task Time	I: -0.59 F: -0.26 TO: -0.19	3616.67	0.04	3.56

Note. **p*<0.05, ****p*<0.001. I=Interest, F=Frustration, and TO=Task Order.

The same multilevel model analysis was also conducted on the heart rate data (Table 58). The data was prepared in a similar time-ordered fashion as with the skin conductance data. As with the model for skin conductance data, the null model was explored, fitted with task time as a random effect, and had the same fixed effects introduced progressively. Once again, these models were refitted from REML to ML in order to run the ANOVA. This analysis again showed that task time as a random effect significantly improved the model. All models were also significantly different from the null model, but in comparing models as fixed effects were added, the model with interest, frustration, and task order (Model 5) was significantly different from the model with only interest and frustration (Model 4). Thus, the best fitting model (and the model with the lowest BIC) was the model with interest, frustration, and task order (Model 5). In both Models 4 and 5 interest was non-significant while frustration was significant (*p*<0.05), indicating that frustration contributed more to heart rate than interest.

Table 58

Fixed and Random Effects, Correlation Coefficients, BIC values, and L-Ratios for Models of Heart Rate Data

Model	Fixed Effects	Random Effects	Coefficients	BIC	ICC	L-ratio
0	Intercept			4787.40	0.94	
1	Intercept	Task Time		4785.49	0.01	15.58***
2	Interest	Task Time	0.32	4791.64	0.01	0.61
3	Frustration	Task Time	-0.70	4788.49	0.01	
4	Interest + Frustration	Task Time	I: 0.15 F: -0.67	4795.10	0.01	0.12
5	Interest + Frustration + Task Order	Task Time	I: 0.24 F: -0.72 TO: -1.08	4755.31	0.00	48.96***

Note. ****p*<0.001. I=Interest, F=Frustration, and TO=Task Order.

4.8.7 Regression Models with Physiological Signals. Physiological signals were added to the regression of search actions to see if they added predictive value above and beyond a model with search actions alone. This presents a challenge, as the physiological data is time-ordered, and important differences can be smoothed in a simple average. However, the physiological data from different time points was isolated, and then each time point of data (i.e. data at 60 seconds, data at 120 seconds, data at 180 seconds) was added to see if this improved the prediction of engagement and frustration. As a methodological note, the models were fitted in R to run the regression despite having missing data values (which occurred with the physiological data, as people finished the tasks at different times) The results of this analysis are summarized below in Tables 59 and 60. In this analysis, I began with the final model explored in the regression, the model with queries, scrolls, SERPs, SERP clicks, clicks on documents, and bookmarks (see Tables 39 and 40). Therefore, all the models discussed below have already included the search actions from the earlier model.

Table 59

Multiple Regression Analysis of Search Actions and Physiological Signals Predicting Engagement

	Model 1		Model 2		Model 3*		Model 4		Model 5		Model 6	
	β	R^2	*β*	R^2	*β*	R^2	*β*	R^2	*β*	R^2	*β*	R^2
EDA60	0.84	0.05	0.81	0.07	0.06	0.09	0.07	0.08	0.06	0.08	0.06	0.07
HR60			0.05		0.03*		0.35		0.38		0.37	
EDA120					0.05*		0.05		0.05		0.05	
HR120							0.76		0.82		0.75	
EDA180									0.45		0.45	
HR180											0.79	

Note. *p<0.05

The results of this analysis show that some physiological signals were significant in the regression models. Specifically, heart rate within the first 60 seconds and skin conductance in the first 120 seconds were significant in the only model that was significantly different from all the rest, Model 3. This is useful because it shows that the while many physiological changes occur in the first 60 seconds of the task, it is possible that a slightly larger window (of 120 seconds) could also offer useful information with regards to disambiguating engagement and frustration. Table 60 details the results of the regression analysis for frustration.

Table 60

Multiple Regression Analysis of Search Actions and Physiological Signals Predicting Frustration

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	β	R^2	β	R^2	β	R^2	β	R^2	β	R^2	β	R^2
EDA60	0.08	0.37	0.08	0.37	0.12	0.37	0.10	0.37	0.13	0.38	0.12	0.37
HR60			0.35		0.44		0.36		0.40		0.43	
EDA120					0.25		0.22		0.83		0.85	
HR120							0.47		0.55		0.67	
EDA180									0.09		0.09	
HR180											0.57	

The regression analysis with respect to frustration revealed no significant additions to the model of frustration by physiological signals.

4.8.8. Summary of Physiological Data. The physiological data analyses reveal that frustration played a prominent role in the physiological signals produced. The first 60 seconds of the task were particularly revealing as there were interaction effects for both signals for interest and frustration. The modeling analyses helped reveal some of the differences in each signal, as frustration contributed more to the models of heart rate data, while interest contributed more to models of skin conductance. This reflects the complex relationship between interest, frustration, and physiological signals as demonstrated in the ANOVAs; for many of the measures computed for both physiological signals, both interest and frustration were significant as well as the interaction between the two.

4.9. Analysis of Dimension Extremes

An analysis of the dimension extremes (i.e. interesting only and frustrating only) was performed to more clearly delineate the difference between tasks in engagement and frustration. This diagonal analysis compared interesting tasks without a frustrator (hereafter referred to as the interesting task) and uninteresting tasks with a frustrator (hereafter referred to as the frustrating task) to determine more distinct differences in both states.

The diagonal analysis of physiological data showed that while there no significant differences in overall skin conductance or heart rate data between the interesting and frustrating tasks, interesting tasks had significantly higher (*t*(59)=7.41, *p*<0.001) skin conductance in the first 60 seconds. Interested participants also had significantly higher heart rate (*t*(59)=4.09, *p*<0.001) in the first 60 seconds between interesting and frustrating tasks. There were no significant differences in number of skin conductance responses between frustrating and not frustrating tasks. Lastly, there were no significant differences in windows of skin conductance between interesting and frustrating tasks, but participants had significantly windows of higher heart rate for frustrating than interesting tasks during (*t*(41)=-5.82, p<0.001).

The diagonal analysis of questionnaire data showed frustrated participants had significantly greater overall stress than interested participants (*t*(39)=-9.44, *p*<0.001), had significantly greater distress than interesting participants (*t*(39)=-10.85, *p*<0.001), but there were no significant differences in worry. Participants who were interested had significantly higher engagement than participants who were frustrated (*t*(39)=19.09, *p*<0.001). In terms of the subscales of engagement, interested participants reported significantly higher *perceived usability* (*t*(39)=9.38, *p*<0.001) significantly higher *endurability* (*t*(39)=10.88, *p*<0.001), *felt involvement* (*t*(39)=5.57, *p*<0.001), and *novelty* (*t*(39)=5.65, *p*<0.001). There were no significant differences in *focused attention* between interested and frustrated participants. Frustrated participants also experienced significantly greater frustration (*t*(39)=-12.22, *p*<0.001) than interested participants.

With regards to search behavior, frustrated participants clicked on the SERP significantly more (*t*(39)=-2.22, *p*<0.05) than interested participants, but there were no significant differences in clicks on documents. Frustrated participants, however, loaded significantly more SERPs (*t*(39)=-3.09, *p*<0.01) than interested participants, and scrolled significantly more (*t*(39)=-4.00, *p*<0.001). Frustrated participants submitted significantly more queries (*t*(39)=-4.39, *p*<0.001), but interested participants bookmarked significantly more documents (*t*(39)=2.11, *p*<0.05).

4.10. Results Summary

The table below summarizes many of the relevant results in this study across all measures: physiological, self-reported, and behavioral.

Table 61

Summary Table of Study Results

	Interest		Frustration		Combinations			
Item	Yes	No	Present	Absent	IyFy	IyFn	InFy	InFn
Skin Conductance (First 60 seconds)	7.38 (0.40)	7.39 (0.31)	7.20 (0.37)	7.58 (0.30)	6.96 (0.32)	7.80 (0.29)	7.43 (0.23)	7.36 (0.36)
Heart Rate (First 60 seconds)	80.43 (2.26)	80.17 (1.90)	79.78 (2.26)	80.82 (1.77)	80.14 (2.66)	80.72 (1.76)	79.42 (1.72)	80.92 (1.78)
Number of SCRs	4.50 (4.18)	3.45 (2.99)	4.79 (4.21)	2.91 (2.71)	5.54 (4.72)	2.97 (3.13)	4.05 (3.53)	2.84 (2.20)
Engagement	3.45 (0.73)	3.15 (0.86)	2.96 (0.82)	3.64 (0.64)	3.12 (0.74)	3.77 (0.58)	2.80 (0.90)	3.51 (0.67)
Frustration	7.24 (3.90)	7.64 (3.58)	9.83 (3.76)	5.04 (2.78)	9.67 (3.54)	4.80 (2.32)	10.00 (4.00)	5.27 (3.16)
Queries	5.65 (4.45)	5.06 (4.13)	7.51 (4.66)	3.20 (2.42)	7.87 (4.73)	3.42 (2.73)	7.15 (4.62)	2.97 (2.08)
Clicks Per Query	7.77 (4.77)	6.79 (3.97)	8.27 (5.03)	6.29 (3.42)	8.32 (5.32)	7.22 (4.16)	8.22 (4.79)	5.35 (2.15)
SERPs displayed	13.80 (9.16)	12.29 (7.88)	15.52 (9.37)	8.56 (5.90)	15.90 (10.09)	9.70 (6.95)	15.15 (8.70)	7.42 (4.42)
SERP clicks	19.52 (14.30)	17.70 (13.75)	23.04 (15.57)	14.19 (10.66)	23.65 (15.31)	15.42 (12.05)	22.45 (15.96)	12.95 (9.03)

CHAPTER V: DISCUSSION

This chapter will discuss the findings of the study. Each research question and hypothesis (where applicable) is presented accompanied by evidence for its support or rejection.

The task rankings showed that participants tended to favor the online communication task over other tasks, and tended to dislike the vehicle purchasing and tattoo removal tasks more than the other tasks. This indicates that there is a distinct lack of homogeneity among task rankings, which shows that though an attempt was made to balance the domains and subject matter of the tasks, participants clearly preferred certain tasks to others. This could have been because of the topicality of online communication; the title of this topic is vague enough that it could have suggested a task about social media, which is popular among the demographic examined in the experiment. The pre-task rankings showed that participants felt significantly more interested in tasks that had some relevance to them, that they had more pre-task knowledge of, and which they felt they had more skill at. The significant differences in these ratings indicate that the manipulation of interest was largely successful, and that higher ratings of interest are reflected in higher ratings of engagement, as found in other work (O’Brien & Lebow, 2013).

