Translating Theory into Methodological Practice

Heather O'Brien

1 Introduction

Despite a decade of devoted research, user engagement (UE) remains a difficult concept to define. Many of us have used various theories and frameworks to guide our understanding of UE and, while there is certainly overlap and congruence amongst our perspectives, key differences are apparent. In the preceding chapter, varied definitions of user engagement were analysed to articulate several challenges: *clarity*, in terms of the unit of analysis (e.g. user, system, content) and level of interaction (e.g. micro or macro); *scope*, the temporal, contextual, and conceptual boundaries of UE; and the defining attributes and accompanying antecedents and outcomes of UE that give the concept *meaning*. An assessment of existing definitions gave rise to a number of emerging propositions and questions to guide future inquiry, yet also an acknowledgement that a unified definition—given the varied applications, settings, and variables of interest in UE research—is difficult to achieve. At the same time, it is this lack of a shared definition of UE that makes the question of how to measure it so arduous.

There are a variety of methodological approaches in human-computer interaction (HCI) that are utilized in user engagement research. These range from self-report methods, such as questionnaires, interviews, focus groups, and verbal elicitations, to neurophysiological methods, including eye tracking, brain imaging, facial expressions, and muscle movements, and observational methods of user behaviour as measured through embodied and on-screen actions, for example, mouse clicks, navigation patterns, etc. [27]. As we will see in this chapter, many researchers are using and combining various methods and measures in innovative ways to further our understanding of user experience. However, we may be putting the "cart before

H. O'Brien (✉)
iSchool, The University of British Columbia, Vancouver, BC, Canada
e-mail: h.obrien@ubc.ca

H. O'Brien, P. Cairns (eds.), *Why Engagement Matters*,
DOI 10.1007/978-3-319-27446-1_2

the horse," so to speak. Research studies are typically focused on a phenomenon of interest; for example: Is online shopping platform A more engaging than B? Do the behavioural patterns of "engaged" searchers differ from those of the "unengaged"? Seldom do we evaluate the methods and measures themselves in terms of their reliability, validity, and generalizability. The question of whether we are measuring what we *think* we are measuring has serious implications for the conclusions we draw about users' experiences (engaged or unengaged) and what precipitated or deterred their engagement.

The purpose of this chapter is to focus on the measurement of UE with respect to two intersecting and fundamental challenges: (1) How do we operationalize UE, a multidimensional, complex quality of subjective user experience with technology? And (2) How do we evaluate the robustness of UE methods and measures?

The chapter is organized as follows. First, I will elaborate on the aforementioned challenges by exploring what is meant by operationalization and robustness in the context of this chapter. Subsequently, I will draw upon the user engagement scale (UES), a multidimensional experiential rating scale, as a case study for illustrating these challenges. The case study will trace the origins of the UES and its adoption and adaptation in a variety of studies, with emphasis on findings related to its reliability, validity, dimensionality, and generalizability. This chapter will conclude with an assessment of the UES as a measurement tool for UE, but will also touch upon the broader challenges that have inspired this chapter with recommendations that UE researchers conduct concurrent research on methods and measures as they continue to investigate UE, and that we articulate the *what* and *why* behind our measurement practices, as well as the *how*, in a more systematic way. Good research is not only about the effective execution of methods and measures, but fostering a reciprocal relationship between theory and practice.

2 Challenges in the Measurement of User Engagement

The chapter "Theoretical Perspectives on User Engagement" of this book explores definitions of user engagement and demonstrates a mixture of overlap as well as a lack of consensus amongst researchers. As such, when we operationalize user engagement in the form of a measurement instrument, we encounter problems. If engagement comprises affective, cognitive, and behavioural elements, then what are these and are they the same in each situation? If our definition of UE is guided by defining features or attributes, then do we agree that these adequately capture the concept and should form the basis of measurement? Furthermore, research design is informed by and informs theory. In the case of quantitative research, experiments are intended to test or verify theory, while qualitative researchers may employ theory to explain observations of the world or to guide inquiry [15]. Many studies of UE incorporate established theories from other disciplines, yet an actual theory of UE is in its infancy. Thus, we do not yet have a theoretical "anchor" for our methodological practices.

Any discussion of measurement must acknowledge the tension between observable and latent variables. Latent variables are “hidden” and must be inferred, since there are phenomena we cannot directly observe. In the case of UE we are trying to account for subjective experiences, how people felt or thought about an HCI, and we must use tools, such as self-report questionnaires, to measure these latent variables. UE research does include variables that can be directly observed and measured, such as behavioural interaction patterns on a website and physiological data based on bodily responses or facial expressions. While more objective, each of these measures still requires a certain level of interpretation: how do we know that specific patterns of muscle movement, behaviours, or electrodermal activity are indicative of UE and not something else?

Kelly [25], in her discussion of the measurement challenges inherent in interactive information retrieval (IIR), emphasized that the dynamic nature of searching and the influence of different contextual factors can mean that the user experience at any one moment is different from the next. One of the key concepts underlying IIR studies is that of relevance, the measures of which “assume [relevance] is stable, independent, binary” (p. 198). This argument could be extended to user engagement—and indeed many other areas of interest in HCI—where we are interested in not only the outcome of interaction but also the trajectory. In Chap. 1 of this book, I discussed user engagement as both a process (journey) and product (outcome) of an interaction. It is difficult to design measures that address both of these aspects simultaneously or that capture definitive shifts in engagement levels over time. This also makes it extremely difficult to examine test-retest reliability, since the dynamism of UE means that no experience—even with the same person using the same system—may be the same.

Lastly, the study of user engagement occurs in numerous settings that vary in location (field or laboratory) and scale, which refers to the number of participants. Scale may range from a qualitative study with 12 people to a log analysis of millions of users’ Web interactions [27]. We cannot employ the same methods in each of these settings. It is not feasible to gather dependable data about user’s emotions, for example, in a large-scale study, or to calculate effect sizes for small-scale studies; field and laboratory-based research represents trade-offs between internal and external validity, and each study must deal with potential confounds and constraints. Thus, a “one-size-fits-all” methodological approach or measure does not exist for UE.

Given the challenges that we face in the operationalization of user engagement, what then constitutes a useful measure of UE? Ideally, measures of user engagement would be relatively easy to administer and interpret, making replication of research designs and findings possible. Ultimately, however, researchers need to know that the measures they are using are robust, and this involves evaluating a measure in terms of its reliability, whether it produces similar results under similar conditions, and validity, its ability to capture what it is intended to measure. Reliability and validity are the cornerstones of effective measures and can help strengthen and expand theory.

In the following section, robustness will be explored using the specific example of the UES. The UES is a 31-item self-report instrument designed to capture six dimensions of user engagement: perceived usability, aesthetic appeal, focused attention, felt involvement, novelty, and endurability, or the overall evaluation of the experience [34, 35, 40]. This case study is not an argument that the UES is the best or only way to measure UE. Rather the UES is used here, in part, because, as its developer, I have a keen interest in its evaluation for the purposes of improving its composition and administration. In addition, it has seen a fair amount of uptake within the research community over the past 5 years. Analysing its multidisciplinary use provides insights regarding its robustness and utility in different computer-mediated contexts.

3 Operationalizing User Engagement with the User Engagement Scale

The UES is one of a handful of self-report questionnaires developed over the past 30 years to measure user engagement.[1] The UES was created during my doctoral work, and I have continued to examine its effectiveness as an experiential rating scale for measuring UE since that time in a range of settings.

When I began my research in the area of user engagement a decade ago, a common reaction was, "Ah, engagement. That's very interesting, but how are you going to measure it?" This was a concern I shared. UE is, after all, not easy to delineate in terms of the user, system, and contextual elements at play; the multidisciplinary literature in which I engaged—from marketing to hypertext systems to games and educational technologies—highlighted different considerations for and attributes of engagement. I unified these varied literatures and examined different theoretical perspectives and their characteristics in an attempt to anchor the concept, and to identify what engagement is and is not. With a theoretical basis in place, I attempted to operationalize UE in the form of a self-report instrument on the basis of its attributes.

