Int. J. Human-Computer Studies 112 (2018) 28–39

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form

Heather L. O'Brien[a,*], Paul Cairns[b], Mark Hall[c]

[a] School of Library, Archival and Information Studies, University of British Columbia, Vancouver, Canada
[b] Department of Computer Science, University of York, York, UK
[c] Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany

ARTICLE INFO

Keywords:
User engagement
Questionnaires
Measurement
Reliability
Validity

ABSTRACT

User engagement (UE) and its measurement have been of increasing interest in human-computer interaction (HCI). The User Engagement Scale (UES) is one tool developed to measure UE, and has been used in a variety of digital domains. The original UES consisted of 31-items and purported to measure six dimensions of engagement: aesthetic appeal, focused attention, novelty, perceived usability, felt involvement, and endurability. A recent synthesis of the literature questioned the original six-factors. Further, the ways in which the UES has been implemented in studies suggests there may be a need for a briefer version of the questionnaire and more effective documentation to guide its use and analysis. This research investigated and verified a four-factor structure of the UES and proposed a Short Form (SF). We employed contemporary statistical tools that were unavailable during the UES' development to re-analyze the original data, consisting of 427 and 779 valid responses across two studies, and examined new data (N=344) gathered as part of a three-year digital library project. In this paper we detail our analyses, present a revised long and short form (SF) version of the UES, and offer guidance for researchers interested in adopting the UES and UES-SF in their own studies.

1. Introduction

User engagement is a quality of user experience characterized by the depth of an actor's investment when interacting with a digital system (O'Brien, 2016a). Engagement is more than user satisfaction: it is believed that the ability to engage and sustain engagement in digital environments can result in positive outcomes for citizen inquiry and participation, e-health, web search, e-learning, and so on. Yet user engagement (UE) is an abstract construct that manifests differently within different computer-mediated contexts, and this has made it challenging to define, design for, and evaluate.

This research is fundamentally focused on the challenge of measuring engagement so that it can be used in design and evaluation. A range of methodological approaches have been utilized to measure engagement, including (Lalmas et al., 2014; O'Brien and Cairns, 2016):

- behavioural metrics such as web page visits and dwell time;
- neurophysiological techniques such as eye tracking and electrodermal activity (EDA);
- self-reports such as questionnaires, interviews, diary entries and verbal elicitation.

All methodological approaches have their advantages and limitations with respect to use with specific populations, settings, and time scales, from a single user-computer interaction to longitudinal observations. In addition, measures may capture interactions formatively or summatively, and subjectively or objectively (Lalmas et al., 2014). In general, there has been advocacy for multiple measures and mixed methods to reliably and validly capture constructs such as user engagement. This requires attention to the robustness of individual measures, as well as to triangulating multiple measures.

Our work is concerned with the User Engagement Scale (UES), a 31-item experiential questionnaire. The UES (or items derived from it) has been used to evaluate engagement in a range of settings: information search, online news, online video, education, and consumer applications, haptic technologies, social networking systems, and video games (see (O'Brien, 2016b) for an overview of this work). Although there is evidence to suggest that the UES is a reliable and valid means of

* Corresponding author.
E-mail addresses: h.obrien@ubc.ca (H.L. O'Brien), paul.cairns@york.ac.uk (P. Cairns), mark.hall@informatik.uni-halle.de (M. Hall).
URL: http://www.heatherobrien.arts.ubc.ca (H.L. O'Brien), https://www-users.cs.york.ac.uk/~pcairns/ (P. Cairns)

https://doi.org/10.1016/j.ijhcs.2018.01.004
Received 21 July 2017; Received in revised form 22 December 2017; Accepted 9 January 2018
Available online 10 January 2018

capturing subjective user engagement, some findings have questioned its effectiveness, which are reported in O'Brien (2016b). Such findings may point to flaws in the UES, the ways it has been administered and analyzed in practice, or some combination of these. For instance, few researchers have used the UES in its entirety, which makes it difficult to assess its factor structure and robustness over time and across different digital applications. On the other hand, the decision to not use all 31 items raises pragmatic issues of using the UES in a study (i.e., length), or poor documentation regarding how to adapt, implement, and make meaning from the measurement tool.

In the current research, we applied state-of-the-art statistical techniques to re-analyze the data originally collected to develop the UES. Based on our findings, we proposed a revised long-form and short-form (SF) of the questionnaire, which we then evaluated with a new data set collected over a three-year period as part of a large digital library project. In the remainder of this paper, we provide background information on the UES and our approach to data analysis; present the revised UES and UES-SF with an explanation of our findings, and conclude with recommendations for the administration and analysis of the UES and UES-SF in future studies.

