Examining the Generalizability of the User Engagement Scale (UES) in Exploratory Search

Heather L. O'Brien*

iSchool, University of British Columbia

1961 East Mall, Vancouver, British Columbia, Canada

Tel: 604-822-6365

Fax: 604-822-6006

h.obrien@ubc.ca

Elaine G. Toms

School of Information, The University of Sheffield

Regent Court, 211 Portobello

Sheffield, United Kingdom S1 4DP

e.toms@sheffield.ac.uk

* Corresponding author

Please cite as: O'Brien, H.L. & Toms, E.G. Examining the Generalizability of the User Engagement Scale (UES) in Exploratory Search. *Information Processing and Management* (In press) DOI: http://dx.doi.org/10.1016/j.ipm.2012.08.005.

Abstract

The user experience is an integral component of interactive information retrieval (IIR). However, there is a twofold problem in the measurement of user experience. Firstly, while many IIR studies have relied on a single dimension of user feedback, that of satisfaction, experience is a much more complex concept. IIR in general, and exploratory search more specifically, are dynamic, multifaceted experiences that evoke pragmatic and hedonic needs, expectations, and outcomes that are not adequately captured by user satisfaction. Secondly, questionnaires, which are typically the means in which user's attitudes and perceptions are measured, are not typically subjected to rigorous reliability and validity testing. To address these issues, we administered the multidimensional User Engagement Scale (UES) in an exploratory search environment to assess users' perceptions of the Perceived Usability (PUs), Aesthetics (AE), Novelty (NO), Felt Involvement (FI), Focused Attention (FA), and Endurability (EN) aspects of the experience. In a typical laboratory-style study, 381 participants performed three relatively complex search tasks using a novel search interface, and responded to the UES immediately upon completion. We used Principal Axis Factor Analysis and Multiple Regression to examine the factor structure of UES items and the relationships amongst factors. Results showed that three of the six sub-scales (PUs, AE, FA) were stable, while NO, FI and EN merged to form a single factor. We discuss recommendations for revising and validating the UES in light of these findings.

1. Introduction

Assessing search systems usually takes a systems-based approach that uses standard information retrieval measures such as precision, recall, and cumulative gain (Järvelin, 2011) to evaluate the output from the system; in interactive information retrieval, evaluation is extended to encompass aspects of user interactivity, where elements of the interaction are measured, including number of queries used and the mean size of a query; performance measures such as number of relevant documents retrieved; and usability measures such as ease of use, effort and preference, among a host of metrics (Kelly, 2009). But how is the user *experience* assessed? In studies of interactive information retrieval, user experience is rarely addressed except in post task questionnaires that inquire about user satisfaction, which is generally evaluated using a single Likert-scaled question: "How satisfied were you …?" (see reports from the Interactive Track at TREC and INEX as examples.) This one-dimensional approach is a limited and imprecise assessment of the rich, multi-dimensional experience that is so typical of user interactivity with any digital product including search engines and exploratory search systems.

Increasingly, researchers are looking to experience-based frameworks as a means of understanding Human Information-Interaction (HII) (O'Brien, 2011b). The concept of user experience (UX) gained prominence in e-commerce and was further popularized in human computer interaction by Norman (2002). UX, defined as "a person's perceptions and responses that result from the use or anticipated use of a product, system or service" (ISO, 2008), represents a more holistic way of approaching people's interactions with technologies than usability:

> An experience is an episode, a chunk of time that one went through—with sights and sounds, feelings and thoughts, motives and actions; they are closely knitted together, stored in memory, labeled, relived and communicated to others (Hassenzahl, 2011, p.8).

UX examines the *quality* of information interactions from the perspective of the user. Like usability, UX is outcomes-based, but this outcome may be tangible (e.g., using your smartphone to text a friend) or intangible (e.g., sharing a joke and feeling connected to that friend). UX also places an emphasis on process. Several researchers have emphasized the idea of "plot" or "story" to describe the way in which an interaction with a system unfolds over time and the affective, cognitive, physical and social aspects of experience (Laurel, 1993; McCarthy & Wright, 2004).

If we begin to view IIR as an experience, then we must re-examine how we measure information searching and retrieval, moving beyond standard metrics of efficiency, effectiveness and user satisfaction to incorporate measures of fulfilment, play and engagement (McCarthy & Wright, 2005). This concept is especially important in exploratory search, which emphasizes learning, discovery, creativity and problem solving (White & Roth, 2009). Exploratory search centres around a complex information need that changes as the searcher encounters and incorporates new information within an information space for the purposes of knowledge acquisition and personal growth. Given the richness of exploratory searching, traditional IIR metrics must be complemented by measures that address learning, discovery, enjoyment and engagement (White & Roth, 2009). In addition to developing such measures, we must also ensure that they are rigorously tested and meet standards of reliability, validity and generalizability in order to accurately reflect the user experience.

The focus of our work has been to define and measure user engagement. User engagement "explain[s] how and why applications attract people to use them" (Sutcliffe, 2010, p. 3). We have found that engagement is a quality of user experience that depends on several factors, including the aesthetic appeal, novelty, and usability of the system, the ability of the user to attend to and become involved in the experience, and the user's overall evaluation of the salience of the experience (O'Brien & Toms, 2008). Engagement depends on the depth of participation the user is able to achieve with respect to each experiential attribute.

The multidimensional nature of user engagement makes it challenging to measure. While we are very comfortable measuring concrete events, such as the number of errors a user makes when interacting with a system or how long it takes to find the answer to a factual search query, we are less firmly seated when it comes to activities for which there are no visible or physical outcomes. Since only the user can evaluate the level of engagement experienced during an interaction with a system, a subjective approach is needed in the development of measures for this construct. We elected to develop a questionnaire, which takes assessment "… away from the usual product-centered towards a more experiential evaluation" (Hassenzahl, 2011, p. 56). Based on an extensive literature review, a qualitative study with users of four types of technologies (video games, e-shopping, e-learning, and web searching), and two large-scale survey studies conducted in the e-shopping domain (O'Brien & Toms, 2008; O'Brien & Toms, 2010a), we developed the User Engagement Scale (UES). The UES is a 31-item questionnaire that taps into six dimensions of experience: Aesthetic Appeal, Novelty, Focused Attention, Felt Involvement, Perceived Usability, and Endurability (i.e., the users' overall impression of the experience).

As part of the scale development and evaluation process, we are interested in generalizing the UES to different research environments, including exploratory search, with the ultimate goal of producing a reliable and valid instrument that can assess user engagement in IIR settings. In the work reported here, we administered the UES to a large group of searchers who interacted with an exploratory search interface to perform decision-making tasks. In the following sections, we elaborate on issues inherent in the measurement in IIR, exploratory search, and user engagement, describe the current study, and present our findings. In light of our results, we discuss implications for the revision, validation, and use of the UES.

2. Literature Review

2.1. Measurement in interactive information retrieval and exploratory search

There are four basic classes of measures commonly employed in Interactive Information Retrieval (IIR): contextual, interaction, performance, and usability (Kelly, 2009). Contextual measures include demographic and socio-cognitive variables (e.g., topic familiarity and search experience), as well as the nature of the search or work task and the setting in which the information interaction occurs. Interaction measures are collected during an IIR session, and are based on users' search strategies (e.g., query construction, number of queries) and their interactions with retrieved documents. Performance measures, such as precision and recall, and usability measures, which gauge users' perceptions of the system and their interactions with it, are outcome oriented and used to evaluate the success of an IIR session (Kelly, 2009). Several researchers have examined the relationship between these four classes of measures. For example, Su (2008) found that usability measures (e.g., users' confidence and satisfaction with the completeness and precision of the retrieved results), along with the value of the search results (i.e., utility), were highly correlated with users' evaluations of system success. Additionally, Al-Maskari and Sanderson (2010) found significant relationships between user satisfaction and user effectiveness (e.g., completeness, accuracy), and user satisfaction and system effectiveness (e.g., precision, recall, relevance). Such studies underscore the user as a fundamental component of IIR evaluation and the need to account for user perceptions and actions in the measurement model.

The usability measures typically seen in IIR studies pertain to users' attitudes toward the IR system or the search results. Although users' may be asked about their confidence in their search abilities or results, level of topical knowledge before and after the search, etc., they are not asked about their affective reactions to the *experience*, specifically their emotional responses to various stages of the interaction (Kelly, 2009). Applying an affective, experiential framework to IIR may be increasingly important to the burgeoning area of exploratory search, which is as much about the journey as the destination (White & Roth, 2009; Dillon & Vaughan, 1997). According to White and Roth, "exploratory search can be used to describe an information-seeking problem context that is open-ended, persistent, and multi-faceted; and to describe information-seeking processes that are opportunistic, iterative, and multi-tactical" (p. 6). Unlike the "look up model" of information retrieval that focuses heavily on task completion and system performance, the outcomes of exploratory search, such as learning, knowledge discovery and personal growth, are less tangible.

As such, traditional measures of IIR, such as efficiency and effectiveness, may not be adequate for evaluating exploratory search. White and Roth (2009) highlight the challenges inherent in evaluating exploratory search interactions, where users may experience varying degrees of uncertainty during the search process and interact with different systems over multiple search sessions. They state that evaluation remains an understudied area of exploratory search, and approaches to its measurement must take into account user behaviours and cognition, as well as subjective assessments of user satisfaction, search tasks, content, and felt engagement. We are interested in the engagement piece of this measurement puzzle.