The first research question in this study was: to what extent can physiological signals be used to disambiguate engagement and frustration? I hypothesized that physiological signals during frustrating episodes would exceed those experienced during engaging episodes. This hypothesis was true for overall skin conductance (though the differences were not significant) and for number of skin conductance responses (which were significant). The heart rate values for interesting and frustrating tasks were virtually similar overall. However, the reverse was true (i.e., there were greater physiological signals during interest than frustration) in the initial task stages; there was greater skin conductance in the first 60 seconds of the task, and greater heart rate in the first 60 seconds of the task for interesting tasks than frustrating tasks. These results show that disambiguation of emotional state is possible using both physiological signals and examining the data at different task stages; frustrated participants had greater amounts of skin conductance responses, but participants who were engaged had greater skin conductance in the first 60 seconds of the task than participants who were frustrated, and also had greater positive change in the first 60 seconds than participants who were frustrated. This time period (the first 60 seconds) is especially significant because work has shown that this is the time period during which the stimulus is experienced most strongly (Boucsein, 2012), and thus the resulting arousal is expressed most strongly. This means that both interested and frustrated participants experienced high levels of arousal in the beginning of the task, but these were expressed differently via different physiological responses.

The findings regarding changes in valence of skin conductance and heart rate are interesting because of what they may indicate. We see that there are stark differences in valence of change with regard to skin conductance between interesting tasks without a frustrator, and interesting tasks without a frustrator. Interesting tasks without a frustrator had a greater increase in skin conductance, pointing to increases in arousal that others have linked to increases in cognition (Yun et al., 2014). However, we see that there were no significant differences in interesting and uninteresting tasks, because participants experienced increases during uninteresting tasks with or without a frustrator as well. Participants may have experienced increases in skin conductance during uninteresting tasks as they use cognitive resources to try to navigate an unfamiliar information space. Interestingly, we also see that participants who were frustrated had greater increases in skin conductance than not frustrated participants, which could indicate that participants were again using cognitive resources to cope with their failed attempts at finding relevant information.

The increase in overall skin conductance during periods of frustration above those of engagement could possibly be explained by subconscious reactivity. Subconscious reactivity refers to physiological arousal not related to any particular cognition. This means that participants may have been reacting to the frustrating situation and thus experienced more skin conductance responses. However, the window analysis of skin conductance signals for both frustrating and interesting tasks points to a different explanation: cognitive activation. Cognitive activation refers to increases in cognition. Studies have shown that cognitive activation is linked to greater levels of arousal (Yun, Shastri, Pavlidis, & Deng, 2009). It is likely that greater cognitive activation occurred during interesting tasks than non-interesting tasks, which is expressed as greater skin conductance. The graphs of the skin conductance data show that skin conductance levels were similar for frustrating tasks, indicating that frustration could have also had an effect on cognitive activation and arousal, but an opposite one. Participants may have spent less time thinking about the task and instead mechanically reformulated their queries to achieve more relevant search results. The strength of these two responses is confirmed by the significant interaction effect for the window analysis of skin conductance data, indicating varying levels of both signals at different points during the task.

Examining the skin conductance data by both task dimensions complicates the cognitive activation hypothesis. Skin conductance was highest for interesting and frustrating tasks, and the lowest for not interesting and not frustrating tasks. The window analysis shows that participants had the highest skin conductance for interesting and not frustrating tasks, and uninteresting and not frustrating tasks had the lowest skin conductance. Interesting and frustrating tasks also had the highest rate of positive increase in skin conductance in the first 60 seconds, while interesting and frustrating tasks had the lowest rate of positive change during the first 60 seconds of the task. As stated before, other studies have associated increases in skin conductance with increases in cognition (Yuan et al., 2014), but with the exception of the data from the initial task stages, there were only significant main effects found for frustration. It is possible that frustration might have dampened the skin conductance responses generated by interest, i.e., participants who were not frustrated may have been more able to engage with the task and experience higher levels of each physiological signal. This seems to be indicative of interest "trumping" frustration; because nothing acted to stop the interest, arousal was highest during interesting and not frustrating tasks in the first 60 seconds.

The results of the multilevel modeling analysis also challenge the idea that greater cognition and interest is expressed as a greater increase in arousal, and by extension, skin conductance. Though interest had a higher coefficient in the best model for skin conductance, it was negative, indicating an inverse relationship between skin conductance and interest. This follows if one examines the overall means of the skin conductance data, upon which the model was based: interesting and not frustrating tasks have the lowest skin conductance values. These tasks represent the most opportune time for participants to experience unhampered by frustration. This supports the conclusions drawn in O'Brien and Lebow's (2013) work, which found that as engagement increases, physiological signals of stress decrease. Therefore, a case can be made for proper contextualization of physiological signals: we know that participants experienced high levels of interest and low levels of frustration for interesting and not frustrating tasks, and while participants experienced lower skin conductance overall during these tasks, they experienced the greatest positive increases in skin conductance during the first 60 seconds of the task. This indicates that though both sets of results indicate arousal, the examination of the signals through different lenses is key to interpretation.

The heart rate data in tasks of different dimensions complicate our understanding of these signals further. Though the heart rate data overall was similar between tasks, the window analysis revealed that participants had the highest heart rate during uninteresting and not frustrating tasks, and the lowest during interesting and not frustrating tasks. Significant main effects for interest and frustration as well as significant interaction effects were found for these data. The data for the first 60 seconds of the task also shows that participants had the greatest positive change for uninteresting and frustrating tasks, and the greatest negative change for interesting and frustrating tasks, and a significant main effect for frustration as well as a significant interaction effect was found. These data suggest that subconscious reactivity may not thoroughly explain the variance in heart rate, especially during the first 60 seconds. Given that the heart rate signal had more variability and more fluctuations than skin conductance data, and also tended to decrease more during both frustrating and interesting tasks, it is possible that heart rate could be indicative of a different kind of arousal than skin conductance.

Other studies have linked changes in heart rate to changes in mood (Moss, 2004) and greater regulation of emotional response (Appelhans & Luecken, 2006), and it is possible that during frustrating and interesting tasks there were greater mood shifts. The greatest declines in heart rate occurred at the two tasks that were polar opposites of each other: interesting tasks with a frustrator, and uninteresting tasks without a frustrator. These are the two tasks during which participants likely experienced the greatest and least (respectively) amounts of stimulation. Participants experienced the greatest decline in heart rate during both interesting and frustrating tasks, which could indicate a drop in mood due to frustration. Participants experienced the second greatest decline in heart rate during uninteresting and not frustrating tasks, which could indicate that participants had drops in mood during tasks that they were uninterested in and did not struggle with. These data reflect the principles of the inverted-U theory, which states that too much stimulus and too little stimulus can both lead to poor performance, while just enough stimulus would create optimal performance.

However, these data again challenge other work that suggests that frustration involves an increase in physiological signals (Scheirer et al, 2002; Partala & Surakka, 2004). The decreases in heart rate may be explained by Russell's (2003) understanding of core affect. Russell stated that core affect is composed of both "hedonic values" referring to pleasure and displeasure, and "arousal," which ranges from sleepy to activated. As skin conductance has previously been tied to changes in cognition (Nourbaksh et al., 2012; Yun et al., 2009), it is possible that heart rate instead encompasses the hedonic portion of affect, and thus differs according to changes in emotion, rather than changes in a state dominated by cognitive activation, such as engagement. Other studies (Appelhans & Luecken, 2006) have found that higher heart rate variability (such as that found in this experiment) is associated with greater emotional regulation. Thus, heart rate may be most useful as a disambiguation tool for emotional state. Participants had greater decreases in heart rate for interesting tasks as well as frustrating tasks, and had the greatest overall decrease for interesting and frustrating tasks. These decreases could be indicative of attempts at emotional regulation that occur in the presence of a frustrator. Evidence for this can be found in the results from the multilevel modeling. The best-fitting model for heart rate included a large negative coefficient for frustration, suggesting that an inverse relationship exists between heart rate and frustration, as seen in the largest decrease for heart rate during the most frustrating task (IF task).

The regression analysis including physiological signals showed that skin conductance at 120 seconds, and heart rate at 60 seconds, were significant contributors (though with small coefficients) to the model of engagement. However, there were no significant physiological predictors of frustration. Interestingly, if we compare this with the original multivariate regression, we see that there were several behavioral signals that significantly predicted frustration, more so than engagement. It may be that frustration is more easily predicted by search actions than physiological signals. This seems in a way to confirm the multilevel modeling analysis with regards to engagement, and undermine the modeling analysis of frustration. In the multilevel modeling analysis, skin conductance was crucial to engagement, and this is reflected in the regression. However, while heart rate was important to the models of frustration, neither heart rate nor skin conductance were significant contributors to frustration. I believe that this shows that frustration can be strongly predicted using a number of signals, but it seems that behavioral signals may be stronger predictors of frustration when used in conjunction with physiological signals.

The diagonal analysis of the physiological data showed that interested participants had significantly higher skin conductance and heart rate in the first 60 seconds than frustrated participants, but there were no significant differences in skin conductance responses or in windows of skin conductance. This seems to confirm research that shows that participants are more likely to experience a strong reaction to a stimulus in the first 60 seconds of the task. One caveat in the analysis and interpretation of the diagonal data (especially with regard to the physiological data) is the possibility of spillover effects, given that this study was not designed to examine the effects of interest and frustration as standalone effects.

This study showed that physiological measures, viewed from a high level, offer very little information without sophisticated modeling techniques; the differences in physiological states were not as discernible when viewing levels of each signal over the course of the entire task. Methodologically, this means that physiological data needs to be parsed finely and modeled carefully in order to be most useful. This study also showed that physiological signals could be used to disambiguate engagement and frustration; during engagement, participants experienced lower skin conductance, and during frustration, participants experienced heart rate. Given the stark differences in means and model coefficients, it is likely that these signals do indeed represent different emotional states.