Our contribution is three-fold:

- firstly, we offer a robust measurement tool to measure user engagement in HCI settings; this tool can be used to guide the design of digital media or to evaluate user experience with computer-mediated systems;
- secondly, the validated UES can be confidently used as a benchmarking and corroborating tool for emerging methodological approaches or process-based metrics; and
- finally, we hope to improve the administration of the UES and other self-report questionnaires by providing guidance on how to adapt and interpret the UES in different research contexts.

2. User engagement

User engagement (UE) is a quality of user experience characterized by the depth of an actor's cognitive, temporal, affective and behavioural investment when interacting with a digital system (O'Brien, 2016a). Over the past two decades, the human-computer interaction (HCI) community has become increasingly interested in understanding, designing for and measuring user engagement with a host of computer-mediated health, education, gaming, social and news media, and search applications (O'Brien and Cairns, 2016). Collectively this work has demonstrated that UE is highly context dependent: each digital environment features unique technological affordances that interact with users' motivations to achieve some desirable end. For instance, in Massive Open Online Courses (MOOCs), learners participate for a variety of reasons, from professional development to curiosity about the topic, and take advantage of digital learning objects, like videos of lectures or quizzes, and opportunities for social interaction, say on discussion forums, to different degrees. Thus MOOC developers must take into account how individuals' goals and needs shape their investment in the course and what they wish to gain from it (Wiebe and Sharek, 2016). Designing for UE in news environments may be quite distinct. While personal goals may drive news interactions to some extent, content (and its presentation) generates situational interest, which in turn fosters engagement (Arapakis et al., 2014; O'Brien and McKay, 2016; Oh and Sundar, 2015). These examples illustrate that digital environments attract users for different reasons (e.g., to learn, to share, to stay current), and seek to sustain engagement for different durations (e.g., a daily ten minute news browsing session, a ten module MOOC) to achieve specific outcomes (e.g., continued loyalty to a news provider, MOOC completion).

The dynamic and variable nature of computer-mediated interactions is compounded by the abstractness of UE. There is some consensus that user engagement is affective, cognitive and behavioural in nature (O'Brien, 2016a; O'Brien and Toms, 2008). This idea is drawn from learning sciences research on student engagement: emotional engagement refers to the positive and negative responses students have to peers, teachers, and so on that influences their attachment to and willingness to work at school; cognitive engagement is the degree of effort students are willing to expend to master ideas and skills; and behavioural engagement involves participation in academic, social and extracurricular activities that discourages negative outcomes, such as dropping out (Fredricks et al., 2004) (p. 58). In HCI, users have emotional reactions to the system (e.g., frustration), content (e.g., shock, interest) or other users operating within the interaction space. Cognitively, the relationship between users' skills and the difficulty of the task determines the degree of mental effort required by users, and whether this results in boredom, engagement or frustration. Lastly, behavioural engagement refers to users' actions, such as clicking or querying, and frequency and duration of use.

Despite the recognition that engagement is multifaceted, a persistent challenge involves understanding what aspects of users' interactions with digital applications are indicative of user engagement. Time on task or physiological arousal may either suggest engagement with an application, or disorientation and frustration (O'Brien and Lebow, 2013; Webster and Ahuja, 2006). Several scholars in recent years have attempted to disambiguate these two contrasting experiences that share similar behavioural and physiological indicators. Edwards (2015) monitored electrodermal activity in participants completing frustrating and non-frustrating search tasks, where frustration was manipulated with different search results response latencies, while Grafsgaard examined facial expressions and body posture/movement as students interacted with an intelligent tutoring systems (Grafsgaard, 2014). Both researchers attempted to show different patterns inherent in engaged and frustrated participants by corroborating physiological data with other measures, including self-report questionnaires. While capable of monitoring interactive processes over time and in real-time (Rowe et al., 1998), neurophysiological methods are still developing as researchers continue to devise techniques for filtering noisy signals, making sense of the large volume of data generated, and syncing signals from different data sources, e.g., eye tracking and performance behaviour (Taub et al., 2017). In addition, interpreting signals to represent a psychological state such as engagement effectively requires an understanding of the concept itself.

One approach to operationalizing the concept of UE has been to isolate user-system attributes that constitute an engaging experience. Working in the area of educational multimedia, Jacques proposed six attributes of UE: attention (divided or focused), motivation, perception of control, needs satisfaction, perception of time ("dragging on" or "flying by") and positive or negative attitude (Jacques, 1996) (p. 67); Webster and Ho distinguished attributes of engagement, such as attention focus, curiosity, and intrinsic interest, from influences on engagement like challenge, control, feedback and variety (Webster and Ho, 1997) in their research on presentation software. Through a systematic multidisciplinary literature review and exploratory interview study with online learners, shoppers, searchers and gamers, O'Brien put forward existing and additional attributes of UE: challenge, aesthetic and sensory appeal, feedback, novelty, interactivity, perceived control and time, awareness, motivation, interest, and affect. These were mapped to a stage-based Process Model of User Engagement consisting of a point of engagement, period of sustained engagement, disengagement, and reengagement, where the attributes were depicted as ebbing and flowing according to the changing needs of users as they moved through dynamic digital interactions (O'Brien, 2008; O'Brien and Toms, 2008).