2.2. User engagement

User engagement is a quality of user experience that describes a positive human-computer interaction. User engagement has been equated with user satisfaction (Quesenbury, 2003), but previous work has demonstrated that it is much more than this (O'Brien & Toms, 2008). While engagement encompasses users' attitudes toward systems (e.g., usability, aesthetic appeal), it also focuses on individual users' thoughts (Laurel, 1993), feelings (Jacques, Carey & Preece, 1995), and their degree of activity (Laurel, 1993; Norman, 1986) during system use. Through a systematic review of interdisciplinary literature and an exploratory study with technology users, we articulated a number of engagement attributes, including perceived attention, challenge, feedback, control, novelty, interest, motivation, and affective and sensory appeal (O'Brien & Toms, 2008). Contextual variables may also be important in user engagement, specifically the social nature of interacting with technologies for the purposes of sharing information or experiences with other people (O'Brien, 2011b).

In addition to the attributes of engagement, we developed a process-based model to describe the way in which engagement fluctuates during the course of an interaction. Users experience a point of engagement, a period of sustained engagement, disengagement, and re-engagement during system use; at each of these stages and with different types of systems, the attributes of engagement may manifest to different degrees (O'Brien & Toms, 2008). For example, the aesthetic appeal of a search system may be more important at the beginning of an interaction in order to capture users' attention, whereas e-shoppers may require high levels of aesthetic appeal throughout a shopping encounter in order to evaluate and form attachments to products.

2.3. Measuring user engagement

Given the multifaceted nature of user engagement and its emphasis on the interaction process, it is an appropriate construct to apply in exploratory search environments. Recent research has focused on pinpointing behavioural indicators of user engagement in large-scale web studies. Using metrics such as dwell time, session utility, interaction duration, number of distinct and returning users, number of page visits, bounce rate, and short- and long-term interaction time (Lehmann, Lalmas, Yom-Tov & Dupret, 2012; Singla & White, 2010), researchers are attempting to construct models of user engagement that take into account the diversity of websites and users. However, not all of these performance metrics may be useful indicators of user engagement with a particular website or for a particular user (Lehmann, Lalmas, Yom-Tov & Dupret, 2012). Self-report measures can complement and aid in the interpretation of behavioural metrics, taking context and individual differences into account.

However, self-report measures are susceptible to demand effects, social desirability, and acquiescence (Kelly, Harper & Landau, 2008). Even more problematic, "most of the questionnaires and scales that are used in IIR do not have established validity and reliability and are often developed ad-hoc" (Kelly, 2009, p. 180). This has motivated our work to develop a self-report instrument to measure user engagement and to test its reliability, validity, and generalizability in information-rich environments.

The User Engagement Scale (UES) built on existing research by Webster and Ho (1997) and Jacques, Carey and Preece (1995) that explored engagement with educational multimedia systems. Both sets of researchers developed and administered questionnaires that included some attributes of engagement (e.g., users' perceptions of challenge, attention, feedback, variety, curiosity, and intrinsic interest). However, our previous research indicated that there may be additional attributes that inform engagement, such as positive affect, endurability, aesthetic and sensory appeal, and interactivity (O'Brien & Toms, 2008). This led us to question the completeness of previous questionnaires and whether there were potential interdependencies amongst the attributes, resulting in the construction of our own scale.

The scale development process, as prescribed by DeVellis (2003), is explained in depth in O'Brien and Toms (2010a), but we summarize it here briefly. Firstly, we identified attributes (e.g., challenge, control, feedback, etc.) that characterized user engagement, and items (i.e., statements to which users would respond) to support each attribute from existing scales and interviews with video gamers, online shoppers, web searchers, and e-learners (O'Brien & Toms, 2008). This process resulted in approximately 400 items, which were systematically assessed to remove duplicates, clarify wording, and evaluate the "fit" of the item with the construct it was intended to measure. Following this screening process, items were pre-tested. The remaining 186 items were administered in a web-based survey to 440 general e-shoppers. Reliability analysis assessed the internal consistency of the attribute-based sub-scales (Aladwani & Preshant, 2002). Exploratory Factor Analysis (EFA), a data reduction technique, identified the most parsimonious set of items and grouped these together according to six underlying factors or dimensions (Table 1).

Table 1
Factors of Engagement and their Definitions

Factor	Definition
Aesthetic Appeal (AE)	The users' perception of the visual appearance of a computer application interface.
Endurability (EN)	Users' overall evaluation of the experience, its perceived success and whether users would recommend the e-shopping site to others. This factor combines concepts related to users' likelihood to return (Webster & Ahuja, 2006) and evaluation of system success (DeLone & McLean, 2003).
Felt Involvement (FI)	Users' feelings of being drawn in, interested, and having fun during the interaction.
Focused Attention (FA)	The concentration of mental activity (Matlin, 1994); contained some elements of Flow, specifically focused concentration, absorption, and temporal dissociation (Csikszentmihalyi, 1990).
Novelty (NO)	Users' level of interest in the task and curiosity evoked by the system and its contents.
Perceived Usability (PUs)	Users' affective (e.g., frustration) and cognitive (e.g., effort) responses to the system.

Secondly, we tested the validity and factor structure of the scale with a sample of 802 shoppers of a specific online book retailer (who requested anonymity). This analysis used Structural Equation Modelling (SEM) to perform Confirmatory Factor Analysis (CFA) and Path Analysis (PA). The purpose of CFA was to verify the six-factor structure of the scale, while PA examined the relationships amongst the factors (Fig. 1). PA showed that Perceived Usability mediated the relationships between Aesthetic Appeal, Novelty, Focused Attention, Felt Involvement and Endurability. In other words, Perceived Usability was an important variable in predicting e-shoppers lasting impressions of the experience as worthwhile, rewarding, etc. Aesthetic Appeal and Novelty were predictors in the model, indicating whether users would chose to invest their attention and become involved in the e-shopping encounter. Overall, users' level of Felt Involvement predicted perceptions of system usability and overall evaluations of the experience (i.e., Endurability).

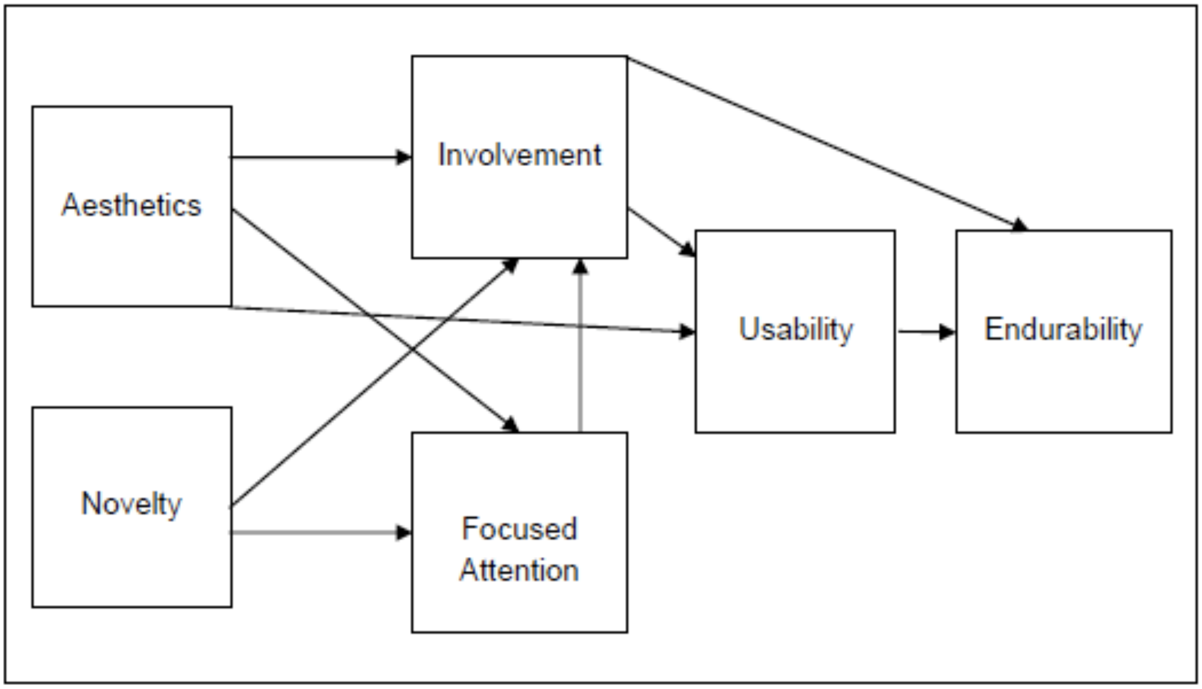


Fig. 1. Path model of UES factors in e-shopping domain

The outcome of this scale development and evaluation process in the e-shopping domain was a 31-item instrument with six underlying dimensions or factors (see Table 1).

2.4. Applications of the User Engagement Scale (UES)

Scale development is a longitudinal process, and an important consideration for any measure is its generalizability, or its "administrative viability and interpretation in different research situations" (Peterson, 2000, p. 79-80). To date, the UES (or components of it) has been administered in studies with different types of technologies, including an archival webcast system (O'Brien & Toms, 2010b), the social networking application, Facebook (Banhawi & Mohamed, 2011), and a simulated travel agency website (Hyder, 2011), and has been examined for its reliability and validity in these settings.