The second research question posed in this study was: to what extent do search actions differ for participants who are engaged versus frustrated? There were no hypotheses for this research question. Participants generally performed more behaviors when they were engaged or frustrated though there were very few significant main effects for interest; engaged and frustrated participants submitted more queries (and longer queries), clicked more, scrolled more on the SERP and spent more time on tasks. Significant main effects were found for frustration for most signals, including queries, clicks, documents bookmarked and total task time as well as time spent on SERP. A significant main effect for interest was found for time between queries, indicating that this could be a potentially useful signal in distinguishing interesting tasks from frustrating tasks. The regression analyses of search behavior were also helpful in this regard. Multiple and logistic regressions of search actions on engagement and frustration showed that signals like scrolls, clicks on documents, and clicks on the SERP could be useful in identifying a frustrating search episode, thereby distinguishing frustration from engagement.

The diagonal analysis allowed for comparison between interested and frustrated states, and showed that frustrated participants had clear markers such as more clicks on the SERPs, more SERPs loaded, more scrolls, and more queries. Interested participants, on the other hand, bookmarked significantly more documents. This confirms the regression analyses, which show that there are clearer markers for frustration than for engagement.

The increases in search behavior observed during engagement are interesting because though there were no significant main effects for interest in search behavior (with the exception of query time intervals and scrolls on the SERP), in some respects they contradict the findings in O'Brien and Lebow (2013) and support the findings of many other studies regarding engagement and search behavior (Jiang, He & Allan, 2014). However, these studies did not successfully link subjective engagement and search actions, as this study has, and there is a theoretical basis for these findings. The increases in search action during engagement (as well as frustration) can be understood as a signal of behavioral activation, part of the appetitive motivation process (Amsel, 1952) outlined earlier in this dissertation. As described earlier, frustration and motivation are related through appetitive motivation, which is a state in which a person experiences increases in behavior - the examples frequently used in the literature are biological drives such as hunger and thirst, which induce behavioral activation to resolve those states. One explanation for the significant main effects for frustration found for many signals is that research has shown that negative emotions (such as frustration) have a stronger effect than positive emotions (Russell, 2003); participants who were frustrated may have felt their frustration more strongly than they felt interest when performing interesting and frustrating tasks. Still, for many facets of search behavior, though there were no significant main effects for interest, there were greater means for interesting tasks than uninteresting tasks, implying that interest could also have had a small effect on an increase of search actions. Tasks that were both interesting and frustrating consistently had much higher search actions than any other task type combinations, which points to strong levels of behavioral activation created by frustration and to a lesser extent interest.

In the behavioral activation hypothesis, frustrated participants may have sought to resolve their frustrated state in the form of performing more behaviors. In a practical sense, participants completed more behaviors because their search tactics were proving fruitless. Since the frustrator took the form of poor search results, participants had to perform more actions to find relevant results as their attempts to find relevant ones usually ended in failure. Participants may have been motivated to solve the task (and persist in their search actions) because of their interest in the task. Contrary to appetitive motivation engendering frustration, aversive motivation can help explain the consistently low number of search actions for uninteresting and not frustrating tasks. Aversive motivation is characterized by behavioral inhibition. When participants were not interested or frustrated, there was no behavioral activation, and they responded by performing fewer behaviors in general.

Querying is another area where behavioral activation caused by both interest and frustration was present. Participants entered more queries for interesting tasks as well as for frustrating tasks, but were likely motivated in two different ways. Interest in the task (and prior knowledge) likely led to participants querying more, and frustration at finding poor search results likely also led to participants submitting more queries and performing more query reformulations. Closer examination of query time intervals demonstrates some of the differences between engagement and frustration. Participants had greater time intervals between queries for interesting tasks than uninteresting tasks, and a significant main effect was found for this, indicating that participants confirmed the finding that, though not significant, more time was spent examining documents during interesting tasks. However, analysis of term uniqueness for queries did not show any differences in specific terms between tasks, which other studies have found (Hassan et al., 2014). Given that the trends in this paper conflict with other work, it is unclear precisely what emotional state is signaled by query reformulation patterns, but it is still an area that holds promise for disambiguation of engagement and frustration.

If we look at the potential of search actions for disambiguating engagement and frustration, utility can be found in querying and click patterns. Participants loaded fewer SERPs for interesting tasks than frustrating tasks, and frustrated participants also had higher clicks on the SERP as well as more clicks per query. This shows that frustrated participants likely looked through more search results pages and clicked more, indicating less focused searching and less concrete search strategy. This partially supports the findings in Hassan et al. (2014), who found that participants had greater clicks per query for exploring rather than struggling sessions, and Feild et al. (2010), who found that frustrated participants exhibited different click behaviors. Though both interesting and frustrating tasks had greater clicks per query than their non-frustrating and uninteresting counterparts, the greatest amount of clicks per query were for interesting and frustrating tasks, during which participants were likely exploring as well as struggling. Interested participants had more clicks on documents than frustrated participants, which could serve as an indicator of greater document exploration than SERP exploration.

Though search actions generally increased during both engaging and frustrating episodes, examining the results for the interesting and not frustrating tasks compared to the uninteresting and frustrating tasks offers the clearest path for disambiguating engagement and frustration, as these tasks varied by each dimension. Comparison of these two types suggests that frustrated participants will submit more queries, will click more per query, load more SERPs, and generally perform more search actions. It also suggests that engaged participants will bookmark more documents, spent more time on documents, and spend more time between queries. In general, the trends suggests that frustrated people will have shorter task times and spend more of that time interacting with the SERP, while people who are engaged with the task will have longer task times and interact with documents more. The regression analyses confirm this hypothesis, and findings in other work regarding frustration and search actions (Feild et al., 2010). The multiple regression analysis showed that clicks on documents was a significant (though small) predictor of engagement. The multiple regression analysis for frustration also showed small but significant contributions of queries, scrolls, SERP clicks, and clicks on documents to frustration, though the coefficient for clicks on documents for frustration was negative, indicating an inverse relationship to frustration. The logistic regression, though it yielded slightly different results, also showed that queries, scrolls, and clicks on the SERP were the best indicators of frustration.

The third research question posed in this study was: how does task interest impact search behavior, engagement, and frustration? How does this, in turn, impact stress? I hypothesized that engagement would be greater during interesting tasks, and lower during frustrating tasks. I also hypothesized that task interest would lower stress. This hypothesis was supported in that engagement was greater during interesting tasks and lower during frustrating tasks. The stress hypothesis was also confirmed, as interesting tasks had lower levels of overall self-reported stress than uninteresting tasks.

The diagonal analysis of the questionnaire data showed that interested participants had significantly higher engagement, with the exception of focused attention. This is an interesting finding because it seems likely that both interested and frustrated participants are focused on the task at hand, but for different reasons. It's also interesting that though frustrated participants had greater overall stress, and distress, there were no significant differences in worry. It seems that there are opportunities for self-reflection and self-consciousness both in periods of interest and frustration. During an interesting episode, a user might be thinking about how the search results fit into their information space or perception of their problem space, as this study showed that participants are more interested in tasks that have some personal relevance for them. Frustrated participants may also naturally worry about their performance if they are struggling.

Though there were no significant main effects for interest, participants performed more search actions for interesting tasks. The lack of a significant main effect for interest could indicate that frustration dampened the effect of interest on most aspects of search behavior. The lack of significant main effects for interest and search behavior could also be due to the artificial nature of the interest manipulation; because participants were forced to rank which tasks were interesting, instead of organically generating interesting tasks, the differences in search behavior may be smaller than if interest were not relative.

Engagement scores were higher for tasks that were interesting than for tasks that were uninteresting. Participants had greater levels of focused attention for interesting tasks, which supports the hypothesis that greater engagement with a task results in greater focus and motivation to complete it (Csikszentmihalyi, 1991). Participants also felt more involved in interesting tasks, indicating that the task may have had had some greater meaning to participants. This fits well with participants’ ratings of the relevance and pre-task knowledge of interesting tasks. It is also curious that participants felt that interesting tasks had greater novelty; it is likely that since they already had some knowledge of the interesting tasks, they were able to discover and determine new information about the task. It is also interesting that there were significant main effects found for both interest and frustration with respect to engagement. While frustration has appeared to dampen the effect of interest in many areas (such as physiological signals and search actions), it appears that interest played a greater role in promoting and sustaining engagement, while frustration significantly lowered it.

Interest had no effect on frustration questionnaire responses. Interesting tasks generally had lower frustration scores than uninteresting tasks. A possible explanation for this could be that uninteresting tasks were more frustration-prone precisely because they are less interesting. However, this seems a bit contradictory, as one might imagine that interesting tasks (or tasks that participants feel more invested in) would present an opportunity for greater feelings of frustration. However, this presents the opportunity for an important methodological note: the results indicate that task interest (overall) was not as susceptible to effects from the frustrator (or perhaps any other experimental manipulation) as one might have thought - interest in the task stayed fairly robust during frustrating episodes. When examining the tasks by interest and frustrator combinations, we see that although uninteresting and frustrating tasks had the greatest levels of frustration, interesting and frustrating tasks also had similarly high levels of frustration, and similar scores across the three frustration questions. This supports the belief that frustration had a demonstrable effect on interesting tasks because participants were more engrossed and felt more motivated during interesting tasks, and thus were less tolerant of interruptions in the information-seeking process. It is also possible that people were more frustrated with uninteresting tasks because they felt more pressure to find an answer for a task they had very little pre-task knowledge about.

There were very few differences in stress between interesting and uninteresting tasks, but very clear differences between frustrating and not frustrating tasks. The differences in stress responses by frustrator can be explained most effectively using Lazarus' (1984) model of stress appraisal. In this model, one uses goal congruence, goal relevance, and ego involvement as three forms of cognitive appraisal to determine if a stimulus is a stressor. One also tries to appraise whether the stimulus is in line with their goals, if the stimulus is relevant to their goals, and whether they experience ego-related feelings such as changes in self-esteem to determine whether they should react with a stress response. In this experiment, participants likely appraised the frustrator (or the frustrating task) as preventing them from reaching their goals during interesting tasks (which had more relevance to participants) as well as literally preventing them from finding documents relevant to their goal (i.e., satisfying the demands of the task). Though verbal utterances were not formally collected during the experiment, several participants stated that they felt "they were so bad at search" or "totally doing this wrong" when they encountered non-relevant results.