3. An attribute-Based approach to user engagement

An attribute-based approach to the definition of UE has the advantage of helping researchers operationalize user experience design guidelines or measurement tools. Jacques constructed ten design principles for engaging educational multimedia based on the attributes of UE he

articulated. For instance, one principle emphasized the needs for users' to feel in control of the interaction:

> Give the user control and support: Users that feel in control are more likely to feel engaged. Part of feeling in control is knowing that the software is supportive and this can achieved through consistency, for instance, knowing that something can always be found in one location; reversibility, that is, being able to go back and change an event; and facilities such as 'help' (Jacques, 1996) (p. 93).

Similarly, Sutcliffe devised recommendations for enhancing user experience, e.g., to attract attention, to persuade, to arouse emotion, through the choice of media (e.g., photographs, characters) or design decisions around colour, visual salience and so on (Sutcliffe, 2009).

With regards to measurement, several researchers have developed self-report questionnaires based on attributes of UE (Jacques, 1996; O'Brien, 2008; Webster and Ho, 1997). Jacques's Survey to Evaluate Engagement (SEE) consisted of 14 questions or items related to the six attributes he identified (attention, motivation, controls, needs satisfaction, time perception, and attitude); Jacques recommended that SEE could be used to yield a global engagement score for users or a score for each of the attributes (Jacques, 1996). Webster and Ho's questionnaire was made up of two questions for each of the Engagement and Influences on Engagement measures, as well as an overall item, "The presentation medium was engaging," for a total of fifteen questions (Webster and Ho, 1997) . O'Brien built on the work of Jacques and Webster and Ho with the User Engagement Scale (UES) (O'Brien, 2008; O'Brien and Toms, 2010a). She elected to devise a new questionnaire to represent and validate the additional UE attributes arising from her systematic review and interview study (O'Brien, 2008; O'Brien and Toms, 2008).

The UES was constructed through an iterative process of scale development and evaluation that involved gathering, refining and assessing the appropriateness of potential items, pretesting items, and conducting two large online surveys in the e-commerce domain. The first online survey was administered to 440 general online shoppers and exploratory factor analysis was used to reduce the original 124 items to a more parsimonious set and to examine the factor structure inherent in the data. The second online survey, consisting of 40–50 items, was targeted to shoppers of a specific company who had made a recent purchase; responses from approximately 800 individuals were used to perform structural equation modelling to confirm the factor structure observed in the first study, and to explore the relationship between factors using path analysis. Overall, this work resulted in a 31 item self-report instrument that comprised six factors or dimensions:

- FA: Focused attention, feeling absorbed in the interaction and losing track of time (7 items).
- PU: Perceived usability, negative affect experienced as a result of the interaction and the degree of control and effort expended (8 items).
- AE: Aesthetic appeal, the attractiveness and visual appeal of the interface (5 items).
- EN: Endurability, the overall success of the interaction and users' willingness to recommend an application to others or engage with it in future (5 items).
- NO: Novelty, curiosity and interest in the interactive task (3 items).
- FI: Felt involvement, the sense of being "drawn in" and having fun (3 items).

While there is overlap amongst the SEE, Webster and Ho's questionnaires, and the UES (e.g., attention is a central component) there are differences. One important structural distinction is that three or more items constitute each dimension of the UES, whereas the other questionnaires include two items per dimension. This has implications, particularly when measures are new and not yet validated, regarding whether a subscale's items adequately and consistently capture the attribute of interest; in other words, more questions increase the reliability of the scale, and a cautious approach is to establish reliability before seeking to reduce the number of items (Devellis, 2003).

4. User Engagement Scale

The User Engagement Scale (UES) has been used widely since its publication (O'Brien and Toms, 2010a). Recently, O'Brien conducted a synthesis of over forty published works that have utilized it to investigate UE in a range of digital domains: information search, online news, online video, education, and consumer applications, haptic technologies, social networking systems, and video games (O'Brien, 2016b). Since few researchers have used the UES in its entirety, it is difficult to evaluate its generalizability. Instead many studies feature a selection of items or specific subscales, which is also problematic for examining the robustness of the scale. As Kazdin notes, "tinkering" with a scale by altering its content or format threatens its construct validity; removing questions may mean that the construct is no longer adequately captured by the remaining items or may result in less variability in scale scores, limiting the potential to detect differences between samples, experimental conditions, or, in the case of HCI, systems (Kazdin, 2016). For those using select subscales only, it must be remembered that what is being measured is not engagement but some component of it, such as focused attention (O'Brien and McCay-Peet, 2017). This may be as relevant as examining all of the UES dimensions depending on a particular study's goals, but what is actually being measured must be articulated.