O'Brien and Toms (2010b) examined the factor structure of the UES in a study that asked participants to perform fact finding and content summary tasks using a multimedia webcast system. Exploratory Factor Analysis (EFA) resulted in six factors, but the composition of these factors was different than in previous research (O'Brien & Toms, 2010a). While Aesthetics, Focused Attention, Novelty and Endurability remained stable constructs, the Felt Involvement sub-scale was eliminated and the Perceived Usability sub-scale became two factors: one factor contained affective items (e.g., "discouraged) and the other consisted of items relating to the effort required to use the application. The clustering of affective items related to the Perceived Usability of the system was also observed in the work of Banhawi and Mohamed (2011) who administered the UES to a sample of Facebook users. In this study, EFA revealed a four-factor model: Focused Attention, Perceived Usability (affective items only) and Aesthetic Appeal emerged as distinct factors. However, items from the Endurability, Novelty and Felt

Involvement sub-scales loaded together on one factor, suggesting that evaluations of Facebook as worthwhile, fun, and stimulating were perceived along the same dimension. This same four-factor model emerged when O'Brien (2010) examined the relationship between e-shoppers UES scores and shopping motivations using an exploratory, rather than confirmatory, factor analysis approach.

In addition to studies that have looked at the factor structure of the UES, Hyder (2011) investigated the criterion validity of the UES in his study of user engagement with a simulated online travel agency. The UES was combined with items from other scales, and some items within the sub-scales were extracted and grouped with items from different UES constructs. For example, the Aesthetics sub-scale and four items from the Perceived Usability sub-scale were tested as "antecedents of engagement;" items from the Focused Attention and Novelty sub-scales formed a measure of "curiosity," and four of the Endurability items formed part of the "measurement of value" (p. 352). UES items were used successfully with other psychometric and behavioural measures to examine website engagement. In addition, Hyder examined the relationship between engagement variables and, concurrent with our work (O'Brien & Toms, 2010a), found that aesthetics predicted focused attention, involvement, and elements of perceived usability (control, challenge). In addition, elements of Endurability were embedded in Hyder's outcome measures of "perceived value" and "return intention."

In summary, applications of the UES in different research environments have demonstrated the reliability of the Aesthetic and Focused Attention sub-scales. Perceived Usability has been an internally consistent sub-scale, though there is some evidence to suggest that affective and cognitive items may be distinct dimensions in some circumstances. In one study, Novelty items loaded together (O'Brien & Toms, 2010b), while in other research these items combined with Felt Involvement and Endurability items to form one factor (O'Brien, 2010; Banhawi & Mohamed, 2011). Hyder's (2010) results demonstrated significant correlations between UES items and other psychometric scales and behavioural measures, and found similar predictive relationships between engagement variables. Thus, some aspects of the UES have been generalized across contexts, while others have not.

2.5. Current study

The current study examines the generalizability of the UES in an exploratory search environment. We administered the UES to participants using an exploratory search system in a laboratory setting to complete complex search tasks. Consideration of the contextual elements of any one human-information interaction quickly highlights the potential impact of task (e.g., externally versus internally motivated), user (e.g., alert versus fatigued), and situational (e.g., alone versus with friends) variables on the user experience. This study was similar to the wikiSearch study (O'Brien & Toms, 2010b) in that participants were completing research-generated tasks in a laboratory setting with a novel interface, so we might expect to see the same factor solution as observed in this setting. However, in an exploratory search environment, we might also expect to see the same six-factor structure that emerged in the e-shopping domain, since tasks conducted in both environments involved gathering information (i.e., comparing products or information sources) and using a combination of searching and browsing strategies.

In this paper, we investigate the internal consistency and factor structure of the UES based on users' perceptions of their experiences with a specific exploratory search system, with the goal of demonstrating its generalizability to the exploratory search domain. Based on previous studies, we expected the Aesthetic Appeal and Focused Attention sub-scales to remain stable, but were less certain about the composition of the Perceived Usability sub-scale, or whether Novelty, Felt Involvement and Endurability would emerge as distinct factors.

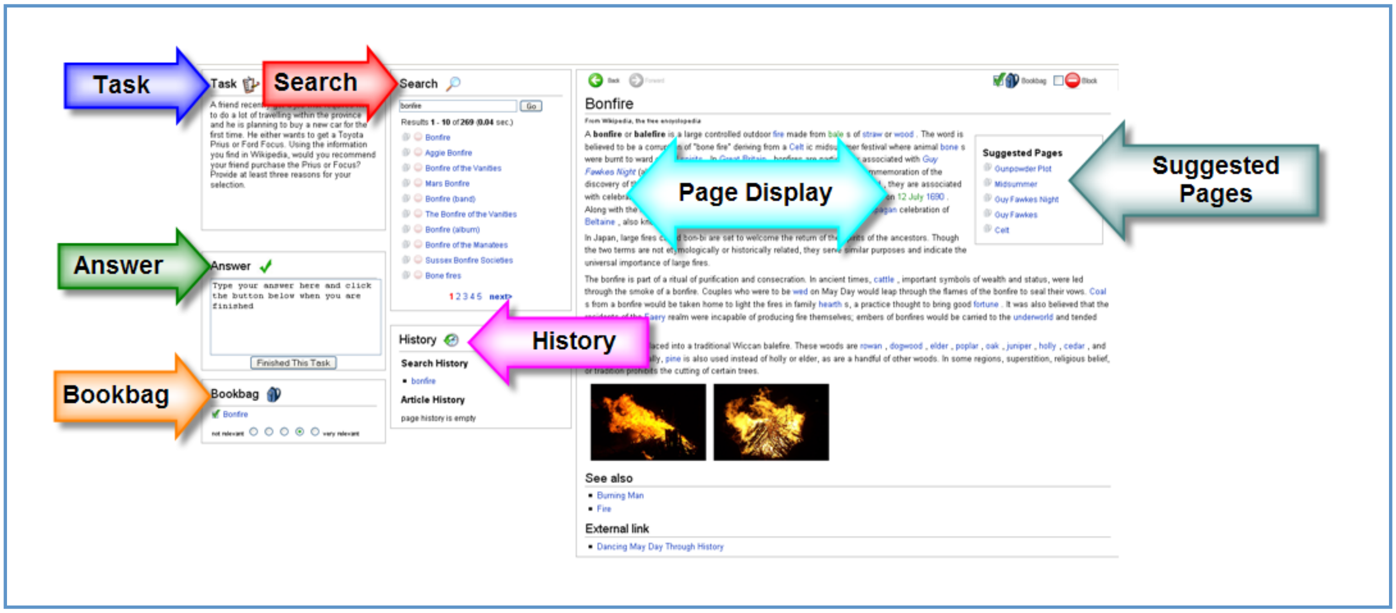
3. Method

The UES was administered at the end of a large laboratory experiment that examined how people performed complex search tasks using a specialized interface to a locally stored version of Wikipedia. In this within-subjects design, three experimental tasks were randomly assigned to participants; each task was designed so that more than one page was required to respond to the task and participants were expected to make and submit their decision identifying which pages were most suitable to address the search scenario.

3.1. Search application

The wikiSearch system has been used in several studies and is fully described in Toms, McCay-Peet and Mackenzie (2009), which is summarised here. The interface (see Fig. 2) was "flattened" into in a three-column representation of the multiple webpage that is typically found in search system. It provided access to a version of the Wikipedia that was used in the INEX Interactive Track.

The three-sections of the interface pertained to (from left to right) task-based activities, search procedures, and detailed views of documents. The Task (extreme left) section contained the assigned experimental task, a "Bookbag" to collect information and a Notebook or Answer pane. The second column, the Search section, contained a search box, a history section to display all queries entered and pages viewed, and the search results, which appeared as an abbreviated list but featured mouseover access to page descriptions. The third column, Page Display, contained the content, a scrollable wiki page with internal and external links and a text box with Suggested Pages that linked to other articles. The Suggested Pages links were created dynamically by entering the first paragraph of the displayed page as a search query. The intent behind the design was to follow some of Shneiderman's (1998) design principles including reducing the number of mouse-clicks, leaving the user in control, and reducing memory load.



Task

A friend recen[illegible] to do a lot of travelling within the province and he is planning to buy a new car for the first time. He either wants to get a Toyota Prius or Ford Focus. Using the information you find in Wikipedia, would you recommend your friend purchase the Prius or Focus? Provide at least three reasons for your selection.

Search

bonfire

Results **1 - 10** of **269** (**0.04** sec.)