The fourth research question in this study was: how does the presence of a frustrator moderate the relationship between task interest and search behavior, engagement, and frustration? I hypothesized that participants would feel more frustrated during interesting tasks than non-interesting tasks, and that this frustration would be expressed in the form of higher physiological signals above and beyond that of engagement as well as higher reports of frustration. This hypothesis was not confirmed, in that participants experienced greater (though non-significant) frustration during interesting tasks than uninteresting tasks. The hypothesis was confirmed for the physiological data in that participants experienced a greater number of skin conductance responses for frustrating tasks than not frustrating tasks, and there were significant main effects for frustration for both skin conductance and heart rate for frustrating tasks in the first 60 seconds of the task. Though there were no significant interaction terms, the presence of a frustrator appeared to moderate the relationship between task interest and search behavior, engagement, and frustration such that it created increases in several areas: interesting and frustrating tasks consistently had higher search actions as well as higher frustration, engagement, and stress scores.

Frustration alone had a significant effect on almost all measurements in this study. Frustrated participants felt that tasks were significantly more difficult, rated their skill at these tasks significantly poorer, felt the system had a significantly poorer ability to retrieve documents, and felt that they were significantly less successful for these tasks than not frustrating tasks. These data are interesting because research has shown that frustration can impact self-perception and mood (Klein, Moon & Picard, 2002), and in this study frustration appeared to lower both the participants' feelings of self-efficacy as well as their ratings of the system. These lowered feelings of self-efficacy and success are confirmed when looking at the post-task data by combinations of frustrator and interest. Participants felt their self-efficacy was lowest for uninteresting and frustrating tasks, likely because they were not as motivated to complete the task, and had less prior knowledge, which could have rendered them less capable of completing the task. Participants felt the least success for tasks that were interesting and frustrating, also likely because these tasks had personal relevance and they experienced difficulty completing them.

Again, though there were no significant interaction effects, examining the differences between interesting and frustrating tasks and the other task types offers the clearest way to discuss the moderating effect of the frustrator. With respect to engagement, interesting and frustrating tasks had high stress and frustration scores, but relatively low engagement. This is important because while the participant may have attempted to become engaged with this task due to their interest, frustration may have prevented the participant from becoming absorbed in the task, and prevented the information gleaned from the task from becoming imprinted on the participant. This is reflected in the *endurability* scores, which were significantly lower for frustrating tasks than not frustrating tasks.

The results were similar for the SSSQ data. Frustration increased levels of distress, worry, and overall stress for interesting and frustrating tasks. Frustration likely created more distress in participants likely because of the difficulty involved in completing frustrating tasks, and also created more worry, possibly due to feelings of lowered self-efficacy (i.e., the skill item on the post-task questionnaire). Participants had the highest levels of stress for uninteresting tasks with a frustrator. This is likely because participants felt less interested in the task (which could have acted to combat frustration) and also experienced a struggle with the task. Lastly, self-reported frustration was higher for frustrating tasks than not frustrating tasks, which confirmed the effect of the frustrator, but was again highest for uninteresting tasks with a frustrator, followed by interesting and frustrating tasks. The questionnaire data seem to indicate that that while the frustrator seemed to increase many negative items, interest dampened these increases somewhat.

Frustration moderated task interest and search behavior by creating an increase in behaviors. Interesting and frustrating tasks consistently had the highest behaviors out of all tasks (for example, the highest number of SERPs displayed, the most queries, the most SERP clicks). Frustrative non-reward theory (Amsel, 1958) is useful to explain this relationship. In this theory, Amsel outlines persistence and invigoration as learned behaviors that occur in the presence of a frustrating situation. Participants experienced the highest increases in search behavior during interesting and frustrating tasks. This is likely because they were exhibiting persistence by continuing to complete more search actions, though these efforts were not as fruitful as they likely had hoped. The invigoration response, characterized by renewed efforts despite the presence of a frustrator, was also present in both types of frustrating tasks, given that they generally had higher amounts of search actions than not frustrating tasks. These efforts can also be thought of as coping strategies (Mansourian, 2008). The increase in behaviors seems to indicate that frustration and interest entered a somewhat “combative” relationship, i.e., the participant’s interest in the task spurred them to continue searching, while the frustrator, in effect, “forced” them to perform more behaviors in order to satisfy the demands of the task.

Participants performed more scrolls on the SERP and loaded more SERPs for frustrating tasks than not frustrating tasks, but this is likely due at least in part to the nature of the experimental manipulation with regards to the frustrator. Since the frustrator in this experiment was created by presenting the 500th ranked results first, participants likely loaded more SERPs because they did not find relevant results on the first or the second pages. Frustration also resulted in greater scrolling, which could again be because of the lack of relevant results presented on the SERP. Click behavior in particular is interesting; participants clicked more on the SERP during frustrating tasks, and clicked more on documents during non-frustrating tasks. This is likely because participants completing frustrating tasks spent more time trying to find relevant documents, while not frustrated participants spent time reading or examining documents unfettered. Again, this effect is most pronounced for interesting and frustrating tasks - despite frustration, participants in this condition had the highest average clicks on the SERP as well as had the highest average clicks on documents. Bookmarks were lower for frustrating tasks, supporting the finding that people found less relevant results for frustrating tasks. As shown in the regression analysis, queries, SERP clicks, clicks on documents, and scrolls were the most predictive of frustration.

Overall, this experiment showed that disambiguating engagement and frustration is a difficult task, given that participants who are both engaged and frustrated tended to exhibit a greater increases in outward indicators of behavior such as search actions. However, it appears that skin conductance and heart rate may indeed be useful as a tool to disambiguate the two, given the finding that participants had significantly greater skin conductance responses for frustrating tasks, and participants had significantly higher overall skin conductance for interesting tasks. The inverted-U theory, which states that a “perfect” amount of stimulus can be useful in creating optimal behavior while greater or lesser amounts of stimuli can produce less than optimal behaviors, is applicable here. The inverted-U theory is useful for explaining the increases in search actions, as well as greater reports of frustration, and stress for interesting and frustrating tasks. These tasks in particular represent an overwhelming amount of stimulus, given that participants are both very interested in the task but also thwarted from completing it. Tasks that present lower levels of stimulus (such as uninteresting and not frustrating tasks) resulted in lower search actions, lower skin conductance, and lower frustration and stress. However, it seems that the optimal “middle ground”, or interesting tasks without a frustrator, resulted in higher levels of skin conductance, lower levels of frustration and stress, and high levels of engagement.

This work fits into the larger work on theories of emotion in information science by confirming many of the principles explored in other work. This work has served to confirm Kracker (2002)’s work using Kulthau’s ISP model, which states that there are affective and cognitive states that affect information-seeking behaviors. This work showed that negative states such as anxiety and uncertainty can alter information-seeking behavior; in this study, negative emotions meant an increase in search actions. This work also confirms Nahl (2004)’s work that information-seeking is laden with emotion. However, this work also serves to add more nuance to the theories of information behavior that incorporate emotion. This work showed that differentiating the type of emotion is important to interpreting the resulting search behavior, specifically differentiating between positive and negative emotions can help contextualize whether increases in behavior are the result of exploration or of struggling. This work also serves to confirm in a larger sense that collecting and understanding information related to emotion is essential to understanding user experience and user search behavior.

CHAPTER VI: CONCLUSION

This dissertation explored the utility of physiological signals in disambiguating engagement and frustration, as well as differences in search behavior between the emotional states of engagement and frustration. Specifically, this work addressed four research questions:

- RQ1: To what extent can physiological signals be used to disambiguate engagement and frustration?
- RQ2: To what extent do search actions differ for participants who are engaged versus frustrated?
- RQ3: How does task interest impact search behavior, engagement, and frustration? How does this, in turn, impact stress?
- RQ4: How does the presence of a frustrator moderate the relationship between task interest and search actions, engagement, and frustration?

An experimental study was conducted with 40 participants. This study indicated that skin conductance and heart rate had the potential for disambiguating engagement and frustration, but this disambiguation must occur at specific analysis points: within the initial stages of the task, within windows of time during the task, and by examining the number of skin conductance responses throughout the task. The data revealed that both engagement and frustration are marked by an increase in skin conductance, but frustrated participants had greater skin conductance responses than engaged participants. The window analysis of the heart rate data revealed significant main and interaction effects for both interest and frustration, indicating that though the means were similar, the fluctuations in in heart rate were important in determining whether the participant was engaged or frustrated. The multilevel modeling analysis was also helpful, confirming that interest was a critical factor in models of skin conductance, while frustration was critical to models of heart rate. This work demonstrated how difficult it is to parse engagement and frustration, as both states are characterized by high levels of physiological arousal.

This study also demonstrated that both engaged and frustrated participants completed greater amounts of search behavior, but experienced varying levels of engagement, frustration, and stress. Specifically, this study found that though participants experienced an increase in search actions during both engagement and frustration, there were no significant main effects for interest for most search actions, while there were significant main effects found for frustration.

It was surmised that frustration had a dampening effect on interesting tasks, and that in general frustration had a greater effect on search actions than engagement. There were significant differences in number of scrolls between interesting and not interesting tasks, as well as a significant main effect for interest for queries, indicating that for a select few types of search behavior, interest was important. It was hypothesized that these two behaviors are perhaps more linked to interest than the other signals, and that these signals are not as sensitive to the effect of the frustrator. It was also hypothesized that behavioral activation was present for both engagement and frustration; participants who were both engaged and frustrated performed more search actions than any of the other task types. However, engagement and frustration manifest themselves differently in some aspects of search behavior. Frustrated participants tended to click more on the SERP than on documents, submit more queries, and scroll more on the SERP.

This study also found differences in reported engagement, frustration, and stress by interest and frustration. Specifically, participants reported higher engagement for tasks that were interesting, and lower engagement for tasks that were frustrating. Interesting tasks had lower (though non-significant) frustration scores, and participants rated frustrating tasks as more frustrating, confirming the success of the experimental manipulation. Greater stress was also reported for frustrating tasks than interesting tasks.

This work fits well with other work on engagement and frustration in interactive information retrieval because it serves to support many of the findings of other works. It supports previous findings that interest and engagement are related (O'Brien & Lebow, 2013). As discussed in the literature review, much of the work on engagement has characterized it by an increase in search actions, and this study offered support for that by linking reports of the cognitive aspects of engagement to increases in search behavior. It also confirms other work such as Feild et al.'s (2010) on frustration by supporting the finding that frustrated participants engage in greater click behaviors. This work has confirmed that search actions are very useful as indicators of engagement, though their interpretation and application must be carefully handled. This work has also served to support other research (Wu et al., 2012; Kelly et al., 2015) that has found that search tasks can be used to create engagement during information seeking in experimental contexts. More specifically, this work confirmed that exploratory search tasks that allow participants to investigate and assemble diverse sources of information are the most effective kinds of tasks for creating engagement, as other work has shown (Jiang, He & Allan, 2014). This work is also well situated among other research that has measured both the subjective experience of engagement as well as objective measures such as search behavior (Arapakis et al., 2014; Barreda-Angeles et al., 2015, adding to a trend of combining objective and subjective measures in order to get a more complete picture of participants.