Studies that have made use of the subscales or entire UES have revealed that the questionnaire has, in general, demonstrated good reliability and validity. Specifically, the UES is associated with other self-report measures in ways we would expect, e.g., the focused attention subscale has been shown to correlate with measures of cognitive absorption and flow (O'Brien and Lebow, 2013; Wiebe et al., 2014), while the perceived usability subscale has been related to other usability questionnaires (O'Brien and Lebow, 2013). Findings on the relationship between the UES and neurophysiological or behavioural measures and on the UES's ability to detect differences in engagement between experimental conditions or systems has been mixed. However, it is unclear whether this was due to the sensitivity of the questionnaire or how it was implemented in the studies (e.g., using select items rather than complete subscales); in addition, some studies demonstrated differences between systems and conditions once they re-analyzed the data with respect to user preferences (Arguello et al., 2012). O'Brien acknowledged that one of the reasons that researchers may select items or subscales is the length of the UES (O'Brien, 2016b); 31 items may not seem like a large number, but this length may be cumbersome and repetitive in experimental studies that involve multiple trials or conditions, or where other questionnaires are also being used. This could result in participant fatigue in experimental settings or attrition in field-based work.

Of the studies that have used the UES in its entirety, there is strong evidence to suggest that it has four factors rather than six as originally proposed. Research conducted with Facebook users, gamers, online news browsers, and exploratory searchers has used factor or principle components analysis to analyze the UES, suggesting a four factor structure (Banhawi and Ali, 2011; O'Brien and Cairns, 2015; O'Brien and Toms, 2013; Wiebe et al., 2014). Three of the UES subscales, aesthetic appeal, focused attention, and perceived usability have been relatively stable, while the endurability, novelty and felt involvement scales have typically combined into a fourth factor. One exception to this was an analysis of webcast users' engagement where there were distinct factors for aesthetic appeal, novelty, focused attention, and endurability, but the felt involvement items were dropped and the perceived usability items split into two groups: cognitive demands of the task and users' affective responses to the system (O'Brien and Toms, 2010b). In addition to aspects of dimensionality, different numbers of UES items were retained in these analyses, and this may be due to the use of reductionist statistical techniques, such as exploratory factor analysis, or the fit of particular items in the context of use. For example, recent work

exploring engagement with different media conditions resulted in removing aesthetic appeal items after an examination of missing values and reflection on the appropriateness of the questions for audio and text presentations of content compared to video (O'Brien, 2017).

The synthesis of the UES's uptake and use since its publication has presented several avenues for future work (O'Brien, 2016b). First, the widespread use of the UES in various HCI domains suggested that there was a need for an instrument to capture UE from the users' perspective, and that the UES's conceptualization of UE as a multi-dimensional construct resonated within the community. Second, the four-factor structure that emerged in those studies that used the UES in its entirety indicated the need to look closely at the dimensions of the UES and either confirm the six factor structure or propose and validate a new four-factor scale. Third, the use of select items or subscales rather than the entire UES signalled that the questionnaire may be too lengthy for some research contexts and a briefer version was warranted. Further, it may be inferred that researchers were unclear about how the implementation of a questionnaire impacts its reliability and validity, and therefore better documentation of how to use and analyze the UES was deemed essential. This paper is a direct response to these identified methodological issues. Specifically, our goals were to: verify the dimensionality of the UES, develop and test the robustness of a brief version or short form (SF), and provide recommendations for the adoption and use of the UES by HCI researchers.

5. Methodology

The original UES (O'Brien and Toms, 2010a) gave a six-factor description, but subsequent evaluations have suggested a four-factor description (Banhawi and Ali, 2011; O'Brien and Cairns, 2015; O'Brien and Toms, 2013; Wiebe et al., 2014). Therefore, the first goal of the analysis was to re-evaluate the original UES data to confirm the four or six factor structure, and determine how the 31 questionnaire items grouped together. The second goal was to use our knowledge of the validated factor structure to propose a short-form of the UES.

To improve on the original analysis, we used the most current statistical tools to provide an analysis more faithful to the data. These tools have only recently become easily available, specifically through a new package called `mirt` (Chalmers et al., 2012) for the R statistics program. As there are several stages to this analysis, we first provide an overview of the analysis including a justification of the techniques used to help the reader understand the overall approach before being given the details and results of the actual analysis.

The `mirt` package provides multidimensional analysis of questionnaire data, or what is commonly called factor analysis. This is used to identify the latent (hidden) concepts that underpin a questionnaire which emerge as groups of items that form a factor (or dimension) in the dataset. The difference between `mirt` and existing tools such as Principal Component Analysis in SPSS is that `mirt` is based on Item Response Theory (IRT) (Embretson and Reise, 2013). IRT differs from traditional Classical Test Theory in that it recognises that individual questionnaire items may be responded to in a way that reflects both difference in the items and differences between respondents. Classical Test Theory conflates these variations (though still produces robust results), but it is more accurate to account for these using IRT. In particular, IRT includes parameters in its analysis to reflect that some items may elicit more positive responses, some may discriminate more between values of the underlying traits, and some are more susceptible to random answers. Because of these extra parameters for every item of a test, analysis based on IRT requires sophisticated numerical algorithms that can be computationally expensive. Relatively recent advances in theory have made the practical analysis of multi-dimensional questionnaires possible and `mirt` is the first package to support this.