- Bonfire
- Aggie Bonfire
- Bonfire of the Vanities
- Mars Bonfire
- Bonfire (band)
- The Bonfire of the Vanities
- Bonfire (album)
- Bonfire of the Manatees
- Sussex Bonfire Societies
- Bone fires

1 2 3 4 5 next>

Answer

Type your answer here and click the button below when you are finished

Bookbag

✓ Bonfire

not relevant ○ ○ ○ ● ○ very relevant

History

Search History

- bonfire

Article History

page history is empty

Bonfire

From Wikipedia, the free encyclopedia

The bonfire is part of a ritual of purification and consecration. In ancient times, cattle , important symbols of wealth and status, were led through the smoke of a bonfire. Couples who were to be wed on May Day would leap through the flames of the bonfire to seal their vows. Coal s from a bonfire would be taken home to light the fires in family hearth s, a practice thought to bring good fortune . It was also believed that the

Suggested Pages

- Gunpowder Plot
- Midsummer
- Guy Fawkes Night
- Guy Fawkes
- Celt

See also

- Burning Man
- Fire

External link

- Dancing May Day Through History

Fig. 2. WikiSearch interface

3.2 Task

The 12 tasks used required participants to make a decision between two options based on a set of criteria. Tasks were crafted so that the participants would either look at two items holistically or using a set of pre-ordained criteria. Each task included a brief background, e.g., "Your friend has decided that he wants to get a dog…;" provided two alternatives, e.g., "he is trying to decide between two dog breeds, Siberian Husky and German Shepherd;" and asked participants to make a choice between the two alternatives. The tasks followed the two patterns below:

Type A. "Your friend has decided that he wants to get a dog. He's never owned a dog before and he is trying to decide between two dog breeds, Siberian Husky and German Shepherd. Using the information you find in Wikipedia, would you recommend that your friend get a Siberian Husky or a German Shepherd? Provide at least three reasons for your decision".

Type B. "Your friend has decided that he wants to get a dog. He's never owned a dog before and he is trying to decide between two dog breeds, Siberian Husky and German Shepherd. Using the information you find in Wikipedia, you offer to answer three of your friend's questions: *Which dog breed is easier to train? *Which dog breed is more prone to health problems? *Which dog has the longer lifespan? Based on this information, would you recommend your friend select a Siberian Husky or a German Shepherd?"

Participants were assigned three of these tasks of either Type A or Type B. Tasks were randomly assigned but counter balanced across the participant group.

3.2. User Engagement Scale (UES)

The wording of the UES (see O'Brien & Toms, 2010a, for the complete set of questions) was modified to fit the current search environment, i.e., the word, "shop" or "shopping" was replaced with "search" or "searching." The UES must be evaluated in its current form for the purposes of validity; thus only small adaptations were made to the items. Participants indicated, for each question, the extent to which they agreed with each statement about their web searching experience and use of wikiSearch using a 7-point Likert scale from strongly disagree (1) to strongly agree (7).

3.3. Participants

Participants (N=381) ranged in age from 18 to 64, though 73.8% were 18-24 years old (M=21, SD=7.1), and the second largest group were between the ages of 25 and 34 (12.6%). There were approximately the same number of males (52.8%) as females (47.1%). Most participants were students, although some indicated they were also employed in other capacities (13.1%). The majority of participants (N=344, 90.3%) indicated that they used search engines one or more times every day. In addition, they were frequent users of Wikipedia: almost three quarters of the sample used Wikipedia one or more times per week (N=201; 52.8%) or per day (N=74; 19.4%); the remaining participants said they used Wikipedia one or more times per month (N=75; 19.7%), per year (N=15; 3.9%), or never (N=14; 3.7%). Participants were recruited through university email lists and signs advertising the study.

3.4. Procedure

Data collection took place in a laboratory setting in a seminar room where five to ten people could participate individually using the same laptop model. The experiment was presented using a modified version of WiIRE (Web Interactive Information Retrieval Experimentation) (Toms, Freund & Li, 2004) which contained study instructions, the consent form, and demographic, pre-task, post-task and post-session questionnaires. The system automatically assigned tasks to each participant, provided a tutorial of the wikiSearch system, and integrated the wikiSearch interface with the experiment. The WiIRE system guided participants seamlessly through the process such that no researcher engagement was required. The UES was administered post-session after participants completed three complex search tasks. At the end of the study, participants were thanked for their time, debriefed and given a small honorarium. A research assistant was present at all times to introduce the study, respond to questions, and oversee participants' activities. The research reported here examines only the analysis of UES data.

3.5. Data analysis

The data was collected into a mySQL database, and exported to SPSS for analysis. While 427 participants were processed, some participants' data was removed because they were pilot participants, had technical difficulties, did not complete the experiment, or did not fully engage in what they were asked to do. After screening the data, 381 respondents remained who on average completed each task in approximately 7 ½ minutes

In preparation for data analysis, eight of the UES items were reverse-coded. Next, descriptive statistics were examined, namely the frequencies of valid responses, means and standard deviations of each item, and inter-item correlations (DeVellis, 2003). Response rates to the majority of items ranged from 96.8% (12 missing values) to 100% (no missing values). The exception to this was the Focused Attention item, "During this search experience, I let myself go," which was not answered by 31 (8.2%) of the participants. This item uses a colloquial expression that may not have resonated with all participants. However, more than 90% of participants did respond to the item and thus it was retained at this point in the analysis.

In order to include all 381 cases in the analysis, we elected to replace missing values using regression imputation. This method was selected after reviewing various imputation methods. We used random regression imputation since linear regression imputation may affect the shape of the distribution and relationships between variables not included in the regression model (Little & Rubin, 1989; Durrant, 2005). Although regression imputation relies on a predicted rather than actual value, this disadvantage was not as significant as those of other methods, such as using the variable mean to replace missing values, which has a tendency to compress variable distributions (Little & Rubin, 1989; Durrant, 2005).

Item means ranged from 2.56 to 6.29 on the 7-point Likert scale and standard deviations ranged from 0.83 to 1.39; thus the data demonstrated some variability but few responses toward the "extremes" of the scale. Examining the inter-item correlations, several observations were made. Firstly, there were low to moderate (>0.1 - <0.6) significant correlations amongst items from the Endurability (EN), Aesthetics (AE), Felt Involvement (FI), and Novelty (NO) sub-scales. Secondly, Focused Attention (FA) and Perceived Usability (PUs) exhibited low to moderate correlations (some of which were significant) with other sub-scales, but there were negative associations between items from these sub-scales. Namely, the Focused Attention (FA) item, "During this search experience, I let myself go" was negatively and significantly correlated with PUs items, "Using wikiSearch was mentally taxing" (r=-0.18, p=0.000) and "This search experience was demanding" (r=-0.16, p=0.001); this was also true for the FA item, "When I was searching, I lost track of the world around me" (r=-0.12, p=0.01; r=-0.16, p=0.001, respectively). As a result of this finding and the lower response rate for these FA items compared to other items, we removed them from the analysis. However, the existence of other negative correlations (though non-significant) between items from these two sub-scales indicated that there may be issues with inter-correlation analysis of the sub-scales.

4. Results

4.1. Reliability analysis

The reliability of the sub-scales was assessed using Cronbach's alpha; values ranging from 0.7 to 0.9 were considered optimal (DeVellis, 2003). Table 2 displays the Cronbach's alpha, mean, and standard deviation values of each sub-scale. Sub-scales are defined here according to the six factors and the items associated with these factors in O'Brien and Toms (2010a). AE, PUs, EN, FA and FI demonstrated "very good" to "excellent" values for Cronbach's alpha and no items were eliminated from these sub-scales. The alpha value for the NO sub-scale (0.69) was in the "minimally acceptable" range. Eliminating the item, "I continued to use wikisearch out of curiosity," improved the internal consistence of the sub-scale.

Table 2
Reliability analysis and descriptive statistics for sub-scales

Sub-scale	No. Items	Cronbach's alpha	Mean	Standard deviation
Aesthetics (AE)	5	0.88	3.76	0.83
Focused Attention (FA)	5	0.79	3.00	0.83
Felt Involvement (FI)	3	0.72	3.47	0.78
Perceived Usability (PUs)	8	0.86	5.61	0.75
Novelty (NO)	2	0.73	3.57	0.88
Endurability (EN)	5	0.8	4.18	0.71

Means were calculated by summing participants' ratings of items within each subscale and dividing by the total number of items for that sub-scale; these individual scores were then calculated to obtain means and standard deviations for each sub-scale. The means for AE, FA, FI and NO were appropriate for a 7-point Likert scale (i.e., toward the mid-point), but the average ratings of the EN (M=4.18, SD=0.88) and PUs (M=5.61, SD=0.75) were relatively high. The Shapiro-Wilk test of normality indicated that the data were not normally distributed.

Next, correlations amongst the UES sub-scales were examined (Table 3). Significant correlations were observed between most sub-scales. Low to moderate associations (<0.5) demonstrated that the sub-scales should remain distinct during factor analysis; correlations above 0.5, as observed between EN and PUs, EN and FI, and NO and FI, indicated that some of the items within these sub-scales may load on more than one factor. There was a negative correlation between the FA and PUs sub-scales (*r*=0.01, *p*=0.7). This indicated that these sub-scales represented distinct dimensions of user experience, but that their relationship may require further exploration.

Table 3
Inter-correlations of UES sub-scales

Sub-scale	Aesthetics (AE)	FA	FI	PUs	NO
Focused Attention (FA)	0.2*				
Felt Involvement (FI)	0.47*	0.48*			
Perceived Usability (PUS)	0.36*	-0.01	0.36*		
Novelty (NO)	0.36*	0.35*	0.64*	0.33*	
Endurability (EN)	0.52*	0.22*	0.6*	0.69*	0.54*

*p<0.001

4.2. Factor analysis

A visual examination of the scree plot and eigenvalues indicated a four or five factor solution (see Fig. 3). Principal axis factor analysis (PAF) with promax rotation was performed. PAF was selected as the method of extraction because the data were not normally distributed. With regard to rotation, orthogonal rotations (e.g., varimax) are commonly performed and noted for simplicity. However, Reise, Waller and Comfrey (2000) advocate for the use of oblique rotation, including promax. They reason that, in oblique rotations, factors are permitted to correlate, which can not only provide researchers with valuable information about the nature of the relationships amongst factors, but is also a more realistic portrait of psychological variables. In addition, oblique rotations are preferred when the goal is factor replicability (Reise, Waller & Comfrey, 2000), making it suited to generalizability studies.