I built on the substantial work of Webster [53] and Jacques [23] and their colleagues, who had both constructed self-report instruments of UE in the domain of multimedia education technologies. Both had been developed in the mid-1990s and were domain specific. I cast my net wider than had Webster and Jacques, articulating a range of attributes and attempting to build and test a rating scale using a combination of existing instruments and interview data. I intended for theory to inform the selection of attributes, and that the attributes would shape the content of the self-report measure; the evaluation of the resulting measure would then feed back into the definition and theoretical framework of UE.

[1] For description of some of these and other self-report questionnaires, please see "Chapter 2: Approaches Based on Self-Report Methods" in [27].

The scale development and evaluation process were guided by the literature in terms of the steps and statistical practices employed [18, 45].[2] Peterson [45], for example, outlines several distinct steps, including (1) reviewing the information requirements necessitating a questionnaire; (2) developing and prioritizing potential questions; (3) assessing potential questions in terms of the types of questions to be asked and how these will be worded; (4) determining the structure of the questionnaire, for example the number of categories and their labels; and, finally, (5) evaluating the design and usefulness of the resulting scale.

My process mirrored these steps. First, I generated a list of over 400 potential items derived from the theoretical and applied research literature and a qualitative study. I and another independent researcher assessed this list for the purposes of screening and prioritizing the potential items. These activities led in steps three and four (above), the outcome of which was further screening and pretesting. For step five, the UES was employed in two large-scale studies in the online shopping environment. These two studies allowed me to make the UES more parsimonious, explore its factor structure, and examine the reliability and validity of these resulting factors. The result was a 31-item instrument with six distinct sub-scales, which were arrived at through factor analysis, interpreted, and labelled as perceived usability, focused attention, felt involvement, novelty, aesthetic appeal, and endurability. The UES was published in its entirety in two academic papers [35, 40] and my doctoral dissertation [34].

Subsequently, I have continued to evaluate the UES in information search [41, 42] and online news [36–38]. In addition, the UES has been adopted and, in many cases, adapted by other researchers for use in different settings. Based on data provided by Google Scholar, publications of the UES have been cited 84 [35], 180 [40], and 16 [34] times as of the writing of this chapter.

For the purposes of looking here at how the work has been drawn upon, I specifically isolated the 180 works citing [40]. These comprised journal articles, conference proceedings, book chapters, Master's theses, and Doctoral dissertations published in 2010–2015. These works were first surveyed to determine whether or not the UES was used and, if so, whether the nature of the use was clear. In screening the articles, I also took into account works by the same author or group of authors and whether they were reporting on the same study across multiple papers. If this was the case, then only one paper was retained for this review to avoid redundancy. Lastly, there were some works that were not written in English that I was unable to read. After examining all works according to relevance, overlap, and language, there were approximately 44 remaining works; these were thoroughly read and annotated.

The papers represented a broad range of applications and studies in the areas of information search, online news, online video, educational applications, haptic interfaces, consumer applications, social networking systems, and video games. Overall, there were three categories of implementation of the UES: (1) use of

[2] For an in-depth description of the UES development and evaluation process, please see [34] or [40].

individual items or a combination of items unrelated to the original sub-scales; (2) use of specific sub-scales of the UES; and (3) use of the UES in its entirety. In the following sections, I will investigate the various implementations of the UES in greater detail, with particular emphasis on the instrument's reliability, validity, dimensionality, and generalizability.

3.1 Evaluation of the User Engagement Scale

The evaluation of an experiential scale involves assessing its usefulness through four distinct procedures: dimensionality, reliability, validity, and generalizability [45, p. 81]. *Dimensionality* refers to the number of underlying constructs being measured with the scale. This is typically assessed using principal components analysis (PCA) or factor analysis (FA), which are "statistical techniques applied to a single set of variables when the researcher is interested in discovering which variables in the set form coherent subsets that are relatively independent of one another" [50, p. 612]. The UES is a multidimensional tool and, as such, should produce a six-factor structure, according to its original configuration [34, 40].

The *reliability* of a measurement scale may be assessed along two lines: internal consistency and longitudinal stability. Internal consistency refers to how well scale items are measuring the same construct and can be assessed statistically using techniques such as Cronbach's alpha. Devellis [18] suggests that the ideal range for Cronbach's alpha is 0.7–0.9; this range represents good reliability without redundancy. Longitudinal stability, or test-retest reliability, examines the results of administering the instrument over time to the same participants and comparing the responses [45].

There are many forms of *validity*, and this evaluation will look specifically at criterion-related validity and construct validity [18]. Criterion-related validity is demonstrated when measures that should be associated with each other are, indeed, related. Construct validity is "the extent to which a measure 'behaves' the way that the construct it purports to measure should behave with regard to established measures of other constructs" [18]. In this chapter, validity is operationalized according to how well the UES can help differentiate between different experimental conditions or systems and its relationship to established measures, including other self-report scales.

Lastly, *generalizability* refers to "the administrative viability [of the scale] and interpretation in different research situations" [45, pp. 79–80]. To examine the generalizability of the UES, I will provide an overview of its success within the variety of technology domains represented in the reviewed articles.

The case study will conclude with an overall assessment of the UES according to its use and usefulness in measuring engagement in different studies and an examination of the limitations of both the tool and its administration.

3.1.1 Dimensionality

The original UES purported that user engagement comprises six dimensions: perceived usability, aesthetic appeal, focused attention, novelty, felt involvement, and endurability. These six dimensions emerged from exploratory factor analysis (EFA) in the first online study of online shoppers; structural equation modelling (SEM) was used in the second study to confirm the factor structure and test a hypothesized path model of the relationship amongst these factors [34, 40]. Subsequently, only a small number of studies have examined the dimensionality of the UES.

Banhawi and Ali [7] surveyed over 100 Facebook users to examine the generalizability of the UES in a social networking system (SNS) setting. Using EFA, they determined that there were four distinct UES factors: focused attention, perceived usability, aesthetics, and novelty-endurability; furthermore, the number of items retained from the original 31-item scales was 28. Wiebe et al. [55] arrived at a similar outcome in their exploration of the UES in gaming environments. Using principle axis factoring (PAF) with promax rotation, the researchers reported that the "UESz" consisted of four factors: focused attention, which included focused attention items and one felt involvement item; perceived usability, which included the perceived usability items and one endurability item; aesthetics; and "satisfaction", which was comprised of items from the novelty, endurability, and felt involvement sub-scales. 28 items were retained in the analysis.

In my own work, I have also observed this four-factor structure with one exception. In studies of exploratory search [42] and online news [37], perceived usability, aesthetic appeal, and focused attention have been "stable" sub-scales, while endurability, felt involvement, and novelty items have combined to form one factor or component, which we have labelled as "hedonic engagement". When we looked at the internal consistency of these four factors post-FA or PCS, we find support for their reliability. The exception to this four-factor finding was a study of educational webcast users [41]. This study found a six-factor UES, but one that differed from the original. The aesthetic appeal, focused attention, novelty, and endurability sub-scales were retained through factor analysis, though some items were eliminated from each sub-scale. The felt involvement sub-scale was eliminated and the perceived usability sub-scale loaded on two factors: one that contained affective items ("frustrated", "discouraged") and the other comprised cognitive items ("taxing").

This handful of studies that have both used the UES in its entirety and examined its factor structure strongly suggest that the instrument is comprised of four dimensions rather than six, but more studies are needed to confirm this finding. All of the aforementioned studies indicated good reliability of post-factor analysis sub-scales, but only two have commented on the relationship between factors. Wiebe et al. [55] conducted regression analysis with the maximum difficulty achieved by players in the game as the outcome variable and the flow state scale (FSS) and the newly derived UESz as predictors. FSS and UESz explained more of the variance in player performance than either scale individually (though the variance explained was still quite low). They noted that all of the post-PAF sub-scales were significantly correlated with each other but that there appeared to be "hedonic/utilitarian divide in the sub-scales" (p. 130), with perceived usability being least correlated with the other factors.