Parallel analysis using polychoric correlations (Galbraith et al., 2002), another technique to recently emerge, was used to analyse two *eShopping* datasets, which we labelled *eShopping1* and *eShopping2* and that were collected as part of the first author's dissertation, to examine the number of factors inherent in the UES. As the whole UES relates to engagement, there was an expectation that there would be a single underlying factor of engagement with different factors reflecting particular aspects of the concept. In other words, though there may be distinct factors of engagement, these factors were expected to correlate because they all relate to engagement.

To examine the data various factor models were explored using `mirt`. First, a single factor model was developed to see if there was a unified concept underyling the data. Secondly 4 and 6 factor exploratory models were produced to see which resulted in the best description of the data. Thirdly, the expectation of 4 and 6 factor models, based on the original analysis and subsequent studies, was considered using confirmatory factor analysis in `mirt`. However, unlike normal confirmatory factor analysis where factors are independent, a bifactor model was fitted (Reise, 2012). The bifactor models represented the expected 4 or 6 factor models but also included a general factor of all items in the questionnaire, again to reflect that the whole questionnaire measured engagement. The quality of the factors in the various models was also assessed both against the original factors and with statistical measures of reliability and the amount of variance accounted for by the factors.

Based on the outcome of our examination of the 4 and 6 factor models, we proposed a Short Form of the UES (UES-SF) and tested its reliability with *eShopping2*. Next, we tested the validity of the UES-SF with a previously unseen data set arising from three years of an information search system evaluation (*Social Book Search (SBS)*). The basic process of validation was to produce UES scores for each factor using the UES-SF. It was not sensible to correlate these with the full UES scores because they included the SF items. Instead, the UES-SF scores were correlated with a remainder score, which was the score generated from the remaining items in each corresponding factor. This therefore considers whether the SF factors adequately reflected the portion of the factors not measured by the SF. Also, as a confirmation, a bifactor four-factor confirmatory model was also produced for *SBS*.

6. Results

All analyses were done using the `mirt` package, v1.21, in R. The main function used was called mirt and this fits an unconditional maximum likelihood factor analysis model under the item response theory (IRT) paradigm (Chalmers et al., 2012). In exploratory analysis, it is possible to specify the number of factors in the model and oblimin rotation was used in all cases so that it was possible for factors to correlate. In confirmatory analysis, a model is used to specify which items belong to which factor and the mirt function generates loadings to best fit that model, in particular items which are modelled to fit a particular factor have their loadings constrained to be 0 on that factor.

6.1. Re-examining the factor structure of the UES

6.1.1. eShopping1

The first dataset examined was *eShopping1*. In the previous study, respondents (N=440) completed a questionnaire with 123 items; through statistical analyses, the 123 questionnaire items were reduced to a more parsimonious set. In the current analysis, due to several substantially incomplete responses, we used the data for 427 participants over the 31 items of the final UES. The following analysis was only conducted with the final 31 items of the UES to see how they functioned as a unit within the original data. Note, the use of 20-20 hindsight made this possible and hence accounts for discrepancies seen in this later analysis from the original analysis. The factor loadings for the 1, 4 and 6 factor models are given in Table 1. Parallel analysis with polychoric correlations suggested that there were four factors underlying the data.

A single factor exploratory analysis was conducted to see if there was a unifying concept throughout all UES items. The single factor accounted for approximately 37% of the total variance in the data. Though many

Table 1
Loadings of factors from the 1, 4 and 6 factor models of eShopping1. Loadings of magnitude less than 0.3 are omitted for clarity.

Scale	One	Four				Six					
Name	F1	F1	F2	F3	F4	F1	F2	F3	F4	F5	F6
FA.1			−0.84				−0.49				
FA.2			−0.85				−0.36				
FA.3			−0.74				−0.38				
FA.4			−0.78				−0.43				
FA.5			−0.74				−0.40				
FA.6	0.41		−0.66				−0.39				
FA.7			−0.71								
PU.1	0.87	0.69				0.86					
PU.2	0.78	0.60									0.69
PU.3	0.87	0.65				0.75					
PU.4	0.83	0.55			0.33						0.74
PU.5	0.54	0.48	0.37								0.57
PU.6	0.63	0.61				0.60					
PU.7	0.69	0.53									0.55
PU.8	0.66	0.44									0.92
AE.1	0.68			−0.92				−0.36			−0.47
AE.2	0.59			−0.85				−0.58			
AE.3	0.60			−0.85				−0.36			−0.40
AE.4	0.64			−0.78				−0.55			
AE.5	0.70			−0.78				−0.61			
EN.1	0.74				0.75				0.39		
EN.2	0.71				0.74				0.47		
EN.3	0.66				0.59	0.37					
EN.4	0.75				0.61				0.44		
EN.5	0.73				0.55				0.36		
NO.1	0.42		−0.39							0.47	
NO.2	0.57		−0.37							0.54	
NO.3	0.62				0.43						
FI.1	0.51	0.54	−0.40								
FI.2	0.53	0.46									
FI.3	0.72	0.58									