This work also poses several questions, in particular: what advantages does disambiguating engagement and frustration offer? This study showed that participants exhibit increases in many signals when they are engaged and frustrated, and it is quite possible that these situations also occur frequently during natural searches. Both states (engagement and frustration) could signal a cause for intervention, and it is possible that determining high levels of arousal or search actions would be the first step in deciding whether to intervene. It is also possible that participants need intervention to encourage them to become engaged, and so identifying extremely low arousal or behavioral periods is also important.

Future work could include trying to identify a particular threshold over which participants pass from engaged to frustrated. This could involve different analyses of physiological data as well as different ways of soliciting user input in order to track changes in subjective experience more closely. There may also 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 by having participants bring in genuine tasks.

APPENDIX A: SEARCH TASKS

Health Task 1: One of your siblings got a spur of the moment tattoo. However, after some careful consideration, they now regret it. You decide to investigate methods for tattoo removal so you can make some suggestions to your sibling about what he might do to get rid of the tattoo. What are the current available methods for tattoo removal, and how effective are they? Which method do you think is best? Why?

Health Task 2: For several years, your friend has complained of periods of extreme fatigue, headaches, and joint pain. After seeing several doctors, a specialist diagnosed her with lupus. What are some other symptoms of lupus? What are different ways to treat lupus, and how effective are they? Which treatment would you recommend to your friend? Why?

Science Task 1: You’re working on an assignment for an environmental science class about different kinds of energy sources and their efficacy. Your essay compares nuclear and solar energy. Which one is most cost-efficient to produce? How do different types of energies compare with regards to environmental impact? Which type of energy do you think is better? Why?

Science Task 2: 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?

Technology Task 1: Your grandparent makes a comment to you about how much time you spend communicating with people via text messages and social media, such as Facebook and Twitter. Your grandparent suggests that your ability to communicate face-to-face might be underdeveloped. In a few weeks, you have a face-to-face job interview with a prospective employer, so you decide to do some research. Which face-to-face communication skills are people concerned about losing because of increased use of text messages and social media? Do you think there is cause for concern? Why or why not?

Technology Task 2: You recently received some money from your grandparents, and have decided to use it to purchase a new car. You are interested in purchasing a sports utility vehicle (SUV) and are trying to decide between the Honda CR-V, the Toyota Rav4 and the Jeep Liberty. The criteria that are most important to you are price, safety, and fuel efficiency. You are also interested to hear what others have to say about these vehicles. Which SUV would you purchase and why?

Entertainment Task 1: For his 14th birthday, your cousin has asked you for a video game that is rated "M" for mature audiences because it contains intense violence. Your are unsure about whether to purchase this game because you recently overheard two people discussing the effects of violent video games on young teenagers. What are some of the reported effects of violent video games on teenagers? What are arguments for and against allowing young people to play these types of games? Given these arguments, do you feel comfortable buying this game for your cousin? Why or why not?

Entertainment Task 2: Your friend is very athletic and is looking for a new sport to try. She is interested in endurance sports, as she has competed in endurance sports events in the past. Specifically, she is interested in sports that will improve cardio-conditioning, strength and agility. What are different types of endurance sports? What are their pros and cons? What sport would you recommend to your friend and why?

APPENDIX B: PRE – SEARCH QUESTIONNAIRE

Q1: How much do you know about this topic?

Nothing I Know Details

❍ ❍ ❍ ❍ ❍

Q2: How relevant is this topic to your life?

Not at all Very Much

❍ ❍ ❍ ❍ ❍

Q3: How interested are you to learn more about this topic?

Not at all Very Much

❍ ❍ ❍ ❍ ❍

Q4: Have you ever searched for information related to this topic?

Never Very Often

❍ ❍ ❍ ❍ ❍

Q5: How difficult do you think it will be to search for information about this topic?

Very Easy Very Difficult

❍ ❍ ❍ ❍ ❍

APPENDIX C: POST – TASK QUESTIONNAIRE

Q1: How difficult was it to find relevant documents?

Very Easy Very Difficult

❍ ❍ ❍ ❍ ❍

Q2: How would you rate your skill and ability at finding relevant documents?

Not Good Very Good

❍ ❍ ❍ ❍ ❍

Q3: How would you rate the system's ability at retrieving relevant documents?

Not Good Very Good

❍ ❍ ❍ ❍ ❍

Q4: How successful was your search?

Unsuccessful Successful

❍ ❍ ❍ ❍ ❍

APPENDIX D: THE USER ENGAGEMENT SCALE*

Focused Attention

Q1: I lost myself in this search experience.

Strongly Disagree Strongly Agree

○ ○ ○ ○ ○

Q2: I was so involved in this search task that I lost track of time.

Strongly Disagree Strongly Agree

○ ○ ○ ○ ○

Q3: I blocked out things around me when I was completing this search task.

Strongly Disagree Strongly Agree

○ ○ ○ ○ ○

Q4: The time I spent searching just slipped away.

Strongly Disagree Strongly Agree

○ ○ ○ ○ ○

Q5: I was absorbed in this search task.

Strongly Disagree Strongly Agree

○ ○ ○ ○ ○

Endurability

Q6: I would recommend this search interface to my friends and family.

Strongly Disagree Strongly Agree

○ ○ ○ ○ ○

Q7: I was really drawn into this search task.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Q8: Completing this search task was worthwhile.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Q9: I consider this search task a success.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Q10: This search task was rewarding.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Q11: This search experience did not work out the way I planned.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Felt Involvement

Q12: I felt involved in this search task.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Q13: This search task was fun.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Perceived Usability

Q14: I felt frustrated while using this search interface.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Q15: I felt annoyed while using this search interface.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Q16: I felt discouraged while using this search interface.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Q17: Using this search interface was mentally taxing.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Q18: This search experience was demanding.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Q19: I felt in control of the search experience.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Q20: I could not do some of the things I needed to do using this search interface.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Novelty

Q26: I felt interested in this search task.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

Q27: The content of this search interface incited my curiosity.

Strongly Disagree Strongly Agree

❍ ❍ ❍ ❍ ❍

*questions from the aesthetics subscale were not included.

APPENDIX E: FRUSTRATION QUESTIONNAIRE

Q1: Trying to complete this task was a frustrating experience.

Not at all   Extremely

❍ ❍ ❍ ❍ ❍

Q2: Being frustrated comes with this kind of task.

Not at all   Extremely

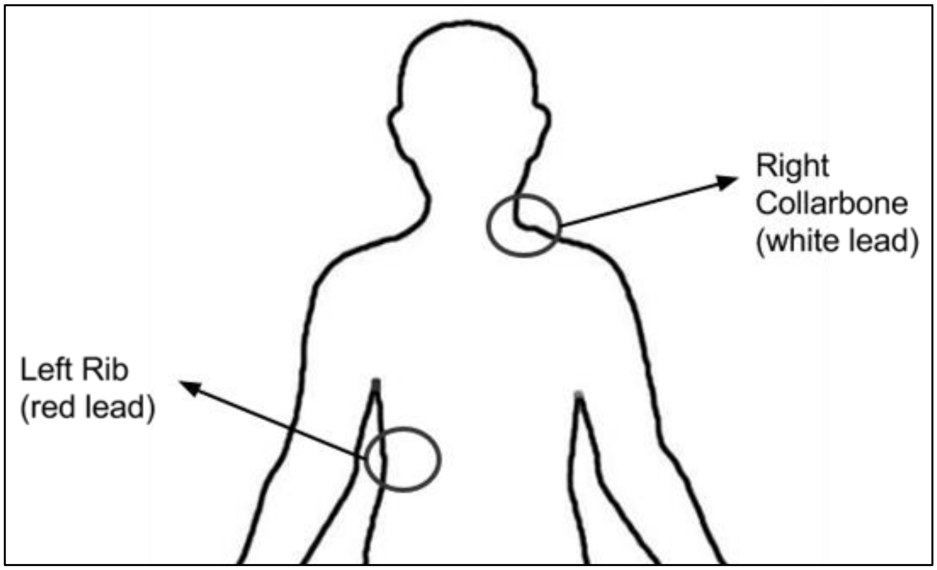
❍ ❍ ❍ ❍ ❍

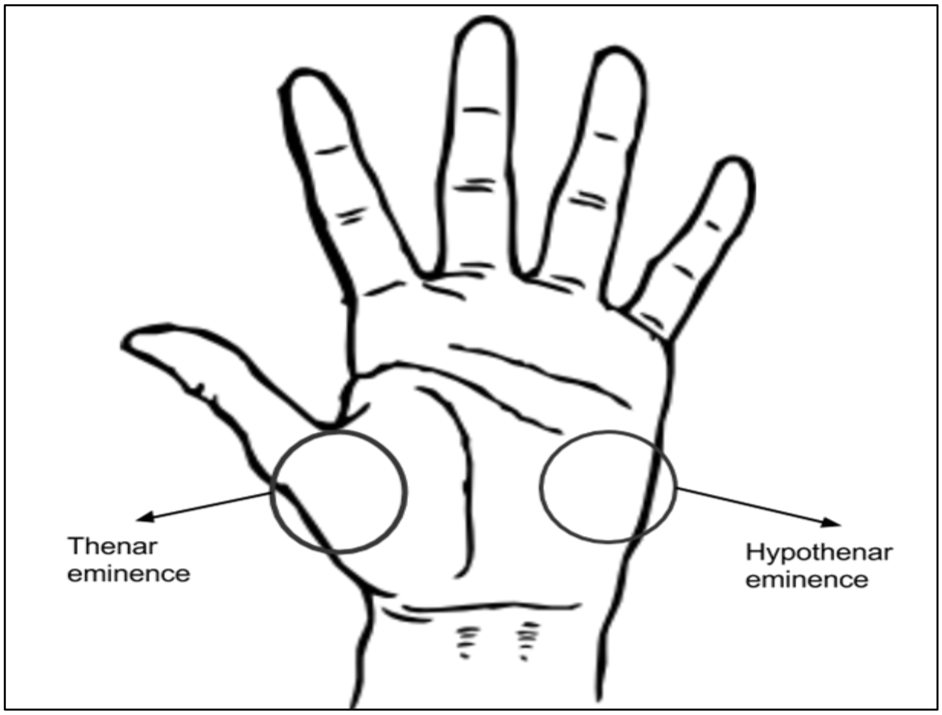
Q3: Overall, I experienced frustration during this task.