In our study of exploratory search [42], we noted a non-significant correlation between focused attention and perceived usability and a significant but low correlation between focused attention and aesthetic appeal. Thus our divide seemed to be not utilitarian/hedonic but system/user. Nonetheless, when we used multiple regression with focused attention, perceived usability, and aesthetic appeal as predictors and the combined endurability/novelty/felt involvement factor as the criterion variable, we found that the strongest model contained all three predictors, which accounted for 55 % of the variance. One way to interpret this is that these empirical findings are in line with the Process Model of User Engagement [34, 39] and that the attributes (or factors) of engagement are varying in intensity, depending on how salient they are in a given context.

The only other study that has utilized SEM in the study of UE is that of Seedorf et al. [47]. They conducted an Amazon Mechanical Turk study where MTurk workers interacted with a shopping website alone or with another remote person; in the collaborative conditions, participants could either communicate via a text chat window or could co-browse the site, view each others' mouse movements, and communicate via the chat feature. Social presence and collaboration usefulness were measured for the participants shopping in pairs, and all MTurkers completed 14 items from the focused attention, endurability, novelty and felt involvement subscales, which had good reliability.

Two different path models were examined for those searching alone (control group) and the participants browsing in the social conditions. For the control group, novelty increased focused attention and involvement, focused attention increased felt involvement, and involvement increased endurability; this model reproduced the relationships we identified between engagement factors in our original work [40]. The path model was similar for the paired searchers, except that there was no relationship between focused attention and felt involvement in this model. However, the path model also took into account new variables, namely social presence and collaboration, which increased the overall explained variance of the basic path model. Social presence increased both endurability and felt involvement, and collaboration usefulness led to higher levels of novelty and social presence. When the two paired conditions were compared, social presence, induced through co-browsing, led to stronger ratings of the endurability of interacting with the shopping website. Although this study did not indicate how they selected the 14 UES items, it replicated and enhanced elements of the original SEM results, confirming focused attention, felt involvement, novelty, and endurability as distinct dimensions and their interrelationships.

These findings are interesting in light of the four-factor structure that emerges in other studies where felt involvement, novelty, and endurability are combining into one factor. One explanation is that EFA, a data reduction technique, is best suited for scale development when the underlying latent constructs of a scale are unknown, whereas confirmatory analysis is more appropriate for testing an existing measurement model with another data set.

3.1.2 Reliability

Reliability, in terms of internal consistency, is related to the above discussion about dimensionality. Studies that have used PCA or FA to examine the factor or component structure of the UES have also examined the reliability of the sub-scales before PCA/FA and the resulting factors/components. Others have calculated Cronbach's alpha even though PCA and FA were not part of their analysis. For example, Arguello et al. [5] used 11 items from the focused attention, felt involvement, perceived usability, and endurability sub-scales and found good reliability with Cronbach's alpha values of 0.714–0.94. In some cases, items were removed during data screening to improve internal consistency, either due to redundancy amongst items or when an item did not correlate well with other items of the same sub-scale [37, 42].

Fewer studies have examined the test-retest reliability of the UES, though there are some noteworthy studies. Bustillo and Garaizar [12] examined the engagement of student teachers with Scratch, an application that combines programming and online community to teach computational thinking and digital literacy. They tested perceived engagement with Scratch in two different academic years and reported no differences; they anticipated higher engagement in year two. However, a focus group conducted with a subset of participants emphasized contextual factors, such as participants' inexperience with programming and limited programming training practices in schools, that may have affected the findings.

Another study in the education domain was conducted by Vail et al. [51], who were interested in whether male and female students might respond to different types of support: cognitive (e.g. problem solving) or affective (e.g. motivation, self-confidence) support offered by intelligent tutoring systems. The researchers tested four versions of a text-based adventure game; the baseline system scaffolded introductory computer programming tasks, and the other three versions were augmented with affective, cognitive, or cognitive-affective support. In all versions, the students completed the learning tasks in five separate but iterative sessions. Some students interacted with a human tutor (human–human), while others received support via the intelligent tutoring system (human–ITS); participants in both groups achieved significant learning gains, according to pre- and post-test scores for each tutorial session. Students rated their level of engagement according to the sum of three UES sub-scales, focused attention, felt involvement, and endurability, and their frustration using the NASA-TLX workload survey after each tutorial session; average engagement, frustration, and learning gain scores were computed for each student across the five sessions [51]. In addition to insightful findings regarding gender differences in preferences for affective or cognitive feedback and human or computer feedback, the authors took consistent measures over five sessions with the same students and examined the relationship between self-reported engagement, frustration, and learning gains. Males and females made similar learning gains, but females preferred affective, system feedback.

Lastly, Bangcuyo et al. [6] compared the user experience of sampling coffee in a traditional sensory environment with a "virtual coffee house" featuring visual, auditory, and olfactory cues; they were interested in contrasting user preferences in these settings and the stability of these preferences over time. Participants rated samples of four different coffees brewed at different strengths on a 9-point hedonic scale and completed an engagement questionnaire (21 items derived from the UES and Witmer and Singer's Presence Questionnaire [56]); total engagement scores were calculated for the virtual and traditional coffee tasting environments. The experiment was repeated twice, with a 1-month lapse between the first and second replications. In the first part of the experiment, there were significant differences in engagement between the traditional and virtual tasting environments, and these findings were stable 1 month later. Based on this finding, the authors concluded that the virtual coffee house remained hedonically appealing to participants over the two trials and that the engagement questionnaire used in the study, of which the UES was a part, was "both a reliable and valid testing instrument" (p. 93).

3.1.3 Validity

Validity, here operationalized according to how the UES functions in relation to other constructs, will also be discussed in the following section on generalizability that looks at the scale's performance in each domain more broadly. This section targets specific examples that support or refute the validity of the instrument.

The UES has been shown to correlate with other self-report instruments, including the FSS [55], system usability scale (SUS) [10], and cognitive absorption scale (CAS) [1, 38]. Flow Theory has informed UE research (see chapter "Theoretical Perspectives on User Engagement"), and the SUS and CAS are akin to the perceived usability and focused attention/felt involvement dimensions of the UES; thus, these measures should and have found to be correlated. Other studies have used elements of the UES and NASA-TLX concurrently to examine the interplay of engagement and its logical opposite, frustration [20, 49].

There have been mixed results concerning the relationship between the UES and objective measures. We found no significant relationship between UES scores and browsing behaviours (time, pages visited, and use of recommended links) or physiological measures (heart rate, electrodermal activity, electromyogram) [38]. However, those who rated their engagement as low spent almost twice as long reading during the session as highly engaged participants and visited more links on average (16 compared to 9.5), suggesting some level of disorientation or inability to engage with the task. Although the lack of statistically significant congruence with the objective measures was disappointing, the study questioned the validity of using some behavioural metrics in isolation, since more time on task was indicative of both low and high engagement. Further, Warnock and Lalmas [52] found no relationship between cursor behaviours and the UES, while [3] discovered a negative correlation between cursor movements and focused attention and affect (PANAS). In other words, “negative emotions [were] more influential on cursor behaviour than positive ones” (p. 1447).

Parra and Brusilovsky [44] looked specifically as users’ interactions with two different interfaces of a novel conference navigation system. Subjective metrics included UES items that corresponded to the focused attention, perceived usability, novelty, and endurability, as well as objective information retrieval metrics, such as average rating of users in particular conditions, precision, mean average precision (MAP), mean reciprocal rank, and normalized discounted cumulative gain. Participants interacted with both versions of the conference navigation system: a baseline system and one that included additional features (e.g. sliders and Venn diagrams) to enhance the controllability of the interface and to assist users in locating relevant papers amongst recommended results. The results of a regression analysis showed that the effects of engagement on usage metrics were dependent upon the order in which people interacted with the baseline or experimental interface. The researchers also demonstrated a relationship between UE and MAP and participants’ understandability of the interface in the task performed with the controllable interface; using the controllable interface after the baseline system also resulted in significant positive effects for subjective impressions of the endurability of the system. In addition, specific user characteristics, namely, experience with Conference Navigator and recommender systems, trust propensity, trust in recommender systems, and expertise in the research domain significantly influenced user engagement.