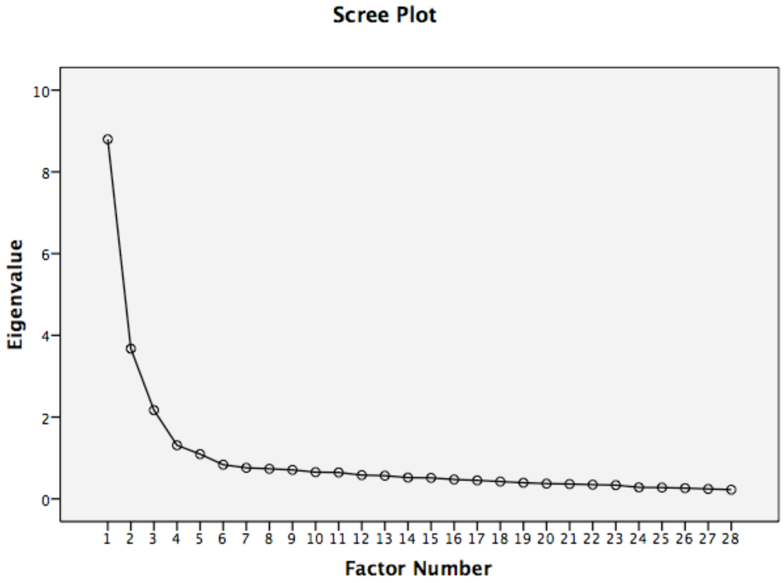


Fig. 3. Scree plot of the factor solution

The Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO = 0.92) indicated that there were distinct factors amongst the items, and Bartlett's Test of Sphericity (x^2=5218.43, df=378, p<0.001) was significant, suggesting there were relationships amongst the factors. In the first iteration, a five-factor solution was obtained that accounted for 60.8% of the total variance. In the five-factor solution, the PUs and one EN item loaded on Factor 1. However, four PUs items also loaded on Factor 5: "I found wikiSearch confusing to use" (0.3); "Using wikiSearch was mentally taxing" (0.49); "This search experience was demanding" (0.43); and "I felt in control during this search experience" (0.25). As a result of these cross-loadings and the small amount of variance (3.9%) the fifth factor contributed to the model, a four-factor solution was specified.

The four-factor model (Table 4) accounted for 56.97% of the total variance. The PUs items and one EN item formed Factor 1, accounting for 31.4% of the variance. The remaining EN, FI and NO items loaded together on Factor 2, which contributed 13.1% to the total variance. AE items (Factor 3) contributed to 7.75% of the variance. FA items made up Factor 4, accounting for 4.68% of the total variance. There were some variables that loaded on multiple factors, specifically the FA item, “I was absorbed in my search task” (Factors 2 and 4), and three EN items, “I would recommend wikiSearch to my friends and family,” “I consider my search experience a success,” and “Searching using wikiSearch was worthwhile” (Factors 1 and 2). Although these variables loaded more strongly on one factor, there presence on two factors is worthy of further discussion.

Table 4
Principal axis factoring with promax rotation of four-factor model

	UES Sub-scale	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
I felt discouraged while using wikiSearch.	PUs	0.801	0.012	-0.023	0.017
I felt frustrated while using wikiSearch.	PUs	0.793	0.079	-0.095	-0.014
I felt annoyed with using wikiSearch.	PUs	0.739	0.119	-0.06	-0.026
This search experience did not work out the way I had planned.	EN	0.703	-0.047	0	0
I could not do some of the things I needed to do using wikiSearch.	PUs	0.682	-0.03	-0.065	0.006
I found wikiSearch confusing to use.	PUs	0.671	-0.132	0.117	0.03
Using wikiSearch was mentally taxing.	PUs	0.628	-0.165	0.059	0.043
This search experience was demanding.	PUs	0.594	-0.176	-0.02	0.039
I felt in control of the searching experience.	PUs	0.497	0.155	0.089	-0.058
I felt interested in my searching tasks.	NO	-0.073	0.81	-0.037	-0.037
The content of wikiSearch incited my curiosity.	NO	-0.153	0.81	-0.112	-0.01
My search experience was fun.	FI	-0.038	0.653	0.048	0.009
I felt involved in the searching tasks.	FI	-0.089	0.609	0.08	0.087
My search experience was rewarding.	EN	0.017	0.603	0.077	0.041
I would recommend wikiSearch to my friends and family.	EN	0.224	0.559	0.068	-0.114
I was really drawn into my searching tasks.	FI	0.011	0.526	0.029	0.284
I consider my search experience a success.	EN	0.372	0.486	-0.057	-0.014
Searching using wikiSearch was worthwhile.	EN	0.222	0.441	0.179	-0.008

The screen layout of wikiSearch appealed to my visual senses.	AE	0.046	-0.165	0.907	0.096
The screen layout of wikiSearch appealed to my visual senses.	AE	-0.013	-0.036	0.868	-0.015
The wikiSearch interface is aesthetically appealing.	AE	0.001	0.014	0.804	0.007
The wikiSearch interface is attractive.	AE	-0.019	0.122	0.791	-0.101
I liked the graphics and images used by wikiSearch.	AE	-0.074	0.162	0.509	-0.023
I was so involved in my searching task that I lost track of time.	FA	0.021	-0.055	0.011	0.77
The time I spent searching just slipped away.	FA	0.093	-0.055	-0.015	0.721
I lost myself in this searching experience.	FA	-0.077	0.087	-0.081	0.635
I blocked out things around me when I was using wikiSearch.	FA	0.016	-0.026	0.04	0.61
I was absorbed in my searching task.	FA	-0.026	0.28	0.035	0.503

The internal consistency of the resulting factors was examined (Table 5). The Cronbach's alpha values were in the very good range, suggesting that the resulting factors were reliable. The means and standard deviations remained unchanged for the Aesthetics and Focused Attention Factors, but were recalculated for Factors 1 and 2. Averages were based on participants' ratings for the nine items that comprised each of these factors.

Table 5
Reliability analysis of four UES factors

Factor	No. Items	Cronbach's alpha	Mean	Standard deviation
1: Perceived Usability	9	0.87	5.64	0.75
2: Novelty, Felt Involvement, Endurability	9	0.87	3.62	0.68
3: Aesthetics	5	0.88	3.76	0.83
4: Focused Attention	5	0.79	3.00	0.83

Correlation analysis demonstrated significant moderate correlations amongst the resulting factors (Table 6). The exception to this was the negative, non-significant relationship between FA and PUs. It is interesting to note that FA and PUs were both correlated with the other dimensions, but not with each other.

Table 6
Correlation analysis of four UES factors

Factor	1	2	3
1: Perceived Usability	1		
2: Novelty, Felt Involvement, Endurability	0.52*	1	
3: Aesthetic Appeal	0.36*	0.54*	1
4: Focused Attention	-0.01	0.42*	0.2*

*p<0.01

5.4. Multiple Regression Analysis

The factor analysis of the wikiSearch data produced a four-factor solution, rather than the six-factor solution obtained in the e-shopping domain. As such, it was not possible to confirm the path model obtained in previous research using structural equation modelling (Fig. 1). However, we were interested in how the four factors related to each other. Based on the correlation analysis of the four factors, we selected Factors 1 (PUs), 3 (AE) and 4 (FA) as predictor variables and Factor 2 (NO, FI, and EN) as the criterion variable. The simultaneous method was used to determine if Factors 1, 3 and 4 would successfully predict Factor 3. The regression model was statistically significant, with the three predictor variables accounting for 55% of the variance in the criterion variable (Adjusted R^2=0.55; $F_{3,377}$ =153.82, p=0.000). Table 7 shows the contribution of each predictor variable to the model (Standardized Beta Coefficient) and to the criterion variable (*t* value).

Table 7
Multiple regression analysis

Predictor variable	Beta	*t*	*p*
Aesthetics	0.32	8.42	0.000
Focused Attention	0.36	10.33	0.000
Perceived Usability	0.41	11.21	0.000

5. Discussion

5.1. Factor structure of the UES

The original sub-scales of the UES (AE, FA, FI, PUs and NO) demonstrated good internal consistency prior to factor analysis. However, factor analysis resulted in a four- or five-factor model. After observing that PUs items pertaining to cognitive effort loaded with other PUs items on Factor 1 and formed their own factor (5), we specified a four-factor solution. While FA, AE, and PUs remained distinct factors, items from the NO, FI, and EN sub-scales converged to form one factor.

Table 8 compares the findings of studies that have employed the UES in its entirety. The four-factor model that emerged in the current study is different from the six-factor model first observed in the general shopping environment (O'Brien & Toms, 2010a), but concurs with the findings of Banhawi and Mohamed (2011) and O'Brien (2010). Across all applications of the UES, AE, FA, and PUs have remained integral factors. However, Banhawi and Mohamed (2011) found that only the affective items (e.g., "I felt frustrated…") made up the PUs component of Facebook use, and O'Brien and Toms (2010b) observed PUs items loading on two distinct factors that distinguished affective and cognitive components of multimedia webcast use. In three administrations of the UES (O'Brien, 2010; Banhawi & Mohamed, 2011; current study), NO, FI and EN items merged to form one factor.