of the UES items did load on this factor, surprisingly, only one of the FA items loaded on the single dimension with a loading above 0.3. This suggests that, in this data, the FA component did not support the notion of a single underlying concept of engagement.

A six factor exploratory analysis failed to produce a good structure with the final UES items. All three of the FI items failed to load on any factor with loadings above 0.3 and the generally strong FA factor had an item which did not load on any factor. Additionally, the PU component split across two distinct factors with no cross loading. Moreover, the six factors accounted for only 34% of the variance in the data, less than the one factor model. This suggested that the six factor structure of the UES is not strongly supported with the original dataset.

By contrast, a four factor exploratory model gave good factor structures with only one item in NO not loading on the same factor as the other items in its scale. There was some sharing of items with FA and NO (excepting one item) loading on the same factor, PU and FI forming another factor and AE and EN both forming separate factors. The model accounted for 49% of the variance in the data. This supported that there was a meaningful four factor model for the UES represented in this dataset, and that the original 6 factors were coherent components within the data but not independent factors.

6.1.2. eShopping2

Turning to the second of the original datasets, *eShopping2*, a similar analysis was conducted. This dataset included the reduced but not final set of 49 UES items alongside other measures that were used to assess the concurrent validity of some of the UES items. The data was gathered from 794 participants, owing to some missing responses. As with *eShopping1*, parallel analysis based on polychoric correlations of the final 31 UES items suggested four underlying factors.

To examine whether there was a single underlying component of engagement, a 1-factor exploratory model was generated (see Table 2). All items of the UES except for NO.1, loaded on this factor with loadings

Table 2
Factor loadings of the 1 and 4 factor models for the *eShopping2* dataset. Loadings of magnitude less than 0.3 are omitted for clarity.

New	Original	One	Four			
Name	Name	F1	F1	F2	F3	F4
FA.1		0.62			−0.95	
FA.2		0.60			−0.87	
FA.3		0.57			−0.75	
FA.4		0.60			−0.87	
FA.5		0.66			−0.84	
FA.6		0.57			−0.52	
FA.7		0.62			−0.84	
PU.1		0.76				0.92
PU.2		0.76				0.90
PU.3		0.77				0.87
PU.4		0.71				0.82
PU.5		0.70				0.80
PU.6		0.72	−0.38			
PU.7		0.67				0.76
PU.8		0.77				0.77
AE.1		0.76		−0.82		
AE.2		0.74		−0.83		
AE.3		0.69		−0.81		
AE.4		0.76		−0.69		
AE.5		0.72		−0.85		
RW.1	EN.1	0.82	−0.66			
RW.2	EN.2	0.79	−0.67			
RW.3	EN.3	0.71				0.70
RW.4	EN.4	0.81	−0.61			
RW.5	EN.5	0.83	−0.53			
RW.6	NO.1					
RW.7	NO.2	0.68			-0.32	
RW.8	NO.3	0.78	−0.54			
RW.9	FI.1	0.69			-0.68	
RW.10	FI.2	0.69	−0.48	−0.33		
RW.11	FI.3	0.77	−0.59	−0.32		

Table 3
The ω reliability estimate with 95% confidence interval of the proposed revised four-factor UES subscales from eShopping2.

Subscale	ω	95% ci of ω	Original α (O'Brien, 2008)
FA	0.92	(0.91, 0.93)	0.92
PU	0.92	(0.91, 0.93)	0.91
AE	0.90	(0.88, 0.91)	0.89
RW	0.87	(0.86, 0.89)	

above 0.5 and all but a further two with loadings above 0.6. This single factor accounted for 50% of the variance in the dataset. The single item which did not load was "I continued to shop on this website out of curiosity." This may not have loaded because the data was gathered in an online setting where participants had made a purchase, which may have been their motivation for continuing their shopping activity rather than browsing out of curiosity; this may therefore be an artefact of the study. Even allowing for this, the single factor does give strong support for the idea of an underlying concept of engagement relevant to all of the items in the UES.

An exploratory 6-factor analysis produced only a weak model. Though the FA, PU and AE clearly emerged in this structure as single factors, the remaining components of the UES did not align with single factors and only weakly loaded on any factor (i.e., loadings of around 0.35). Also, the sixth factor consisted of only a single PU item that cross-loaded with the factor on which PU loaded as a whole. This model only accounted for 33% of the variance in the UES data.