Grafsgaard [20] conducted a series of studies to investigate the relationship between affect and nonverbal behaviours in tutoring systems. Facial action units were recorded as students interacted with JavaTutor and coded across seven tutoring sessions. Significant facial movement patterns were observed between focused attention, involvement and endurability (UES), frustration (NASA-TLX), and learning. Brow lowering intensity, for example, was associated with higher levels of frustration and lower endurability, namely a reluctance to return to future tutoring sessions; average intensity of inner brow raising was associated with students’ rating of the session as worthwhile. Postural shifts were also used to examine disengagement. Specifically, body position during student questioning, tutor responses, and positive feedback from the tutor mapped to higher self-reported engagement. Based on the findings, Grafsgaard [20] devised a predictive model of user engagement whereby students’ initial computer science self-efficacy scores, one-hand-to-face gestures after successful compile (i.e. a programming subtask), and brow lowering after sending a student dialogue message led to higher post-session engagement; more facial movements (and perhaps more intense affective reactions) were associated with lower engagement.

In a portion of the studies, the UES has been shown to distinguish between conditions or experimental systems. We manipulated news source familiarity [37] and news media [36] and detected differences in users’ experience, showing the UES to be sensitive to different study conditions. Moshfeghi et al. [33] developed a news search system with blog and news entries and associated images to determine whether enhanced visual search features would improve user engagement. Log files collected queries, mouse clicks for the distinct components of the user interface, overall time spent interacting with the news system, and time spent reading articles. All MTurk participants (*n* = 63) interacted with both versions (baseline and enriched) in counterbalanced order and were asked about their system preference and level of engagement. Findings indicated a clear preference for the enriched system, and participants found it more engaging than the baseline system on most of the UES dimensions; there was, however, no difference in the perceived usability of the systems.

However, other researchers have found less support for the UES's validity. Kajalainen [24] reported no significant differences in perceived engagement across five experimental conditions that altered the presentation of a satirical news show; the engagement questionnaire was derived from various instruments, including the UES.

Sharek [48] tested three types of game designs with Amazon Mechanical Turk (AMT) workers: static, based on a linear progression of difficulty; user controlled, where users selected a difficulty level at the beginning of the game; and adaptive, which manipulated the level of difficulty algorithmically. MTurk workers described their experience using the UES, the NASA-TLX cognitive load sub-scale, and items pertaining to interest and enjoyment ("personal affect") from the Intrinsic Motivation Inventory based on Self-Determination Theory [17]. He did not find any significant differences for any of the self-report measures across the three gameplay conditions, though there were performance differences. For example, those in the adaptive conditions played fewer levels of the game yet achieved greater difficulty, and those in the linear condition took more time to complete each level and react to the secondary task. Sharek speculated about the lack of significance for the self-report measures, reasoning that MTurkers are paid to participate and may be therefore less intrinsically motivated and involved with the game. In addition, the game seemed to have good entertainment value, regardless of the condition of play. However, Sharek cautioned that reliance on self-reported experience is problematic and "highlights a limitation in the diagnosticity of these cumulative self-report measures and strengthens the case for including real-time measures when possible" [48, p. 85].

Warnock and Lalmas [52] also drew upon MTurkers, asking them to carry out low- and high-interest search tasks (based on pre-task assessments of topical interest) using a "normal" or "ugly" website; the websites contained the same content, but the "ugly" website was made to be unappealing with changes to the colour, font, and presence of ads. Findings most pertinent to the validity of the UES were that no differences were found in participants' aesthetic appeal ratings of the websites. The authors questioned the reliability of the user experience data, since the "ugly" website violated so many aesthetic conventions. A possible explanation is that the usability of the ugly site was not affected by the cosmetic alterations made and that the ability to carry out the tasks proficiently led to less deterrence that expected. In addition, since participants interacted with only one of the websites, they had no basis of comparison in terms of what was "normal" or "ugly". Nonetheless, the fact that aesthetic ratings were not significantly different between the two websites challenges the validity of the aesthetic appeal sub-scale.

In another AMT study, McCay-Peet et al. [32] manipulated the visual catchiness of entertainment news headlines or topics to examine the effects of task-relevant saliency on focused attention, affect, and search performance. Self-report measures included the PANAS, the focused attention sub-scale of the UES, and questions about interest in news items, confidence in search effectiveness, and task difficulty. The focused attention sub-scale was shown to be highly reliable in three pilot studies, but did not detect differences between the salient and non-salient tasks. However, the authors noted that those in the non-salient condition said they were more distracted by the non-task-relevant features of the websites. Thus, the focused attention of both groups may have been similar overall, but they were attending to different aspects of the interface.

In addition to studies that support or refute the validity of the instrument, there are those that suggest more complexity. Arapakis et al. [2], for example, found no significant differences in responses to the UES's focused attention sub-scale across article interestingness levels (as determined by participants' pre-ranking) in their online news study. However, when participants were grouped according to their perceived interest in the articles, the "interesting" group reported significantly higher levels of focused attention than the "uninteresting" group, but these groupings did not affect actual or perceived time spent on the news reading task. Overall, Arapakis et al. [2] found that interesting news content increased positive affect and led to more focused attention and longer fixations on new and popular news item comments, which demonstrated congruency between subjective self-report and objective eye tracking data. These findings revealed differences when situational interest was fostered in the experiment, but not when it was based on preconceived notions of what news articles participants thought would be interesting to read at the outset. In another study of engagement and search performance with aggregated search displays, Arguello et al. [5] demonstrated significant findings for their objective measures of search performance and task complexity. However, user perceptions of search effectiveness and engagement were not significantly different for two experimental interfaces, even though post-session interviews confirmed that the majority of participants did notice a difference between the two interfaces. Yet participants indicated clear preferences for one of the two interfaces, and, when this preference was taken into account, there was congruency between user interface preferences and their user experience ratings, especially for perceived usability and endurability.

A small number of studies have treated UE as a mediating variable, exploring antecedents and outcomes of engagement with companies' Facebook pages [46], advertisements [28], social media [21], and health information seeking [22]. Reitz [46] adopted and modified 11 UES items to measure cognitive and affective aspects of online consumer engagement, and three items from the narrative engagement scale [11] were included to measure presence. SEM was used to examine the relationship between information quality, enjoyment, interactivity, affective/cognitive engagement, behavioural engagement, loyalty, and (re)purchase intent. Results indicated that perceived information quality, enjoyment, and interactivity predicted cognitive/affective engagement and participation; these, in turn, predicted brand loyalty, which led to intention to (re)purchase. In other words, users' content-based and physical interactions with online consumer Facebook pages led to affective, cognitive, and behavioural engagement, which influenced how they thought about and intended to interact with the company in future.

Another study based in the online consumer domain manipulated the perceived authorship of an ad for the Amazon eBook reader, Kindle [28]. Some people were told the ad was created by a communications firm on behalf of Amazon, while others were led to believe the ad was made by a Kindle user, "Angela", who was either motivated by her enthusiasm for the product or the potential to win a 20,000 dollar prize. Self-report items were derived from the UES and other sources. Based on the positive correlations observed between all of the engagement dimension and ad performance, the authors concluded there was a definitive link between engagement and the effectiveness of the ad, particularly when the ad was perceived to be made by a fellow consumer rather than the communications firm.