Not at all   Extremely

❍ ❍ ❍ ❍ ❍

APPENDIX F: DIAGRAMS OF ELECTRODE PLACEMENT





APPENDIX G: SHORT STRESS STATE QUESTIONNAIRE*

Q1: I thought about how others have done on this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q2: I was trying to figure myself out during this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q3: I felt angry while I was completing this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q4: I felt irritated while I was completing this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q5: I felt grouchy while I was completing this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q6: I was reflecting about myself during this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q7: I felt concerned about the impression I was making during this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q8: I felt annoyed while I was completing this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q9: I felt impatient while I was completing this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q10: I felt self-conscious during this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q11: I daydreamt about myself during this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q12: I thought about how I would feel if I were told how I performed on this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q13: I felt sad while I was completing these tasks.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q14: I felt depressed during this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q15: I was worried about what other people would think of me.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

Q16: I felt dissatisfied while I was completing this task.

Not at all Extremely

❍ ❍ ❍ ❍ ❍

*questions that loaded on the engagement factor were excluded.

APPENDIX H: DEMOGRAPHIC QUESTIONNAIRE

Q1: What is your age?

Q2: What is your sex?

- ○ Female
- ○ Male

Q3: What is your status at UNC?

- ○ Freshman
- ○ Sophomore
- ○ Junior
- ○ Senior

Q4: What is your major?

Q5: How long have you been conducting online searches for information?

Less than 1 year	1-3 years	4-6 years	7-9 years	10+ years
○	○	○	○	○

Q6: How often do you conduct online searches for information?

Less than once per week	1-3 times per week	4-6 times per week	1-3 times per day	4-6 times per day
○	○	○	○	○

7+ times per day
○

APPENDIX I: RECRUITMENT EMAIL

Subject: Take Part in a Study on Search Behavior!

----- Message Text ----

Do you regularly perform Internet searches? Then you can be part of a research study on search behavior. You will receive a sum total of $15.00 for participating. This study takes approximately 60 minutes to complete.

This study takes place on- campus, in the Interactive Information Systems Laboratory in Manning Hall. To participate, please email aedwards@unc.edu.

You will not be offered or receive any special consideration if you take part in this research; it is purely voluntary.

IRB Study Number: 15-0956

Thank you,

Ashlee Edwards, Ph.D. Candidate
School of Information and Library Science
University of North Carolina at Chapel Hill

REFERENCES

Amershi, S., & Morris, M. R. (2008). CoSearch: a system for co-located collaborative web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1647-1656.

Amsel, A. (1958). The role of frustrative nonreward in noncontinuous reward situations. *Psychological Bulletin*, *55*, 102-119.

Amsel, A., & Roussel, J. (1952). Motivational properties of frustration: I. Effect on a running response of the addition of frustration to the motivational complex. *Journal of Experimental Psychology, 43*, 363-366.

Anders, T. F., Sachar, E. J., Kream, J., Roffwarg, H. P., & Hellman, L. (1970). Behavioral state and plasma cortisol response in the human newborn. *Pediatrics*, *46*, 532-537.

Anderson, L. W., & Krathwohl, D.R., (Eds.)(2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. White Plains, NY*:* Longman.

Anttonen, J., & Surakka, V. (2005). Emotions and heart rate while sitting on a chair. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 491-499.

Appelhans, B. M., & Luecken, L. J. (2006). Heart rate variability as an index of regulated emotional responding. *Review of General Psychology, 10*, 229-240.

Arapakis, I., Athanasakos, K., & Jose, J. M. (2010). A comparison of general vs. personalised affective models for the prediction of topical relevance. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 371-378.

Arapakis, I., Jose, J. M., & Gray, P. D. (2008). Affective feedback: An investigation into the role of emotions in the information-seeking process. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 395-402.

Arapakis, I., Moshfeghi, Y., Joho, H., Ren, R., Hannah, D., & Jose, J. M. (2009). Integrating facial expressions into user profiling for the improvement of a multimodal recommender system. In *Proceedings of the IEEE International Conference on Multimedia and Expo,* 1440-1443.

Arapakis, I., Konstas, I., & Jose, J. M. (2009). Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *Proceedings of the 17th ACM International Conference on Multimedia*, 461-470.

Arapakis, I., Konstas, I., Jose, J. M., & Kompatsiaris, I. (2009). Modeling facial expressions and peripheral physiological signals to predict topical relevance. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 728-729.

Arapakis, I., Lalmas, M., Cambazoglu, B. B., Marcos, M. C., & Jose, J. M. (2014). User engagement in online News: Under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology, 65*, 1988-2005.

Arguello, J., Wu, W. C., Kelly, D., & Edwards, A. (2012). Task complexity, vertical display and user interaction in aggregated search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 435-444.

Aula, A., Khan, R. M., & Guan, Z. (2010). How does search behavior change as search becomes more difficult?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 35-44.

Barber, C., Brown, B. H., & Smallwood, R. H. (1984). Dictionary of physiological measurement. Madison, WI: MTP Press.

Barreda-Ángeles, M., Arapakis, I., Bai, X., Cambazoglu, B. B., & Pereda-Baños, A. (2015). Unconscious Physiological Effects of Search Latency on Users and Their Click Behaviour. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 203-212.

Bateman, S., Teevan, J., & White, R. W. (2012). The search dashboard: How reflection and comparison impact search behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1785-1794.

Baumeister, R. F., Vohs, K. D., DeWall, C. N., & Zhang, L. (2007). How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review, 11*, 167-203.

Bian, J., Dong, A., He, X., Reddy, S., & Chang, Y. (2013). User Action Interpretation for Online Content Optimization. *IEEE Transactions on Knowledge and Data Engineering, 25*, 2161-2174.

Bilal, D. (2002). Children's use of the Yahooligans! Web search engine. III. Cognitive and physical behaviors on fully self-generated search tasks. *Journal of the American Society for information science and technology, 53*, 1170-1183.

Bliese, P. (2006). Multilevel Modeling in R (2.2)–A Brief Introduction to R, the multilevel package and the nlme package.

Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, *56*, 71-90.

Borlund, P., Dreier, S., & Byström, K. (2012). What does time spent on searching indicate?. In *Proceedings of the 4th Information Interaction in Context Symposium*, 184-193.

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, *53*, 225-250.

Boucsein, W. (2012). *Electrodermal activity*. Springer Science and Business Media.

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments. *Psychophysiology, 49*, 1017-1034.

Brennan, K., Kelly, D., & Arguello, J. (2014). The effect of cognitive abilities on information search for tasks of varying levels of complexity. In *Proceedings of the 5th Information Interaction in Context Symposium*, 165-174.

Brosschot, J. F., Gerin, W., & Thayer, J. F. (2006). The perseverative cognition hypothesis: A review of worry, prolonged stress-related physiological activation, and health. *Journal of Psychosomatic Research*, *60*, 113-124.

Brownlow, S. (2009). *The effects of recalling and ruminating about previous humiliating experiences on cortisol levels* (Doctoral dissertation, Walden University).

Byström, K., & Hansen, P. (2005). Conceptual framework for tasks in information studies. *Journal of the American Society for Information Science and Technology*, *56*, 1050-1061.

Cai, J., Liu, G., & Hao, M. (2009). The research on emotion recognition from ECG signal. *IEEE International Conference on Information Technology and Computer Science, 1*, 497-500.

Cannon, W. B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology, 39*, 106-124.

Capra, R., Arguello, J., Crescenzi, A., & Vardell, E. (2015) Differences in the Use of Search Assistance for Tasks of Varying Complexity. SIGIR 2015.

Capra, R., Sams, B., & Seligson, P. (2011). Self-Generated Versus Imposed Tasks in Collaborative Search. *Collaborative Information Seeking: Bridging the Gap Between Theory and Practice (CIS).*

Chen, J. V., Lin, C., Yen, D. C., & Linn, K. P. (2011). The interaction effects of familiarity, breadth and media usage on web browsing experience. *Computers in Human Behavior, 27,* 2141-2152.

Chen, H., Wigand, R. T., & Nilan, M. (2000). Exploring web users' optimal flow experiences. *Information Technology & People*, *13,* 263-281.

Clore, G. L. (1994). Why emotions are felt. In Ekman, P., & Davidson, R.J. (Eds.), *The Nature of Emotion: Fundamental Questions*, (pp. 103-111). New York: Oxford University Press.

Cockburn, A., Kristensson, P. O., Alexander, J., & Zhai, S. (2007). Hard lessons: effort-inducing interfaces benefit spatial learning. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1571-1580.

Cole, M. J., Gwizdka, J., & Belkin, N. J. (2011). Physiological data as metadata. In *Workshop on Enriching Information Retrieval, Special Interest Group in Information Retrieval (SIGIR).*

Conklin, J. E. (1951). Three factors affecting the general level of electrical skin-resistance. *The American Journal of Psychology, 64*, 78-86.

Csikszentmihalyi, M. (1991). *Flow: The psychology of optimal experience* (Vol. 41). New York: Harper-Perennial.

Csikszentmihalyi, M. (2014). Flow. In *Flow and the Foundations of Positive Psychology* (pp. 227-238). Springer Netherlands.

De Silva, P. R., Kleinsmith, A., & Bianchi-Berthouze, N. (2005). Towards unsupervised detection of affective body posture nuances. In *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction*, 32-39.

Dirican, A. C., & Göktürk, M. (2011). Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science*, *3*, 1361-1367.

Dollard, J., Miller, N. E., Doob, L. W., Mowrer, O. H., & Sears, R. R. (1939). Frustration and aggression.

Doob, A. N., & Kirshenbaum, H. M. (1973). The effects on arousal of frustration and aggressive films. *Journal of Experimental Social Psychology, 9*, 57-64.

Duin, A. H., & Archee, R. (1997). Distance learning via the World Wide Web: Information, engagement, and community. In S. Selber (Ed.), *Computers and Technical Communication: Pedagogical and Programmatic Perspectives*, (pp. 149-169). Hillsdale, NJ: Erlbaum.

Edwards, A., Kelly, D., & Azzopardi, L. (2015). The Impact of Query Interface Design on Stress, Workload and Performance. *Advances in Information Retrieval*, 691-702.