Table 8
Comparison of factor analysis across studies using the UES

UES Factors	E-Shopping (O'Brien & Toms, 2010a)	E-Shopping (O'Brien, 2010)	Facebook (Banhawi & Mohamed, 2011)	Webcast (O'Brien & Toms, 2010b)	wikiSearch
Perceived Usability	Yes	Yes	Yes	Yes	Yes
Aesthetics	Yes	Yes	Yes	Yes	Yes
Focused Attention	Yes	Yes	Yes	Yes	Yes
Novelty	Yes	Merged to form one factor	Merged to form one factor	Yes	Merged to form one factor
Felt Involvement	Yes			No	
Endurability	Yes			Yes	
Number of items	31	26	26	19	28

The Novelty sub-scale contains items pertaining to curiosity in the content of the application and interest in the task. The Felt Involvement sub-scale addresses users' felt involvement in their task and their overall assessment of the experience as fun. Lastly, Endurability items ask users to consider the outcome of their experience with the application, i.e., whether it was successful, worthwhile, and rewarding, as well as whether they would recommend it to family and friends. These sub-scales have demonstrated good internal consistency prior to factor analysis in the majority of studies that have employed the UES. One exception to this was the NO sub-scale in O'Brien's (2010) study of e-shopping. However, these items have consistently loaded on one factor across e-shopping (O'Brien, 2010), social networking (Banhawi & Mohamed, 2011), and wikiSearch applications. This finding has motivated us to consider revising the scale in order to enhance its validity. However, we must determine whether the underlying issue is the factor structure of the UES, or the ability of the Novelty, Felt Involvement, and Endurability items to adequately represent these constructs (Reise, Waller & Comfrey, 2000).

Firstly, the items that comprise the NO, FI and EN sub-scales focus on users' evaluation of their experience (e.g., Felt Involvement: "fun"; Endurability: "rewarding") with the application or how they felt about their task (e.g., Novelty: "I felt interested…," Felt Involvement: "I felt involved…"). It is possible that these assessments, distinct from their impressions of the application's usability (PUs) or aesthetic appeal (AE), or their ability to reach a state of flow

(FA), are equated with the value and success they ascribe to the interaction. Thus, novelty, felt involvement and endurability may not be distinct dimensions of the experience, but form one dimension that characterizes their felt engagement in the task and with the experience as a whole.

A second possibility is the nature of the NO, FI and EN items. Previous research has suggested that each of these dimensions encourages user engagement. Endurability, which represents users' evaluation of system success (DeLone & McLean, 2003) and the likelihood they will use an application in future (Webster & Ahuja, 2006), influenced users' attitudes toward re-engaging with technologies they felt involved with or that provided something new (O'Brien & Toms, 2008). Other studies have demonstrated that felt involvement and novelty are powerful predictors of engagement. For example, in a longitudinal laboratory experiment that tested different interfaces of a digital newspaper, Toms (2000) found that participants read more novel items regardless of the web interface. More recently, O'Brien (2011a) found that novelty and felt involvement played a strong role in maintaining user engagement with online news content, whereby newsreaders were prone to disengage from news stories that did not offer anything new on a topic or that they could not relate to on a personal or societal level, i.e., become interested or invested in. Thus, the issue may not be the inclusion of novelty, felt involvement, and endurability as facets of user experience, but the ability of the UES items to represent them as distinct constructs. UES items mostly pertain to users' perceptions of the system and their experience using it to perform a task, be it shopping, searching, etc. Yet few items pertain to the content of the system, and this may be a severe shortcoming, since both novelty and felt involvement are important characteristics of interest, an important criterion for evaluating information interactions (c.f., Ruthven, 2008).

5.2. Relationships amongst the factors

Due to the four-factor model that emerged, we were unable to fit the wikiSearch data to the path model (Fig. 1) developed in the e-shopping context (O'Brien & Toms, 2010a). However, we examined the correlations amongst the factors and used multiple regression analysis to understand more about how the factors of the UES are related. We observed significant, positive correlations amongst most of the factors, which is consistent with previous research (O'Brien & Toms, 2010a). However, the negative relationship between Focused Attention and Perceived Usability warrants further discussion.

The negative correlation between FA and PUs is inconsistent with our research in the e-shopping domain, but similar to results obtained in the webcast study (O'Brien & Toms, 2010b). Focused Attention is a quality of flow theory (Csikszentmihalyi, 1990), characterized by temporal dissociation and complete absorption in an activity. However, perceived control, challenge (Webster, Trevino & Ryan, 1993) and interactivity (Finneran & Zhang, 2003) are also aspects of flow that have been shown to relate positively to usability. Thus the negative association between these two factors is surprising. One explanation for this may be situational: both the wikiSearch and webcast studies were conducted in laboratories with researcher assigned tasks. The nature of the setting and the tasks may have made a state of flow difficult to achieve. Indeed, in the current study, the average PUs rating (M=5.6) was significantly higher than that of the FA sub-scale (M=3). In naturalistic environments that are not constrained by time or externally motivated tasks, such as e-shopping or exploratory searching, flow and usability may act in a more complementary manner.

Despite the negative association between Perceived Usability and Focused Attention, both play an important role in overall experience. Multiple regression analysis, using Factor 2 (Novelty/Felt Involvement/Endurability) as the criterion variable) showed that Aesthetics, Perceived Usability, and Focused Attention all contributed to the criterion variable, and that the model was not as strong when one of these predictor variables was removed. As a result, it is important not to discount the contribution of both usability and focused attention to user experience.

5.3. Implications for scale validation and revision

The administration of the UES across different applications has resulted in three relatively stable factors: Perceived Usability, Aesthetics and Focused Attention, and three sub-scales (Novelty, Felt Involvement, Endurability) merging to form one factor. Our ultimate goal is to produce a reliable, valid psychometric scale that can be used to gauge the level of engagement users feel when interacting with an information system. Given the findings of this and other studies, we see the way forward as a combination of scale revision and delineation of *what* is being measured, i.e., validation. In this section, we discuss these objectives at the item, dimension, and scale levels.

5.3.1. UES Items

The number of items that retained reliability and factor analyses varied across different applications of the UES. This may reflect the quality of the items or their saliency in a particular context. In the current study, two FA items, “During this search experience, I let myself go,” and “When I was searching, I lost track of the world around me” were eliminated during data screening. One possibility is that the expressions “let myself go” and “lost track…” are culturally nuanced and did not resonate with all participants. However, similar expressions were present in FA items that were not eliminated, such as “I was so involved in my search task that I lost track of time” and “I lost myself in the search experience.” Thus, another reason why these items may have been eliminated is that they did not suit the context. The average FA score was low (M=3, SD 0.83) compared to the means of other sub-scales, especially PUs (M=5.61, SD=0.75), which may suggest that participants were more focused on the functionality of the system or completing the assigned task during the study, rather than on achieving a “flow” state. In addition, one NO item, “I continued to use WikiSearch out of curiosity,” was removed during the reliability analysis. In this case, participants may have continued to use the search system because they had no alternative for finishing the task. Such items may be more appropriate for exploratory search studies conducted in the field where there are no parameters around the system being used or task being executed.

Hence, at the item level, we need to investigate the generalizability of individual items to different research contexts, specifically whether the wording of items reflect Western ways of describing an experience and the applicability of items to both laboratory and more naturalistic settings. However, since different items have been eliminated in each administration of the UES, it is possible that some items fit the application and/or the user group in a particular setting better than others. For example, while the Novelty item, "I continued to use wikiSearch out of curiosity," was eliminated in this study it was included in the Novelty factor in the webcast study (O'Brien & Toms, 2010b), potentially because the wikiSearch interface had functional tools to assist with search whereas the Webcast interface had tools that juxtaposed its information objects in different and perhaps unusual ways. Therefore, we recommend continuing to use all 31 items and using statistical techniques to determine the items that are most salient to user engagement in each circumstance.

5.3.2. Dimension Level

The factor structure and regression analysis provide insight into the UES at the dimension level. Firstly, the results of the factor analysis in this and other studies indicate that NO, EN and FI require further investigation. Specifically, we must examine the UES items that represent these constructs in order to ascertain whether they truly capture each dimension, or if the quality of user experiences is best captured in four factors. In other words, should these be distinct constructs of engagement, or do they belong together? We speculate that content will play an important role in how users think about novelty and involvement with respect to task (O'Brien, 2011a), and that the relationship between content and engagement should be explored in greater depth and be better reflected in the UES.

Secondly, the relationship between PUs and FA has been positive in some studies, and negative in others, including the current work. Prior research would suggest that a system must be usable in order to support absorption and engagement (Huang, 2003; Finneran & Zhang, 2003; O'Brien & Toms, 2008; Webster, Trevino & Ryan, 1993). However, the relationship between these two dimension may be predicated on the needs of the user: searchers who have pragmatic goals may not be interested in being in a state of flow during their interactions, while those with more hedonic needs – or longer time periods to spend engaged in an interaction - may report a stronger link between these constructs.