Halpern [21] studied the relationship between user engagement, cognitive involvement, and collective efficacy, the shared belief held by individuals about the group's capabilities and skills for performing a collective action. This study collected participants' (*n* = 151) comments on the White House and other US federal agencies' Facebook and YouTube accounts over a 2-week period. Pre- and post-task questionnaires were used to look at the three primary variables of interest, as well as participants' demographic characteristics, social media use, and interest in political affairs; user engagement items and cognitive involvement were derived from some of the UES's felt involvement items and Kwak et al.'s [26] work on political engagement. The author demonstrated that social networking sites have the potential to positively affect collective efficacy, particularly when the media enables networked information access that supports the formation of an online public sphere. Further, user engagement and cognitive involvement, along with participants' preference for social media channels, helped to explain increases in cognitive efficacy. UE was related to the types of behaviours participants engaged in: those who participated in more interactive conversations and replied to others' messages were more engaged than those who did not.

Hong [22] included UES items in a multifaceted study of online health information seeking, capturing both click stream and user perception data. Participants (*n* = 106) were randomly assigned to interact with health information in one of four message conditions where motivation orientation (health promotion or prevention) or message frame (health outcome gain or loss) was manipulated. The researcher examined the content selected by participants and the extent of their search, as measured by search session length and number of pages viewed. Participants were also asked about their impressions of task engagement (three focused attention and perceived usability UES items) and message quality (two items addressing stylistic quality and interestingness). Hong [22] found that those in the promotion orientation/gain frame condition were more engaged, and this had a mediating effect on message quality. The relationship between content presentation, engagement, and content assessment has important implications for the design of online content in health and potentially other fields.

3.1.4 Generalizability

This section looks specifically at the generalizability of the UES according to domain areas: online shopping, online news, online video, educational applications, haptic and consumer applications, social media, and video games. The caveat in examining the generalizability of the UES is that few studies use the scale in its entirety, and this makes it difficult to draw definitive conclusions. Therefore, this section focuses more on the fit and success of the UES as it relates to the researchers' goals and outcomes of interest and varied domain-based settings.

Online Search

Studies conducted in the online search domain have used UES items to investigate subject-specific information retrieval and aggregate search systems. These studies have tended to use a selection of UES items [22] or one or more sub-scales [4, 5, 9, 44]. These studies largely support the utility of the UES. Where reliability assessments were conducted, UES sub-scales showed good internal consistency [5] and adequate validity. The UES (or components of it) helped to distinguish user experience when the motivation and message frame of the information seeking interface was were manipulated [22], and differentiated between the parallel and dependent search conditions tested by Bron et al. [9] and the fast and slow search systems introduced by Arapakis et al. [4]. However, they did not show effects of search latency [4] or discriminate search results presentation [5] or two versions of an information retrieval conference systems [44] *unless* user system preference or the order of system use was taken into account.

These latter findings question the sensitivity of the UES, but Arguello et al. [5] offer an alternative view, suggesting that user experience may be more person dependent than system/interface dependent. This idea has some support in Parra and Brusilovsky's [44] findings that certain user characteristics influenced participants' perceived usability of the conference system, at least amongst those who interacted with the baseline system first. These studies do demonstrate some validity for the UES: both Hong [22] and Parra and Brusilovsky [44] found relationships between UES items and performance measures or other self-report variables (e.g. message quality and understandability). In the case of Hong [22], UE was shown to be a mediating variable between message frame and motivation and perceived message quality.

None of the search studies profiled used the UES in its entirety. However, aspects of the scale that were used demonstrate utility in helping researchers explore variables of interest, differentiate experimental conditions or interfaces, and gain an understanding of user search behaviour, user characteristics, and system order and preferences on subjective experiences and search behaviour.

Online News

Many of the studies conducted in the online news domain have focused intensely on the relationship between focused attention, emotion, and user behaviour [2, 3, 32, 52]. Arapakis et al. [2, 3] and McCay-Peet et al. [32] did not find differences in focused attention across different levels of article interestingness, and Warnock and Lalmas [52] found that aesthetic appeal items failed to differentiate an obvious manipulation of a news website's aesthetic conventions. Further, while Warnock and Lalmas [52] did not report a connection between self-reported focused attention and cursor behaviours as measured by mouse clicks, Arapakis et al. [2] successfully linked focused attention and eye movements: self-reported focused attention and eye gaze movements should be and indeed were related in their study. This collection of studies also underscored important connections between interest and user engagement, which we have also observed [38]. They specifically emphasized the relationship between interest, negative affect, and cursor behaviour [3] and interest, emotion, and focused attention [2].

Additional work in the online news domain has used more of the UES subscales beyond focused attention and has looked at user experience in relation to the presentation of news search results [33] or how people might "think" about news content [43]. Moshfeghi et al. [33] were able to distinguish a baseline and enriched news system on every dimension of the UES except perceived usability. Further, they showed that user characteristics, previous search experience, and performance data collected during the study were able to predict UE (with the exception of focused attention). Okoro [43] did not reveal significant differences in user engagement when performing a news selection task in freeform, timeline, or argumentation conditions with a news corpus; however, there were also no performance differences across the three interaction modes, and this may indicate that the manipulation was not successful overall.

Online Video

Online video may be part of online searching and news reading, but several studies have isolated video interaction. Lee et al. [29] and Zhu et al. [57] explored the social dimensions of online video viewing; Kajalainen [24] investigated the effects of different amounts and types of interactivity; and De Moor et al. [16] compared self-report and physiological data in this domain.

Lee et al. [29] observed different UE levels between a baseline (static) and dynamic version of a video system that featured affective and social commentary. Although they did not look at UE, learning, and social interaction in concert, the individual analyses of these pairs of variables suggested that content-related comments provided learning benefits, as well as social interaction and engagement. Similar to findings in online news, De Moor et al. [16] showed a clear relationship between interest and UE, and intuitively that engagement is higher when video viewing is error-free and contributes to perceived video quality. Interestingly, Zhu et al. [57] did not find differences in users' evaluation of their experience when they manipulated perceived usability through bitrate speed, but did find a connection between perceived endurability of the experience, genre (e.g. comedy), and viewing videos with friends.

None of the online video studies used the UES in its entirety. It is also not clear what items were used in Kajalainen's [24] study, and this makes it challenging to determine if the lack of differences between conditions in this study is due to the quality of the UES. In the case of the other studies, UES components utilized by De Moor et al. [16], Zhu et al. [57], and Lee et al. [29] were useful for understanding the video experience, particularly illuminating the role of socialization in engagement.

Educational Applications

Studies that have used the UES in the education realm have been technology centered and in the classroom. Studies reviewed here examine engagement, and its antithesis, frustration, with tutoring systems, as well as applications and workshops designed to be more novel than traditional classroom lessons.

The dichotomous relationship between user engagement and frustration, measured with the established NASA-TLX, provides evidence of the UES's validity in this setting. Two studies [51, 54] focused on an especially salient outcome—learning. Vail et al. [51] found that although students had higher learning gains with human tutors, they found them less engaging; there were also no differences in learning when students interacted with the baseline and enhanced systems, which provided cognitive, affective, or cognitive/affective feedback. However, they observed gender differences in students' responses to the type of feedback and mode of feedback delivery. Since learning was not affected, but engagement was, this implies that tutoring systems can be personalized to the preferences of groups of learners to provide an enjoyable experience without compromising learning. Whitman [54] had similar revelations when learning gains were made for students interacting with a baseline or interactive tutorial, even though the latter was more engaging. In this case, however, the interactive tutorial allowed students to perform well on declarative and procedural knowledge tests and to do so faster than the baseline condition while still enjoying the experience. Thus, while there were no differences in learning outcomes for the static and interactive system, the students performed more efficiently and experienced greater enjoyment with the interactive system. The author did not look at long-term retention of the information gleaned from the tutorials, and this would be an interesting and informative investigation of the longitudinal effects of engagement and interactivity on learning.