A confirmatory bifactor 6-factor model, reflecting the original structure of the UES, did suggest that the original UES factors were coherent within the original data but they only formed a weak relationship with a unified engagement factor. This model only accounted for 40% of the variance in the UES.

Given the reasonable 4-factor model seen in *eShopping1*, an exploratory 4-factor model was also produced for the UES in *eShopping2* (see Table 2). This produced three strong factors corresponding to each of FA, PU and AE in the original with all but one PU item strongly loading on their respective factors. The fourth factor was in part made up of the remaining items though there was some cross-loading with other factors; some items from NO that did not strongly load on any factor. This model accounted for 56% of the variance in the UES items. Thus, it provided a strong model, but not necessarily a strong fourth factor.

The internal reliability on *eShopping2* was evaluated using ω following the guidance (and using the MBESS package) recommended by (Dunn et al., 2014) (see Table 3). For those not familiar with ω, it is worth noting that in this context the Cronbach α values were similar to those reported in (O'Brien, 2008). The Reward sub-scale was more robust than the novelty and felt involvement sub-scales, with Cronbach alpha values of 0.58 and 0.7, respectively; the original endurability factor had an alpha value of 0.86.

Across both *eShopping1* and *eShopping2*, the fourth factor included most of the original EN factor but the other two factors, NO and FI, loaded differently in the two datasets. This might because they are quite context dependent aspects of engagement.

6.2. Devising the short form

The re-analysis of the original two datasets suggested that FA, PU and AE were coherent factors. The remaining items did not strongly emerge from *eShopping1* and *eShopping2* as forming a single factor. At the same time, the confirmatory analysis showed that the original components were each identifiable, albeit weak, factors in the data. Other research, for example, O'Brien and Cairns (2015) also found that the FA, PU and AE components emerged as distinct factors in other contexts, with a fourth factor made up of the other items, and with some degree of cross-loading onto other factors. A closer examination of the remaining items demonstrated some conceptual overlaps. In particular, there were aspects of a valued experiential outcome amongst all items. For this reason, we grouped the remaining items into a Reward factor, labelled RW, which is a single set of items made up of the EN, NO and FI components in the original UES. Thus, our revision of the UES did not change the items, but rather the factor structure.

Table 4
Proposed UES-SF based on the items used in *eShopping2*.

SF	Item Id	Item
FA-S.1	FA.1	I lost myself in this shopping experience
FA-S.2	FA.5	The time I spent shopping just slipped away
FA-S.3	FA.6	I was absorbed in my shopping task
PU-S.1	PU.1	I felt frustrated while visiting this shopping website
PU-S.2	PU.2	I found this shopping website confusing to use
PU-S.3	PU.5	Using this shopping website was taxing
AE-S.1	AE.1	This shopping website is attractive
AE-S.2	AE.2	This shopping website was aesthetically appealing
AE-S.3	AE.4	This shopping website appealed to my senses
RW-S.1	EN.1	Shopping on this website was worthwhile
RW-S.2	EN.4	My shopping experience was rewarding
RW-S.3	NO.3	I felt interested in my shopping task

Table 5
The ω reliability estimate with 95% confidence interval of the proposed UES-SF subscales and their correlation with remainder of the corresponding component.

Subscale	ω	95% ci of ω	Correlation (r) with remainder
FA-S	0.82	(0.74, 0.84)	0.85
PU-S	0.86	(0.83, 0.88)	0.87
AE-S	0.84	(0.82, 0.87)	0.81
RW-S	0.81	(0.78, 0.84)	0.81

In devising the Short Form of the UES (UES-SF), we considered 3 items for each of the constituent factors to give a final set of 12 items. This is sufficiently short to be useful to other researchers without being open to the problems of single item scales (Cairns, 2013). In addition, RW was deliberately refined with a view to giving a more conceptually coherent and robust scale.

In selecting suitable items, the primary consideration was that each selected item reflected the latent construct. Secondly, there when trying to capture a latent construct, it is essential that items represented different manifestations of the latent construct to reliably measure it (Devellis, 2003). For instance, there is an item in RW about the system being "fun" to use, but fun is not always indicative of engagement. Based on these considerations, we proposed the UES-SF (Table 4).

To check that the items produced relevant subscales, the internal reliability on *eShopping2* was evaluated using ω. Also, the items were correlated with the remainder of their original factors to see if the subscales captured the variance in the remaining items of the original components. The results are summarised in Table 5.

As can be seen, each subscale showed good internal reliability. Furthermore, each subscale highly correlates with the remaining items in the scale. This suggested that each subscale was suitable to accurately represent the value of the overall components from which the subscales were derived. Accordingly, these items were defined to be the UES-SF and were taken forward for evaluation.