In addition, although the PUs items consistently factored together across studies, there have been instances where the affective and cognitive items loaded separately (Banhawi & Mohamed, 2011; O'Brien & Toms, 2010b). Originally, we distinguished affect as a distinct attribute of user engagement (O'Brien & Toms, 2008), but during the process of scale development and analysis, these items became integrated into other sub-scales. It may be worth revisiting the possibility of an affective sub-scale, or using an established affective scale with the UES. Another possibility is that affective and cognitive aspects of experience will be more relevant depending upon the application. For example, e-learning environments may be more cognitively demanding whereas social networking sites may elicit more emotional responses. Uncertainty plays a major role in exploratory search, where information needs shift over the course of interaction as searchers articulate their goals, acquire knowledge, and reformulate their search trajectories (White & Roth, 2009). Thus, affect and cognition may also fluctuate, as users move from uncertain to "eureka" moments: engagement may vary over the course of the experience and negative engagement may be an important part of the process.

Lastly, we considered additional attributes of engagement in earlier work (O'Brien & Toms, 2008) that did not emerge as distinct factors in the e-shopping environment (O'Brien & Toms, 2010a), but may be of value in IIR. One example is interactivity, which Huang (2003) deemed "key to creating experiential flow…and is the most important attribute for demanding hedonic Web performance" (p. 433). Interactivity, however is also a multi-dimensional element that may be construed as the way in which an experience is achieved. As IIR interfaces become increasingly complex, incorporating additional tools, multimedia and social media features, it will be imperative to understand the relationship between interactive features and user engagement with those features. Another variable to consider is user motivation, which has been shown to predict some attributes of engagement (O'Brien, 2010), and may be tied to information needs, and may as indicated earlier, impact flow. In order to development a more holistic measure of IIR engagement, the UES may need to be extended or augmented with additional measures.

5.3.3. Scale level

As part of the refinement process, we also feel that it is important to look more closely at what is being measured by the UES, especially in the context of IIR and exploratory search. Hassenzahl's (2011) framework of user experience includes hedonic and pragmatic aspects and system and user-specific variables that coalesce to form an evaluation of an experience. The UES includes items that relate to how the user perceived the system (e.g., PUs and AE sub-scales), their state of mind during system use (e.g., FA sub-scale), and their overall evaluation of the experience (e.g., EN sub-scale); in addition, the items are a blend of the pragmatic (e.g., "I found this system confusing to use" [PUs]) and hedonic (e.g., This experience was fun [FI]). Therefore, the UES is a holistic instrument for assessing user experience.

However, some of the UES items relate not only to the system and the user, but to the task that is being accomplished as part of the interaction, e.g., "I felt involved in the search tasks." Previous user engagement research has shown that users' motivation and interest is intertwined with task, whether it is tangible (e.g., purchasing a book), or intangible (e.g., playing a computer game for enjoyment and escape) (O'Brien & Toms, 2008). In addition, task influences users' system preferences. Jacques, Carey and Preece (1995) examined participants' use of educational multimedia and found that visually based media, such as animations, photographs, and videos, were more engaging for browsing tasks, whereas text-based media were favoured for search tasks. In IIR, searching and browsing are embedded in larger activities that shape users' expectations and use of systems (Ruthven, 2008), suggesting that the relationship between engagement and task should be delineated further.

In addition to examining engagement at the task (versus user and system) level, the role of content in engagement must be more clearly articulated. Recent research has shown that online news users may associate a dimension such as novelty with content (i.e., "Show me new information about a story I have been following"), whereas others may be more interested in novel features of the system (i.e., "Show me different ways in which to interact with news") (O'Brien, 2011a). In information retrieval, content is paramount, with the focus on the quality and relevance of the retrieved results. The exploratory search process is characterized by periods of exploration and uncertainty, knowledge acquisition and insight, focused searching and the synthesis of resources (White & Roth, 2009). The success of an exploratory search is defined by its ability to "clarify vague information needs, learn from exposure to information in document collections, and investigate solutions to information problems" (White & Roth, 2009, p. 3). Thus, content is fundamental to traditional and exploratory interactive information retrieval, and must be investigated more thoroughly with user engagement: Does engagement fluctuate depending on the users' evolving understanding of the content they are seeking, or the availability and presentation of that content from the information system? How do users' judgements of relevance, interest, credibility, etc. equate with their perceived engagement?

As a result, we need to look at the relationship between user, system, content, and task in order to enhance the validity of the UES. To improve the criterion validity of the scale, we could develop additional items to ensure that system, user, task and content are represented by the scale and examine the relationship between these sets of items. In addition, we must look at the UES in conjunction with other measures in order to explore its concurrent validity. Although questionnaires play an essential role in IIR research, they are susceptible to self-report biases. Such issues may be difficult to detect or prevent, and pose challenges to researchers trying to collect a "true" picture of experience (Kelly, Harper & Landau, 2008). Thus, triangulated approaches to measurement are imperative. In IIR in general, and exploratory search more specifically, these measures may be behavioural, physiological, or subjective. With regard to the former, recent research has adopted metrics such as dwell time, number of page views, number of distinct and returning users, time spent interacting with a website over single or multiple sessions, etc. (Lehmann, Lalmas, Yom-Tov, Dupret, 2012; Singla & White, 2010) to explore user engagement. Physiological metrics (e.g., electrodermal activity, heart rate) are an emerging area of evaluation in information science, and have been used in human-computer interaction research to explore affective responses to systems (Mahlke & Minge, 2008). Lastly, subjective measures, such as those traditionally used to assess the relevance of retrieved results (c.f., Ruthven, 2008), would illuminate the relationship between felt experience and content, while cognitive variables (e.g., cognitive style, perceptual speed) (c.f., Al-Maskari & Sanderson, 2011) would provide further insight into the role of individual differences in user engagement.

5.4. Limitations

This study was situated in a laboratory setting, where participants interacted with the wikiSearch system to accomplish researcher-generated tasks. Thus, the general set-up of the study may be a limitation. To evoke engagement in more controlled settings, we may need to expend more effort developing the simulated task scenario (Borlund, 2000), or provide users with more choice in and control over their search tasks, since motivation and intrinsic interest are important qualities of engagement (Jacques, Carey & Preece, 1995; O'Brien & Toms, 2008) and IIR (Ruthven, 2008). We may also consider more naturalistic environments and longitudinal designs for evaluating search engagement (Kelly, 2009; White & Roth, 2009).

The majority of participants were familiar with search engines and Wikipedia, However, the wikiSearch system introduced new features for retrieving and managing results. Participants were given a tutorial before beginning their experimental tasks, and previous studies have not uncovered any issues with the use of this novel interface (see Toms, McCay-Peet and Mackenzie, 2009). Thus we do not believe that the interface especially contributed to the findings. Indeed, PUs subscale items, was on average higher than the mid-point.

Unlike in other studies, the UES was completed post-session after participants completed the tutorial, demographic, pre- and post-task questionnaires, and three search tasks. They may have been experiencing fatigue at this point in the study, which may have led to satisficing, whereby they may have focused on one aspect of the experience rather than the whole experience as instructed (Kelly, Harper & Landau, 2008).

Lastly, the questionnaire items were modified from the initial one used in the e-shopping study. Did the subtle (and what we believed to be accurate) changes lead to a different interpretation of the item? See the following pairs:

Example 1
e-shopping: "Shopping on this website was worthwhile."
wikiSearch: "Searching using wikiSearch was worthwhile."

Example 2
e-shopping: "I blocked out things around me when I was shopping on this website."
wikiSearch: "I blocked out things around me when I was using wikiSearch."

In the these examples, e-shoppers may have focused on their shopping experience – the interaction of shopping task with the technology – at an online bookstore; the wikiSearch participants may have only focused on the technology when answering what we perceived as the same item. The challenging aspect of adapting an instrument constructed for one purpose is in making valid and reliable modifications. At the same time it highlights the tight integration of task, technology and content.

6. Conclusion

In the current study, we administered the User Engagement Scale (UES) to users of an interactive search system and contrasted these results with previous administrations of the Scale in e-shopping, webcast, and social networking environments. To date, the context in which the UES has been administered has varied in several important ways, including the setting (laboratory versus online), sample (university pool versus general public), task (assigned versus self-generated), and time lapse between task completion and responding to the survey (immediately versus within six months). According to Serenko and Turel (2007), it is "impossible to find measures that do not vary over time and across contexts" (p. 657). This is true of the UES, as we have seen differences in the number of items retained across contexts and in the factor structure. However, three sub-scales (Perceived Usability, Focused Attention, and Aesthetic Appeal) have demonstrated stability across several studies.

The configuration of items from the Novelty, Felt Involvement, and Endurability sub-scales have been less straightforward, and there is a need to review the items that make up these sub-scales to ensure they adequately reflect the constructs they represent, or if the UES is a four-factor instrument. This is an important next step, as it will determine how researchers who wish to use the UES will calculate UES sub-scale and overall scores.

In this paper, we examined the generalizability of the UES to a new, exploratory search environment. The four-factor model that emerged led us to consider scale revision and further validation activities at the item, dimension, and overall scale levels. We provide recommendations for improving the UES, including investigating the representativeness of items for constructs and with non-Western users, solidifying the dimensions that make up user engagement, delineating the relationship between user, system, task, and content aspects of user experience; and examining the relationship between the UES and other measures. Scale development and evaluation is a longitudinal process, and only by testing the UES in different environments and under different circumstances can we hone its psychometric properties and produce a reliable, valid and generalizable instrument for understanding users' experiences with information systems.