In sum, the UES was effective for helping authors in this domain explore UE in different settings and therefore showed utility. However, the findings highlight the complexity of learning environments, where more engagement does not necessarily equate with increased learning and where previous experience in the domain area or contextual factors influences learners' motivations and ability to learn [12]. Grafsgaard's [20] work, highly innovative and robust in its own right, isolated specific facial units gathered during the learning process and related these to self-reported frustration and engagement. The relationship between summative and formative and objective and subjective measures is an exciting finding related to the validity of the UES and self-report methods more generally.

Haptic Applications

Haptic applications, technologies that use vibrations or motion to convey tactile feedback to users, are featured in two of the reviewed studies that employed the UES. Levesque et al. [30] did not use the UES in its entirety, but did draw items from each of the six sub-scales. They showed that haptic versus non-haptic interactions resulted in no performance differences but did impact users' perceptions of the four widgets they tested (alarm clock, text editor, game, and file manager), with higher perceived engagement when the widgets featured friction. Shirzad [49] also used an assortment of UES items, along with the self-assessment manikin (SAM), NASA-TLX, and Godspeed questionnaire (user responsiveness to the robot), to explore performance differences in a robotic reaching task. The UES was related to various dimensions of NASA-TLX, in that the experimental group were less frustrated, exerted less perceived effort, and achieved higher task performance than the control group; the UES was also associated with the SAM, which examined task satisfaction and attentiveness. The coupling of UES, NASA-TLX, and SAM showed good criterion validity for the UES. For this haptic application, which could be employed in a clinical setting as part of rehabilitation therapy, there is a real impetus to increase people's willingness to use it. Therefore, the performance itself may be less relevant to actual and continued use than engagement.

Consumer Engagement

The range of applications of the UES and related measures in consumer research is quite fascinating. In addition to studying how people interact with companies in a social networking or online shopping setting, researchers have focused on company logos, ads, and virtual tasting environments.

In consumer engagement, UE has been explored along with presence [6, 46], a pairing that may not be suitable in all domains. However, UES items and sub-scales have assisted researchers in testing diverse research questions, such as the dynamism conveyed in brand logos [14] and how the creator of an ad (corporate versus fellow consumer) influenced users' perceptions of trust and overall engagement [28]. Along with the source of the information, studies have looked at the perceived quality of the information [46]. Both Reitz [46] and Seedorf et al.

[47] constructed and validated path models featuring engagement. In the case of the former, engagement was a mediating variable: information quality, along with level of interactivity and enjoyment experiences, predicted engagement, which in turn influenced company loyalty and intention to (re)purchase. Seedorf et al. [47] confirmed the path model we originally tested between some of the UES subscales [40] and added to this “social presence” in their study of collaborative online shopping. The inclusion of the UES in these structural equation models is a boost for the scale’s validity, and Bangcuyo et al.’s [6] finding that users’ experiences with the virtual coffee house were stable after a 1-month period supports the longitudinal stability of the UES.

Social Networking Applications

Although some of the studies included in other sections of this chapter feature social networking sites (SNS), the three studies discussed here focus specifically on personal relationships through technology (as opposed to responding to a company through SNS).

Banhawi and Ali [7] tested the factorization of the UES and indicated a four-factor structure for the scale (discussed further in Sect. 3.1.1) and also found that interactions with Facebook using mobile devices were more engaging than those using a computer. Other SNS studies reviewed did not test the entire UES, but did show that self-reported engagement was linked with online behaviours [21] and that UE was one element (along with SNS preference and cognitive involvement) that predicted cognitive efficacy. This is a particularly salient finding given the plethora of research investigating crowdsourcing applications and provides additional insights into work using analytic data to study such phenomenon. Fuchsberger et al.’s [19] finding that user engagement persisted despite poor perceived system usability and lack of computer skills amongst older adults is interesting. In previous research, I made a case that a minimum amount of usability is necessary for engagement to occur [40], but, in this case, the desire to connect socially puts this hypothesis into question and warrants further examination of the trade-off between usability and perceived social gains in predicting UE.

Video Games

Findings in the area of video games are mixed. Neither Choi [13] nor Sharek [48] found significant differences in perceived engagement for participants in their studies across experimental manipulations. However, Choi [13] also did not see hypothesized performance differences for participants training to do maze tasks using other different video games; thus, if all conditions were equally enjoyable or arduous, then there may have been no procedural or perceived engagement differences. None of the self-report measures tested by Sharek [48] were significantly different across the linear, user-controlled, and adaptive gameplay systems tested.

Wiebe et al. [55] lend further evidence to a four-factor UES with their study and did show the reliability of the six UES sub-scales (pre-PAF) and factors (post-PAF). Further, they demonstrated a relationship between the UES and FSS, but their regression analysis with game level as the criterion variable did not include focused attention and resulted in a small amount of variance explained. This raises the questions of what else should be measured and examined in conjunction with engagement and flow to account for the outcome variable, and whether the outcome variable chosen for this study was the most suitable for looking at UE.

In summary, in the video game domain, the UES failed to be a sensitive measure for [13, 48] and weakly predicted the criterion variable in Wiebe's [55] study. However, it did correlate well with the FSS and presented a four-factor structure similar to other studies subsequent to the original publication of the UES [37, 42].

3.2 Discussion of the User Engagement Scale

HCI rating scales have been aptly called a "tricky landscape" to traverse [31]. The case study of the UES reinforces this statement.

UES items, sub-scales, and the instrument as a whole have been used by researchers exploring UE with a range of applications and outcomes, including behavioural intentions for continued use, brand loyalty, learning, and system preferences. On a positive note, the adoption and adaptation of the UES in various domains implies that others have found the tool to be a useful instrument that resonates with their notion of engagement. Components of the UES have been combined with other self-report and objective measures (e.g. eye tracking, user behaviours) to generate interesting research questions, examine differences between experimental systems or conditions, and understand the relationship between user characteristics, engagement, and perceptions of hedonic and utilitarian technologies. Those authors who tested antecedents and outcomes of UE in specific contexts allow us to understand what predicts and is predicted by user engagement in these contexts. Overall, the literature reviewed in this chapter suggests that the UES is flexible, appropriate, and useful in terms of helping researchers achieve their goals and objectives. When reliability (i.e. internal consistency) was specifically tested, the UES passed the test. It demonstrated reasonable validity in most cases, correlating with other measures, such as the NASA-TLX, and helping to distinguish conditions or systems.

The UES also demonstrated its limitations across this set of studies. It was not always able to distinguish between experimental systems or conditions and did not correlate with cursor behaviour in some studies; person-dependent characteristics, such as preferences, seemed to factor heavily into perceived engagement, sometimes independent of the system or construct of interest in the research study. Studies that have examined the dimensionality of the UES support a four-factor structure with distinct focused attention, aesthetic appeal, and perceived usability sub-scales and items from the remaining three sub-scales (felt involvement, novelty, and endurability) loading on one factor. While we need more research to ascertain whether this reduction of factors is due to the underlying concept of engagement, the use of exploratory versus confirmatory factor/components analysis or a signal of a problem with the UES items [42], these findings do indicate that researchers who cannot perform factor/component analysis in their own studies and want to use the UES should adopt the four-factor structure.

The UES operationalized UE according to a set of attributes, a challenge since the definition of the concept is still maturing. The same range of biases and demand effects that affect all self-report instruments limits the UES, although no method is immune to shortcomings. Yet these limitations must be tempered with the fact that few researchers have used the instrument as a whole, and this leaves an incomplete picture of the UES's robustness. Some of the researchers do not address why they selected particular items or sub-scales, that is, what, in terms of their system, context, objectives, or theoretical orientation, motivated their choices. Therefore, we are unable to link theory and application in these cases. The use of one sub-scale cannot necessarily be equated with studying overall engagement. Studies could recognize formally that they are exploring one of several dimensions of UE.