Ekman, P. (1992). Are There Basic Emotions?. *Psychological Review*, *99*, 550-553.

Ellick, W., Mirza-Babaei, P., Wood, S., Smith, D., & Nacke, L. E. (2013). Assessing user preference of video game controller button settings. *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, 1107-1112.

Epp, C., Lippold, M., & Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 715-724.

Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting With Computers*, *21*, 133-145.

Feild, H. A., Allan, J., & Jones, R. (2010). Predicting searcher frustration. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 34-41.

Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology, 69*, 153.

Fletcher, R. R., Dobson, K., Goodwin, M. S., Eydgahi, H., Wilder-Smith, O., Fernholz, D., ... & Picard, R. W. (2010). iCalm: Wearable sensor and network architecture for wirelessly communicating and logging autonomic activity. *IEEE Transactions on Information Technology in Biomedicine*, *14*, 215-223.

Fogg, B. J., & Nass, C. (1997). Silicon sycophants: the effects of computers that flatter. *International Journal of Human-Computer Studies*, *46*, 551-561.

Fowles, D. C. (1987). Application of a behavioral theory of motivation to the concepts of anxiety and impulsivity. *Journal of research in personality*, *21*, 417-435.

Freeman, G. L. (1940). A method of inducing frustration in human subjects and its influence upon palmar skin resistance. *The American Journal of Psychology,* 117-120.

Frijda, N. H. (1988). The laws of emotion. *American Psychologist*, *43*, 349.

Gentry, W. D. (1970). Sex differences in the effects of frustration and attack on emotion and vascular processes. *Psychological Reports*, *27,* 383-390.

Gilroy, S. W., Cavazza, M. O., & Vervondel, V. (2011). Evaluating multimodal affective fusion using physiological signals. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, 53-62.

Gohm, C. L., Corser, G. C., & Dalsky, D. J. (2005). Emotional intelligence under stress: Useful, unnecessary, or irrelevant?. *Personality and Individual Differences*, *39,* 1017-1028.

Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2013). Automatically Recognizing Facial Expression: Predicting Engagement and Frustration. In *Proceedings of the 6th International Conference on Educational Data Mining*, 43-50.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, *85*, 348-362.

Gwizdka, J., & Lopatovska, I. (2009). The role of subjective factors in the information search process. *Journal of the American Society for Information Science and Technology*, *60*, 2452-2464.

Hassan, A., White, R. W., Dumais, S. T., & Wang, Y. M. (2014). Struggling or exploring?: disambiguating long search sessions. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining,*. 53-62.

Heinström, J. (2006). Broad exploration or precise specificity: Two basic information-seeking patterns among students. *Journal of the American Society for Information Science and Technology*, *57*, 1440-1450.

Helton, W. S. (2004). Validation of a short stress state questionnaire. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1238-1242.

Helton, W. S., Shaw, T., Warm, J. S., Matthews, G., & Hancock, P. (2008). Effects of warned and unwarned demand transitions on vigilance performance and stress. *Anxiety, Stress, & Coping*, *21*, 173-184.

Hernandez, J., Paredes, P., Roseway, A., & Czerwinski, M. (2014). Under pressure: sensing stress of computer users. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, 51-60.

Hertzum, M. (2010). Frustration: A common user experience. In Hertzum, M., and Hansen, M. (Eds.), *DHRS2010: Proceedings of the Tenth Danish Human-Computer Interaction Research Symposium*, 11-14.

Hobfoll, S. E. (1989). Conservation of resources: A new attempt at conceptualizing stress. *American psychologist*, *44*, 513-524.

Hoppmann, T. K. (2009). Examining the 'point of frustration'. The think-aloud method applied to online search tasks. *Quality & Quantity*, *43*, 211-224.

Hwang, M. I., & Thorn, R. G. (1999). The effect of user engagement on system success: A meta-analytical integration of research findings. *Information & Management*, *35*, 229-236.

Imai, K., King, G., & Lau, O. (2009). Zelig: Everyone's statistical software. *R package version*, *3*.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation*, *52*, 3-50.

Ingwersen, P. (2011). The user in interactive information retrieval evaluation. *Advanced Topics in Information Retrieval*, 83-107.

Jacques. R., Precce. J., and Carey. J. T. (1995). Engagement as a Design Concept for Hypermedia. *Canadian Journal of Educational Communications, 24,* 49-59.

Jiang, J., He, D., & Allan, J. (2014). Searching, Browsing, and Clicking in a Search Session: Changes in User Behavior by Task and Over Time. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 607-616.

Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic prediction of frustration. *International Journal of Human-Computer Studies, 65*, 724-736.

Kelly, D., Arguello, J., Edwards, A., & Wu, W. C. Development and Evaluation of Search Tasks for IIR Experiments using a Cognitive Complexity Framework.

Kligfield, P., Gettes, L. S., Bailey, J. J., Childers, R., Deal, B. J., Hancock, E. W., ... & Wagner, G. S. (2007). Recommendations for the Standardization and Interpretation of the Electrocardiogram Part I: The Electrocardiogram and Its Technology A Scientific Statement From the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society Endorsed by the International Society for Computerized Electrocardiology. *Journal of the American College of Cardiology*, *49*, 1109-1127.

Klein, J., Moon, Y., & Picard, R. W. (1999). This computer responds to user frustration. In *CHI'99 Extended Abstracts on Human Factors in Computing Systems*, 242-243.

Kracker, J. (2002). Research anxiety and students' perceptions of research: An experiment. Part I. Effect of teaching Kuhlthau's ISP model. *Journal of the American Society for Information Science and Technology, 53*, 282-294.

Krantz, S. E. (1983). Cognitive appraisals and problem-directed coping: A prospective study of stress. *Journal of Personality and Social Psychology*, *44*, 638.

Kuhlthau, C. C. (1993). A principle of uncertainty for information-seeking. *Journal of Documentation*, *49*, 339-355.

Laird, J. D. (1974). Self-attribution of emotion: The effects of expressive behavior on the quality of emotional experience. *Journal of Personality and Social Psychology*, *29*, 475-486.

Lazarus, R. S., & Alfert, E. (1964). Short-circuiting of threat by experimentally altering cognitive appraisal. *The Journal of Abnormal and Social Psychology, 69*, 195-205.

Lazarus, R. S., & Launier, R. (1978). Stress-related transactions between person and environment. In Pervin, L., and Lewis, M. (Eds.), *Perspectives in Interactional Psychology* (pp. 287-327). New York, NY: Plenum Press.

Lazarus, R. S. & Folkman, S. (1984). *Stress, appraisal, and coping*. New York: Springer Pub. Co.

Lazarus, R. S. (1991). Cognition and motivation in emotion. *American Psychologist, 46*, 352-367.

Lazarus, R. S. (2007). Stress and emotion: A new synthesis. In Monat, A., Lazarus, R.S., and Reevey, G. (Eds.), *The Praeger Handbook on Stress and Coping, Vol 1* (pp. 33-49). Santa Barbara, CA: Praeger.

Lehmann, J., Lalmas, M., Baeza-Yates, R., & Yom-Tov, E. (2013). Networked user engagement. In *Proceedings of the 1st workshop on User engagement optimization*, 7-10.

Lehmann, J., Lalmas, M., Dupret, G., & Baeza-Yates, R. (2013). Online multitasking and user engagement. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, 519-528.

Lehmann, J., Lalmas, M., Yom-Tov, E., & Dupret, G. (2012). Models of user engagement. *User Modeling, Adaptation, and Personalization*, 164-175.

Lee, J. Y., Paik, W., & Joo, S. (2012). Information resource selection of undergraduate students in academic search tasks. *Information Research: An International Electronic Journal*, *17*, 1.

Lewis, M., Ramsay, D. S., & Sullivan, M. W. (2006). The relation of ANS and HPA activation to infant anger and sadness response to goal blockage. *Developmental Psychobiology, 48*, 397-405.

Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management, 44*, 1822-1837.

Lok, C. F., & Bishop, G. D. (1999). Emotion control, stress, and health. *Psychology and Health, 14*, 813-827.

Lopatovska, I. (2009). Searching for good mood: Examining relationships between search task and mood. *Proceedings of the American Society for Information Science and Technology, 46*, 1-13.

Lopatovska, I. (2011). Emotional correlates of information retrieval behaviors. In *IEEE Workshop on Affective Computational Intelligence,* 1-7.

Lopatovska, I., & Arapakis, I. (2011). Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction. *Information Processing & Management*, *47*, 575-592.

Malik, M., Bigger, J. T., Camm, A. J., Kleiger, R. E., Malliani, A., Moss, A. J., & Schwartz, P. J. (1996). Heart rate variability standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, *17*, 354-381.

Mandler, G. (1990). A constructivist theory of emotion. In Stein, N.L., Leventhal, B., & Trabasso, T.R. (Eds.), *Psychological and Biological Approaches to Emotion* (pp. 21-43).

Mansourian, Y. (2008). Coping strategies in web searching. *Program*, *42*, 28-39.

Matthews, G., Campbell, S. E., Desmond, P. A., Huggins, J., Falconer, S., & Joyner, L. A. (1999). Assessment of task-induced state change: Stress, fatigue, and workload components. In Scerbo, M.W., & Mouloua, M. (Eds.), *Automation Technology and Human Performance: Current Research and Trends* (pp. 199-203).

Matthews, G., Warm, J. S., Reinerman, L. E., Langheim, L. K., & Saxby, D. J. (2010). Task engagement, attention, and executive control. *Handbook of Individual Differences in Cognition*, 205-230.

Matsumoto, D., & Sanders, M. (1988). Emotional experiences during engagement in intrinsically and extrinsically motivated tasks. *Motivation and Emotion*, *12*, 353-369.

McCrae, R. R., & Costa Jr, P. T. (1999). A five-factor theory of personality. *Handbook of Personality: Theory and Research*, *2*, 139-153.

McDuff, D., Karlson, A., Kapoor, A., Roseway, A., & Czerwinski, M. (2012). AffectAura: an intelligent system for emotional memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 849-858.

McLaughlin, M., Chen, Y. S., Park, N., Zhu, W., & Yoon, H. (2004). Recognizing user state from haptic data. In *International Conference on Information Systems Analysis and Synthesis*.

Mooney, C., Scully, M., Jones, G. J., & Smeaton, A. F. (2006). Investigating biometric response for information retrieval applications. In *Advances in Information Retrieval*, 570-574.