The complexity of search requires more holistic metrics to assess the users' experience. In this regard, the UES is a useful tool for gauging the pragmatic and hedonic facets of exploratory search. Although we are recommending improvements to the UES, it is useful in its current form for evaluating users' level of engagement along a number of user and system dimensions.

Acknowledgements

Research supported by Social Sciences and Humanities Research Council (SSHRC) grant to O'Brien and GRAND Network of Centres of Excellence grants to O'Brien and Toms, under the nGAIA project. The wikiSearch experiment was conducted at the iLab, Dalhousie University, Canada and funded by grants to Toms: Canada Research Chairs, Canada Foundation for Innovation, and the National Science and Engineering Research Council, and also from the Social Sciences and Humanities Research Council Canada to M. Haufner & B. Detlor (McMaster University) and E. Toms, and V. Trifts (Dalhousie University). The funding sources had no role in the design, conduct or analysis of this work. The authors acknowledge the work of Chris Jordan, Sam Hall, and Tayze Mackenzie who built the wikiSearch system and re-created WiIRE, and of Lab Manager, Lori McCay-Peet, and research assistants, Jason Smith, Alexandra MacNutt, Emilie Dawe, Samantha Dutka, Jennifer Weldon, and Sarah Gilbert.

References

Aladwani, A. M., & Prashant, C. P. (2002). Developing and validating an instrument for measuring user-perceived web quality. *Information and Management, 39*, 467-476.

Al-Maskari, A. & Sanderson, M. (2011). The effect of user characteristics on search effectiveness in information retrieval. *Information Processing and Management, 47*, 719-729.

Al-Maskari, A., & Sanderson, M. (2010). A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology, 61*(5), 859–868.

Banhawi, F. & Mohamad Ali, N. (2011). Measuring User Engagement Attributes in Social Networking Application. In *Proceedings of the 2011 International Conference on Semantic Technology and Information Retrieval, 28-29 June 2011, Putrajaya, Malaysia*, 297-301.

Borlund, P. (2000). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research, 8*(3), Available, http://informationr.net/ir/8-3/paper152.html.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper & Row.

DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean Model of Information Systems Success: A ten-year update. *Information Systems Research, 19*(4), 9-30.

DeVellis, R. F. (2003). *Scale development: Theory and applications, 26 (2nd ed.).* Thousand Oaks, CA: Sage.

Dillon, A. & Vaughan, M. (1997). It's the journey and the destination: Shape and the emergent property of genre in digital documents. *New Review of Multimedia and Hypermedia*, 3, 91-106.

Durrant, G. (2005). Imputation methods for handling item-nonresponse in the social sciences: A methodological review. ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute (S3RI), University of Southampton (37 pp.). Retrieved July 11, 2012, from http://eprints.ncrm.ac.uk/86/1/MethodsReviewPaperNCRM-002.pdf.

Finneran, C. M., & Zhang, P. (2003). A person-artefact-task (PAT) model of flow antecedents in computer-mediated environments. *International Journal of Human-Computer Studies, 59*, 475-496.

Hassenzahl, M. (2011). User experience and experience design. In M. Soegaard & R.F. Dam (Eds.). *Encyclopedia of human-computer interaction.* Available online at http://www.interaction-design.org/encyclopedia/user_experience_and_experience_design.html

Huang, M. (2003). Designing website attributes to induce experiential encounters. *Computers in Human Behavior, 19*, 425-442.

Hyder, J.A. (2010). *Proposal of a website engagement scale and research model: Analysis of the influence of intra-website comparative behavior*. Doctoral Dissertation, Faculty of Economics, University of Valencia, Valencia, Spain.

ISO DIS 9241-210 (2008). Ergonomics of human system interaction - Part 210: Human-centered design for interactive systems (formerly known as 13407). International Organization for Standardization (ISO), Switzerland.

Jacques, R., Preece, J., & Carey, T. (1995). Engagement as a design concept for multimedia. *Canadian Journal of Educational Communication, 24*(1), 49-59.

Järvelin, K. (2011). Evaluation. In I. Ruthven & D. Kelly (Eds). *Interactive information seeking, behavior and retrieval* (pp. 113-138). London: Facet Publishing.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval, 3*, 1-224.

Kelly, D., Harper, D.J., & Landau, B. (2008). Questionnaire mode effects in interactive information retrieval experiments. *Information Processing and Management, 44*, 122-141.

Laurel, B. (1993). *Computers as theatre*. Reading, MA: Addison-Wesley.

Little, R.J.A. & Rubin, D.B. (1989). The analysis of social science data with missing values. *Sociological Methods and Research, 18*(3), 292-326.

Mahlke, S. & Minge, M. (2008). Consideration of multiple components of emotions in human-technology interaction. In C. Peter & R. Beale (Eds.). *Affect and emotion in human-computer interaction* (pp. 51-62). Berlin: Springer-Verlag.

Matlin, M. W. (1994). *Cognition (3rd ed.).* Orlando, Florida: Harcourt Brace.

McCarthy, J. & Wright, P. (2005). Putting "felt-life" at the centre of human-computer interaction (HCI). *Cognition, Technology & Work 7*, 262-271.

McCarthy, J. & Wright, P. (2004). *Technology as experience*. Cambridge, Mass: MIT Press.

Norman, D. A. (1986). Cognitive engineering. In D. A. Norman & S. W. Draper (Eds.), *User centred system design* (pp. 31-61): Lawrence Erlbaum.

Norman, D. A. (2002). Emotion and design: Attractive things work better. Interactions Magazine, 9 (4), 36-42.

O'Brien, H.L. (2011a). Exploring engagement in online news interaction. In *Proceedings of the Annual Meeting of the American Society of Information Science and Technology, 48*, 1–10, DOI: 10.1002/meet.2011.14504801088.

O'Brien, H.L. (2011b). Weaving the threads of experience into human information interaction (HII): Probing User Experience (UX) for new directions in information behaviour. In A. Spink and J. Heinström (Eds.) *New directions in information behaviour* (pp. 69-92). Emerald Publishing.

O'Brien, H.L. (2010). The influence of hedonic and utilitarian motivations on user engagement: The case of online shopping experiences. *Interacting with Computers: Special Issue on User Experience, 22*(5), 344-352.

O'Brien, H.L. & Toms, E.G. (2010a). The development and evaluation of a survey to measure user engagement in e-commerce environments. *Journal of the American Society for Information Science & Technology, 61*(1), 50-69.

O'Brien, H.L. & Toms, E.G. (2010b). Measuring interactive information retrieval: The case of the User Engagement Scale. In *Proceedings of Information Interaction in Context (IIiX)* (pp. 335-340). ACM Digital Library.

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 & Technology, 59*(6), 938-955.

Peterson, R. A. (2000). *Constructing Effective Questionnaires*. Thousand Oaks, CA: Sage.

Quesenbery, W. (2003). Dimensions of usability. In M. Albers & B. Mazur (Eds.), *Content and complexity: Information design in technical communications* (pp. 81-102). Mahwah, N.J.: Lawrence Erlbaum.

Reise, S.P., Waller, N.G., & Comfrey, A.L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*(3), 287-297.

Ruthven, I. (2008). Interactive information retrieval. *Annual Review of Information Science and Technology (ARIST), 42*, 43-92.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction* (3rd ed.). Reading, MA: Addison-Wesley Publishing.

Serenko, A. & Turel, O. (2007). Are MIS instruments stable? An exploratory consideration of the computer playfulness scale. *Information & Management, 44*(8), 657-665.

Su, L.T. (1992). Evaluation measures for interactive information retrieval. *Information Processing and Management, 28*(4), 503-516.

Sutcliffe, A. (2010). *Designing for user engagement: Aesthetic and attractive user interfaces*. In J. M. Carroll & E. M. Frymoyer (Eds.) *Synthesis lectures on human-centered informatics, 5*. Morgan Claypool.

Toms, E.G. (2000). Understanding and facilitating the browsing of electronic text. *International Journal of Human Computer Studies*, 52(3), 423-452.

Toms, E. G., Freund, L., & Li, C. (2004). WiIRE: the Web interactive information retrieval experimentation system prototype. *Information Processing & Management, 40*, 655-675.

Toms, E.G., McCay-Peet, L., and Mackenzie, R.T. (2009). WikiSearch: From access to use. In *Proceedings of the 13th European Conference on Research and Advanced Technology For Digital Libraries (Corfu, Greece, September 27 - October 02, 2009).* M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas, (Eds.) *Lecture Notes In Computer Science* (pp. 27-38). Springer-Verlag.

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*(3), 661-678.

Webster, J., & Ho, H. (1997). Audience engagement in multimedia presentations. *The DATA BASE for Advances in Information Systems, 28*(2), 63-77.

Webster, J., Trevino, L. K., & Ryan, L. (1993). The dimensionality and correlates of flow in human-computer interactions. *Computers in Human Behavior, 9*, 411-426.

White, R.W. & Roth, R.A. (2009). Exploratory search: Beyond the query-response paradigm. In G. Marchionini (Ed.) *Synthesis lectures on information concepts, retrieval and services, 3* (108 pp). Morgan Claypool.