Another challenge is the summation of engagement items to create an overall engagement "score". While this may be useful and appropriate in some cases, authors should do this with the understanding that the UES is multidimensional and its gradations may be lost by looking at the scale in a summative manner. This is reinforced by Devellis [18] who writes, "items must share one and only one underlying variable if they are to be combined into a scale… If a set of items is multidimensional (as a factor analysis might reveal), then the separate, unidimensional item groupings must be dealt with individually" (p. 159). In other words, aesthetic appeal, focused attention, perceived usability, etc., should be examined discretely with other variables of interest in the study. This reinforces the Process Model of User Engagement [34, 39] that proposes that UE attributes vary in intensity and significance depending on the context of use, yet all are necessary for examining UE holistically.

Pragmatically speaking, the 31-item UES may be cumbersome for researchers to use, particularly those testing multiple systems or asking participants to complete multiple trials. Participants are often asked to complete several questionnaires during an experiment, with the UES being one of many, and only so much can be included in user studies without running the risk of fatiguing users and compromising their responses. This is the reason why questionnaires such as the "quick and dirty", time-tested, and easy-to-use ten-item SUS [10] remain so appealing. Based on this review and the number of studies utilizing the UES in some capacity over the past 5 years, it is apparent that there is a need for a questionnaire that measures UE (as opposed to usability or other subjective experiences), but we must ensure that this instrument is robust and measures what it is supposed to be measuring. Increasing the number of studies that use the whole scale may provide insights into how we can create a brief version of the UES without compromising its reliability, validity, or dimensionality.

4 User Engagement Research: A Measurement Agenda

Methods and measures are growing in response to the complex phenomena of "third wave" HCI [8]. In the case of user engagement, for example, Grafsgaard's [20] work with patterns of facial expressions demonstrates the potential to capture objective data over the course of an interaction and to disambiguate positive (engagement) and negative (frustration) user experiences. Increasingly, studies are employing mixed methods and sophisticated analyses to examine the relationship between performance (cursor movements), physiology (eye tracking), and self-reports [3]. As many of the technologies used to capture physiology and user behaviour become more commercially available, and large data sets increase in accessibility, development of UE measurement practices is sure to be rapid.

The studies reviewed in this chapter and those on the measurement of UE more broadly suggest that researchers are working to address some of the challenges identified earlier in this chapter. Process-based measures, such as eye tracking, facial expression analysis, and other neurophysiological observations, are attempting to capture engagement as a dynamic concept that changes over time. Rather than looking at individual measures, researchers are instead identifying patterns that are more reliable indicators of user experience. Furthermore, there is also an attempt to bridge the issue of scale. The work of Arapakis et al. [3] shows great potential in this regard. Mouse clicks are relatively easy to collect in large-scale studies, whereas it is not feasible to collect self-report or individual physiological data in these environments. If mouse movements can be used as a reliable proxy of attention, both observed and latent, then this increases the potential for more in-depth analyses of attention across millions of users and would allow comparisons across Web domain areas at scale.

While existing and emerging work is promising, a challenge we continue to face is to demonstrate the construct validity of our measures: do they measure what they were designed to measure? The only way we will tackle this challenge is to ground our research studies theoretically and to use the findings of our work to inform theory. Looking within and beyond our own research domains, as this book attempts to do, opens up the conversation of what engagement is and how it can adequately be captured methodologically.

As the case study of the UES has shown, it is difficult to create a cohesive picture of measurement when instruments are used only in part or studies lack a rationale for why specific measures were chosen. As we forge ahead to look at the role of user engagement as a predictor, mediator, or outcome of other variables of interest, we must be mindful that the quality of this work and the ability to draw accurate conclusions depend upon the robustness of our measures.

In conclusion, a measurement agenda for UE would include support for the exciting and emerging work that is attempting to capture the dynamic nature of UE, the concurrent use of subjective and objective measures, and the interpretation of patterns rather than individual actions. However, this agenda would be furthered through the development of a parallel stream of research that looks intentionally at

UE methods and measures. This stream would focus on the reliability and validity of measures, replication of research findings with different populations or domain areas, and the assessment of methodological "fit" given the location, scale, and context in which the research takes place. Further research in this direction would provide researchers with a basis for comparison for their own work and the ability to make predictions about user engagement on the basis of others' findings. It would provide a solid basis upon which to conduct UE research, allow for the incorporation and assessment of new techniques and technologies into measurement practices as advances occur, and contribute to the evolution of user engagement theory.

Acknowledgements I wish to express appreciation to the Social Science and Humanities Research Council of Canada and the Networks of Centres of Excellence in Graphics, Animation and New Media Project for funding support.

References

1. Agarwal, R., Karahanna, E.: Time flies when you're having fun: cognitive absorption and beliefs about information technology usage. Manag. Inf. Syst. Q. **24**(4), 665–694 (2000)
2. Arapakis, I., Lalmas, M., Cambazoglu, B., Marcos, M., Jose, J.M.: User engagement in online news: under the scope of sentiment, interest, affect, and gaze. J. Assoc. Inf. Sci. Technol. **65**(10), 1988–2005 (2014)
3. Arapakis, I., Lalmas, M., Valkanas, G.: Understanding within-content engagement through pattern analysis of mouse gestures. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 1439–1448. ACM, New York (2014)
4. Arapakis, I., Bai, X., Cambazoglu, B.B.: Impact of response latency on user behavior in web search. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 103–112. ACM, New York (2014)
5. Arguello, J., Wu, W.-C., Kelly, D., Edwards, A.: 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, pp. 435–444. ACM, New York (2012)
6. Bangcuyo, R.G., Smith, K.J., Zumach, J.L., Pierce, A.M., Guttman, G.A., Simons, C.T.: The use of immersive technologies to improve consumer testing: the role of ecological validity, context and engagement in evaluating coffee. Food Qual. Prefer. **41**, 84–95 (2015)
7. Banhawi, F., Ali, N.M.: Measuring user engagement attributes in social networking application. In: Proceedings of the 2011 International Conference on Semantic Technology and Information Retrieval (STAIR), pp. 297–301. IEEE, New York (2011)
8. Bödker, S.: When second wave HCI meets third wave challenges. In: Proceedings of the 4th Nordic Conference on Human-Computer Interaction (NordCHI '06), pp. 1–8. ACM, New York (2006). doi:10.1145/1182475.1182476
9. Bron, M., van Gorp, J., Nack, F., Baltussen, L.B., de Rijke, M.: Aggregated search interface preferences in multi-session search tasks. In: Proceedings of the 36th International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 123–132. ACM, New York (2013)
10. Brooke, J.: SUS-A quick and dirty usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (eds.) Usability Evaluation in Industry, pp. 4–7. Taylor and Francis, London (1996)