Morris, D., Morris, M., & Venolia, G. (2008). SearchBar: a search-centric web history for task resumption and information re-finding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1207-1216.

Moshfeghi, Y., & Jose, J. M. (2013). An effective implicit relevance feedback technique using affective, physiological and behavioural features. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval,* 133-142.

Moshfeghi, Y., Matthews, M., Blanco, R., & Jose, J. M. (2013). Influence of timeline and named-entity components on user engagement. In *Advances in Information Retrieval*, 305-317.

Müller, S. C., & Fritz, T. (2015). Stuck and Frustrated or In Flow and Happy: Sensing Developers' Emotions and Progress. ICSE.

Muse, L. A., Harris, S. G., & Feild, H. S. (2003). Has the inverted-U theory of stress and job performance had a fair test?. *Human Performance, 16*, 349-364.

Näätänen, R. (1973). The inverted-U relationship between activation and performance: A critical review.

Nahl, D. (2004). Measuring the affective information environment of web searchers. *Proceedings of the American Society for Information Science and Technology*, *41*, 191-197.

Nahl, D., & Bilal, D. (Eds.). (2007). *Information and emotion: The emergent affective paradigm in information behavior research and theory*. Medford, NJ: Information Today, Inc.

Nes, L. S., Segerstrom, S. C., & Sephton, S. E. (2005). Engagement and arousal: Optimism's effects during a brief stressor. *Personality and Social Psychology Bulletin, 31*, 111-120.

Niu, J., Zhao, X., Zhu, L., & Li, H. (2013). Affivir: An affect-based Internet video recommendation system. *Neurocomputing, 120*, 422-433.

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, 420-423.

O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology, 59*, 938-955.

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

O'Brien, H. L., & Lebow, M. (2013). Mixed‐methods approach to measuring user experience in online news interactions. *Journal of the American Society for Information Science and Technology, 64,* 1543-1556.

O'Brien, H. L., & Toms, E. G. (2013). Examining the generalizability of the User Engagement Scale (UES) in exploratory search. *Information Processing & Management*, *49*, 1092-1107.

Otis, J., & Ley, R. (1993). The effects of frustration induced by discontinuation of reinforcement on force of response and magnitude of the skin conductance response. *Bulletin of the Psychonomic Society, 31*, 97-100.

Ohira, H., Nomura, M., Ichikawa, N., Isowa, T., Iidaka, T., Sato, A., ... & Yamada, J. (2006). Association of neural and physiological responses during voluntary emotion suppression. *Neuroimage*, *29*, 721-733.

Pace, S. (2004). The roles of challenge and skill in the flow experiences of web users. In *Proceedings of the 2004 Informing Science and Information Technology Education Joint Conference,* 341-358.

Pan, M. K., Chang, J. S., Himmetoglu, G. H., Moon, A., Hazelton, T. W., MacLean, K. E., & Croft, E. A. (2011). Now where was I?: physiologically-triggered bookmarking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 363-372.

Papez, J. W. (1937). A proposed mechanism of emotion. *Archives of Neurology & Psychiatry*, *38*, 725-743.

Parkes, K. R. (1994). Personality and coping as moderators of work stress processes: Models, methods and measures. *Work & Stress*, *8*, 110-129.

Partala, T., & Surakka, V. (2004). The effects of affective interventions in human–computer interaction. *Interacting with Computers*, *16*, 295-309.

Poddar, A., & Ruthven, I. (2010). The emotional impact of search tasks. In *Proceedings of the Third Symposium on Information Interaction in Context*, 35-44.

Prinz, J. (2003). Emotion, psychosemantics, and embodied appraisals. *Royal Institute of Philosophy Supplement*, *52*, 69-86.

Reeves, B., & Nass, C. (1996). *The Media Equation: How people treat computers, television, and new media like real people and places*. New York, NY: Cambridge University Press.

Reid, N (2015). Random and Mixed Effects Models [PDF]. Retrieved from: www.utstat.utoronto.ca/**reid**/sta410/mar24.pdf.

Reisenzein, R., & Hofmann, T. (1990). An investigation of dimensions of cognitive appraisal in emotion using the repertory grid technique. *Motivation and Emotion*, *14*, 1-26.

Ren, P., Barreto, A., Gao, Y., & Adjouadi, M. (2013). Affective assessment by digital processing of the pupil diameter. *IEEE Transactions on Affective Computing, 4*, 2-14.

Riche, Y., Henry Riche, N., Isenberg, P., & Bezerianos, A. (2010). Hard-to-use interfaces considered beneficial (some of the time). In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, 2705-2714.

Rickles, W. H., & Day, J. L. (1968). Electrodermal activity in non-palmar skin sites. *Psychophysiology*, *4*, 421-435.

Roth, W. T., Dawson, M. E., & Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, *49*, 1017-1034.

Roseman, I. J., Dhawan, N., Rettek, S. I., Naidu, R. K., & Thapa, K. (1995). Cultural differences and cross-cultural similarities in appraisals and emotional responses. *Journal of Cross-Cultural Psychology*, *26*, 23-48.

Russell, D. M., & Grimes, C. (2007). Assigned tasks are not the same as self-chosen Web search tasks. In *40th Annual Hawaii International Conference on System Sciences, 2007*, 83-83.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*, 145-172.

Sakuragi, S., Sugiyama, Y., & Takeuchi, K. (2002). Effects of laughing and weeping on mood and heart rate variability. *Journal of Physiological Anthropology and Applied Human Science*, *21*, 159-165.

Saponas, T. S., Tan, D. S., Morris, D., & Balakrishnan, R. (2008). Demonstrating the feasibility of using forearm electromyography for muscle-computer interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 515-524.

Savolainen, R. (1995). Everyday life information-seeking: Approaching information-seeking in the context of "way of life". *Library & Information Science Research*, *17*, 259-294.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*, 379-399.

Schaufeli, W. B., & Salanova, M. (2008). Enhancing work engagement through the management of human resources. In Naswall, K., Rellgren, J., & Sverke, M. (Eds.), *The Individual In the Changing Working Life*, (pp. 380-402).

Scheirer, J., Fernandez, R., Klein, J., & Picard, R. W. (2002). Frustrating the user on purpose: A step toward building an affective computer. *Interacting With Computers*, *14*, 93-118.

Selye, H. (1975). Stress without distress. In Serban, G. (Ed.), *Psychopathology of Human Adaptation*, (pp. 137-146).

Selye, H. (1976). The stress concept. *Canadian Medical Association Journal*, *115*, 718.

Seward, J. P., Pereboom, A. C., Butler, B., & Jones, R. B. (1957). The role of prefeeding in an apparent frustration effect. *Journal of Experimental Psychology*, *54*, 445-450.

Solovey, E. T., Zec, M., Garcia Perez, E. A., Reimer, B., & Mehler, B. (2014). Classifying driver workload using physiological and driving performance data: two field studies. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, 4057-4066.

Song, Y., Shi, X., & Fu, X. (2013). Evaluating and predicting user engagement change with degraded search relevance. In *Proceedings of the 22nd International Conference on World Wide Web*, 1213-1224.

Starkweather, J. (2010). Linear mixed effects modeling using R. *Unpublished Manuscript*.

Sun, D., Paredes, P., & Canny, J. (2014). MouStress: detecting stress from mouse motion. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, 61-70.

Sun, Q., & Spears, N. (2011). Frustration Theory: toward an understanding of keyword search effectiveness and consumer responses. *Journal of Customer Behaviour*, *10,* 35-48.

Sundar, S. S., Xu, Q., Bellur, S., Oh, J., & Jia, H. (2011). Beyond pointing and clicking: How do newer interaction modalities affect user engagement?. *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 1477-1482.

Svarre, T., & Lykke, M. (2014). Simulated work tasks: the case of professional users. In *Proceedings of the 5th Information Interaction in Context Symposium*, 215-218.

Szafir, D., & Mutlu, B. (2012). Pay attention!: designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 11-20.

Teevan, J., Collins-Thompson, K., White, R. W., Dumais, S. T., & Kim, Y. (2013). Slow Search: Information Retrieval without Time Constraints. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, 1-10.

Vakkari, P. (2001). A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *Journal of Documentation*, *57*, 44-60.

Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science and Technology*, *37*, 413-464.

Vollrath, M. (2001). Personality and stress. *Scandinavian Journal of Psychology*, *42*, 335-347.

Wang, H., Chignell, M., & Ishizuka, M. (2006). Empathic tutoring software agents using real-time eye tracking. In *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications*, 73-78.

Wang, P., Hawk, W. B., & Tenopir, C. (2000). Users' interaction with World Wide Web resources: An exploratory study using a holistic approach. *Information Processing & Management*, *36*, 229-251.

Ward, R. D., Marsden, P. H., Cahill, B., & Johnson, C. (2002). Physiological responses to well-designed and poorly designed interfaces. In *Proceedings of CHI 2002 Workshop on Physiological Computing*.

Webster, J., & Ahuja, J. S. (2006). Enhancing the design of web navigation systems: The influence of user disorientation on engagement and performance. *MIS Quarterly*, *30*, 661-678.

White, R. W., & Dumais, S. T. (2009). Characterizing and predicting search engine switching behavior. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 87-96.

Wiener, E. L., Curry, R. E., & Faustina, M. L. (1984). Vigilance and task load: In search of the inverted U. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 26*, 215-222.

Wu, W. C., Kelly, D., Edwards, A., & Arguello, J. (2012). Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proceedings of the 4th Information Interaction in Context Symposium*, 254-257.

Xie, I. (2009). Dimensions of tasks: influences on information-seeking and retrieving process. *Journal of Documentation*, *65*, 339-366.

Yuan, J., Ding, N., Liu, Y., & Yang, J. (2014). Unconscious emotion regulation: Nonconscious reappraisal decreases emotion-related physiological reactivity during frustration. *Cognition and Emoti*on, 1-12.

Yun, C., Shastri, D., Pavlidis, I., & Deng, Z. (2009). O'game, can you feel my frustration?: improving user's gaming experience via stresscam. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2195-2204.

Zemack-Rugar, Y., Bettman, J. R., & Fitzsimons, G. J. (2007). The effects of nonconsciously priming emotion concepts on behavior. *Journal of Personality and Social Psychology*, *93*, 927-939.

Zhang, Y. (2008). Undergraduate students' mental models of the Web as an information retrieval system. *Journal of the American Society for Information Science and Technology*, *59*(13), 2087-2098.