11. Busselle, R., Bilandzic, H.: Measuring narrative engagement. Media Psychol. **12**(4), 321–347 (2009)
12. Bustillo, J., Garaizar, P.: Scratching the surface of digital literacy…but we need to go deeper. In: 2014 IEEE Frontiers in Education Conference Proceedings, pp. 1440–1443. IEEE, New York (2014)
13. Choi, H.S.: The impact of visuospatial characteristics of video games on improvements in cognitive abilities. Unpublished Doctoral Dissertation, North Carolina State University (2013)
14. Cian, L., Krishna, A., Elder, R.S.: This logo moves me: dynamic imagery from static image. J. Mark. Res. **51**(2), 184–197 (2013)
15. Creswell, J.W.: Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage, Los Angeles (2003)
16. De Moor, K., Mazza, F., Hupont, I., Ríos Quintero, M., Mäki, T., Varela, M.: Chamber QoE: a multi-instrumental approach to explore affective aspects in relation to quality of experience. In: Proceedings of SPIE 9014, Human Vision and Electronic Imaging XIX, 90140U (2014). doi:10.1117/12.2042243
17. Deci, E.L., Ryan, R.M.: Handbook of Self-Determination Research. University Rochester Press, New York (2012)
18. Devellis, R.F.: Scale Development: Theory and Applications, 2nd edn. Sage, Thousand Oaks (2003)
19. Fuchsberger, V., Sellner, W., Moser, C., Tscheligi, M.: Benefits and hurdles for older adults in intergenerational online interactions. In: Miesenberger, K., Karshmer, A., Penaz, P., Zagler, W. (eds.) Computers Helping People with Special Needs. Lecture Notes in Computer Science, vol. 7382, pp. 697–704. Springer, Berlin/Heidelberg (2012)
20. Grafsgaard, J.F.: Multimodal affect modeling in task-oriented tutorial dialogue. Unpublished Doctoral Dissertation, North Carolina State University (2014)
21. Halpern, D.: Towards a networked public sphere: how social media triggers civic engagement through news consumption and political discussion. Unpublished Doctoral Dissertation, Rutgers University (2013)
22. Hong, T.: Internet health search: when process complements goals. J. Am. Soc. Inf. Sci. Technol. **63**(11), 2283–2293 (2012)
23. Jacques, R.D.: The nature of engagement and its role in hypermedia evaluation and design. Unpublished Doctoral Dissertation, South Bank University (1996)
24. Kajalainen, K.: Increasing the enjoyment of online video increasing the enjoyment of online video content with topical interactivity. Unpublished Doctoral Dissertation, Aalto University (2015)
25. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. Found. Trends Inf. Retr. **3**(1–2), 1–224 (2009)
26. Kwak, N., Williams, A.E., Wang, X., Lee, H.: Talking politics and engaging politics: an examination of the interactive relationships between structural features of political talk and discussion engagement. Commun. Res. **32**(1), 87–111 (2005)
27. Lalmas, M. O’Brien, H., Yom-Tov, E.: Measuring user engagement. Synth. Lect. Inf. Concepts Retr. Serv. **6**(4), 1–132 (2014)
28. Lawrence, B., Fournier, S., Brunel, F.: When companies don’t make the ad: a multimethod inquiry into the differential effectiveness of consumer-generated advertising. J. Advert. **42**(4), 292–307 (2013)
29. Lee, Y.-C., Lin, W.-C., Cherng, F.-Y., Wang, H.-C., Sung, C.-Y., King, J.-T.: Using time-anchored peer comments to enhance social interaction in online educational videos. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 689–698. ACM, New York (2015)
30. Levesque, V., Oram, L., MacLean, K., Cockburn, A., Marchuk, N.D., Johnson, D., Colgate, J.E., Peshkin, M.A.: Enhancing physicality in touch interaction with programmable friction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2481–2490. ACM, New York (2011)

31. Lindgaard, G., Kirakowski, J.: Introduction to the special issue: the tricky landscape of developing rating scales in HCI. Interact. Comput. **25**(4), 271–277 (2013)
32. McCay-Peet, L., Lalmas, M., Navalpakkam, V.: On salience, affect and focused attention. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 541–550. ACM, New York (2012). doi:10.1145/2207676.2207751
33. Moshfeghi, Y., Matthews, M., Blanco, R., Jose, J.M.: Influence of timeline and named-entity components on user engagement. In: Serdyukov, P., Braslavski, P., Kuznetsov, S., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) Advances in Information Retrieval. Lecture Notes in Computer Science, vol. 7814, pp. 305–317. Springer, Berlin/Heidelberg (2013)
34. O'Brien, H.: Defining and measuring user experiences with technology. Unpublished Doctoral Dissertation, Dalhousie University (2008)
35. O'Brien, H.: The influence of hedonic and utilitarian motivations on user engagement: the case of online shopping experiences. Interact. Comput. **22**(5), 344–352 (2010)
36. O'Brien, H.: The role of story and media in user engagement with online news. In: Proceedings of the Annual Conference of Canadian Association of Information Science (CAIS), CAIS, Victoria (2013)
37. O'Brien, H., Cairns, P.: An empirical evaluation of the user engagement scale (UES) in online news environments. Inf. Process. Manag. **51**(4), 413–427 (2015)
38. O'Brien, H., Lebow, M.: A mixed methods approach to measuring user experience in online news interactions. J. Assoc. Inf. Sci. Technol. **64**(8), 1543–1556 (2013)
39. O'Brien, H., Toms, E.G.: What is user engagement? A conceptual framework for defining user engagement with technology. J. Am. Soc. Inf. Sci. Technol. **59**(6), 938–955 (2008)
40. O'Brien, H., Toms, E.G.: The development and evaluation of a survey to measure user engagement in e-commerce environments. J. Am. Soc. Inf. Sci. Technol. **61**(1), 50–69 (2010)
41. O'Brien, H., Toms, E.G.: Is there a universal instrument for measuring interactive information retrieval? The case of the user engagement scale. In: Proceedings of Information Interaction in Context (IIiX), pp. 335–340. ACM, Rutgers (2010). doi:10.1145/1840784.1840835
42. O'Brien, H., Toms, E.G.: Examining the generalizability of the user engagement scale (UES) in exploratory search. Inf. Process. Manag. **49**(5), 1092–1107 (2013)
43. Okoro, E.M.: A study of different representation conventions during investigatory sensemaking. Unpublished Masters Thesis, Middlesex University (2014)
44. Parra, D., Brusilovsky, P.: User-controllable personalization: a case study with SetFusion. Int. J. Hum. Comput. Stud. **77**, 43–67 (2014)
45. Peterson, R.A.: Constructing Effective Questionnaires. Sage Publications, Thousand Oaks (2000)
46. Reitz, A.R.: Online consumer engagement: understanding the antecedents and outcomes. Unpublished Doctoral Dissertation, Colorado State University (2012)
47. Seedorf, S., Thum, C., Schulze, T., Pfrogner, L.: Social co-browsing in online shopping: the impact of real-time collaboration on user engagement. In: Proceedings of the Twenty Second European Conference on Information Systems, Tel Aviv 2014. AIS Electronic Library (2014)
48. Sharek, D.J.: Investigating real-time predictors of engagement: implications for adaptive video games and online training. Unpublished Doctoral Dissertation, North Carolina State University (2012)
49. Shirzad, N.: The use of physiological signals and motor performance metrics in task difficulty adaptation: improving engagement in robot-assisted movement therapy. Unpublished Doctoral Dissertation, The University of British Columbia (2013)
50. Tabachnick, B.G., Fidell, L.S.: Using Multivariate Statistics, 6th International edition (cover) edn. Sage Publications, Thousand Oaks (2013)
51. Vail, A.K., Boyer, K.E., Wiebe, E.N., Lester, J.C.: The Mars and Venus effect: the influence of user gender on the effectiveness of adaptive task support. In: Ricci, F., Bontcheva, K., Conlan, O., Lawless, S. (eds.) User Modeling, Adaptation and Personalization. Lecture Notes in Computer Science, vol. 9146, pp. 265–276. Springer, Berlin/Heidelberg (2015)

52. Warnock, D., Lalmas, M.: An exploration of cursor tracking data. arXiv preprint (2015). arXiv:1502.00317
53. Webster, J., Ho, H.: Audience engagement in multimedia presentations. Data Base Adv. Inf. Syst. **28**(2), 63–77 (1997)
54. Whitman, L.: The effectiveness of interactivity in multimedia software tutorials. Unpublished Doctoral Dissertation, North Carolina State University (2013)
55. Wiebe, E.N., Lamb, A., Hardy, M., Sharek, D.: Measuring engagement in video game-based environments: investigation of the user engagement scale. Comput. Educ. **32**, 123–132 (2014)
56. Witmer, B.G., Singer, M.J.: Measuring presence in virtual environments: a presence questionnaire. Presence: Teleoper. Virtual Environ. **7**(3), 225–240 (1998)
57. Zhu., Y., Heynderick, I., Redi, J.A.: Alone or together: measuring users' viewing experience in different social contexts. In: Proceedings of SPIE 9014, Human Vision and Electronic Imaging XIX, 90140W (2014). doi:10.1117/12.2042867