7. Validation with *SBS*

7.1. Social book search data

The *Social Book Search (SBS)* UES data was collected over three years as part of the CLEF (Conference and Labs of the Evaluation Forum) interactive Social Book Search tasks[1] (see Table 6 for an overview). The aim of this series of experiments was to investigate how people use both

[1] In year 1 the task was part of the INEX (Initiative for the Evaluation of XML retrieval) lab, while in years 2 and 3 it was part of the Social Book Search lab.

Table 6
Overview over the three years interactive Social Book Search teams, participants, interfaces (FS - Faceted Search, MS - Multi-Stage), tasks (NG - Non-Goal, GO - Goal-oriented), and task/interface structure. In *Y1* the Latin-square was between participants for the interfaces and within participants for the tasks.

Y	Teams	Participants	Interfaces	Tasks	Structure
1	4	41	FS or MS	NG & GO	Latin-Square
2	7	192	MS	NG & GO	Random order
3	7	111	MS	NG or GO	Random choice

professional and user-generated meta-data in a range of book-search tasks (Bellot et al., 2014; Koolen et al., 2016; 2015). The three experiments used a data-set consisting of approximately 1.5 million books combining both professional and user-generated content from Amazon, user-generated content from LibraryThing, and professional meta-data from the British Library and Library of Congress.

The SBS experiments used a web-based system with a standard experimental structure across all three years (Hall and Toms, 2013). The system initially acquired background information about participants, then presented the task(s), and finally delivered the UES as a post-experiment section. The UES was displayed as a continuous, randomized list of questions on a single-page, though a fixed order of the UES was kept through all three years.

In all three years both a non-goal and a goal-oriented task were used. The non-goal task was the same across all three years:

> Imagine you are waiting to meet a friend in a coffee shop or pub or the airport or your office. While waiting, you come across this website and explore it looking for any book that you find interesting, or engaging or relevant. Explore anything you wish until you are completely and utterly bored. When you find something interesting, add it to the book-bag.[2] Please add a note (in the book-bag) explaining why you selected each of the items.

The goal-oriented task was changed between *Y1* and *Y2*. In *Y1* the task was

> Imagine you are looking for some interesting physics and mathematics books for a layperson. You have heard about the Feynman books but you have never really read anything in this area. You would also like to find an "interesting facts" sort of book on mathematics.

As this led to generally very short interactions, for *Y2* and *Y3* a longer goal-oriented task was used:

> Imagine you participate in an experiment at a desert-island for one month. There will be no people, no TV, radio or other distraction. The only things you are allowed to take with you are 5 books. Please search for and add 5 books to your book-bag that you would want to read during your stay at the desert-island:
>
> - Select one book about surviving on a desert island.
> - Select one book that will teach you something new.
> - Select one book about one of your personal hobbies or interests.
> - Select one book that is highly recommended by other users (based on user ratings and reviews).
> - Select one book for fun.
>
> Please add a note (in the book-bag) explaining why you selected each of the five books.

In year 1, *Y1*, two interfaces were used, one a standard faceted search interface and the second a multi-stage search interface based on the search stages identified in Vakkari (2001); in years 2 and 3, *Y2* and *Y3*, only the multi-stage interface was used. The number of tasks and interfaces each participant saw varied across the three years. *Y1* used a Latin-square structure where participants either used the standard or multi-stage interface (between participants) and undertook both the non-goal and goal-oriented tasks (within participants). In *Y2* all participants used the multi-stage interface and undertook both tasks in a randomly assigned order. Finally in *Y3* participants used the multi-stage interface with either the non-goal or goal-oriented task. Additionally in *Y3* participants could optionally undertake one additional task.

The Social Book Search lab was run as a shared evaluation task, where participating teams contributed participants to a shared pool for each year's experiment in order to create an experiment with a wider range of participants. Teams received the data for all participants. As a result the number of participating teams, which teams participated, and the number of participants recruited varies across the three years. In all three years, the full version of the UES (31 items) was administered to all participants after completing their tasks and so was a summative measure of engagement with the interaction.

7.2. SBS Project data validation and pre-processing

All three years of data represent useful and distinct datasets suitable for validating the UES, but two of the sets *Y1* and *Y3* are relatively small. Even for evaluating the 12 items of the UES-SF, it is still the case that larger datasets are better and therefore it was ideal to merge the data across the three years. However, this was done with caution because systematic differences in UES responses on between years could produce artefacts in the merged data and interfere with validation. In particular, the correlations used to validate the UES-SF against the full scale could be susceptible to the clustering of the UES data into year groupings.

To mitigate against this, we looked first at the location of each item in each dataset. That is, for each item in a given year, the mean score for that item was calculated. To facilitate comparison, the mean scores were plotted as a boxplot, see Fig. 1. As can be seen, mean items scores varied between 1 and 2.5 across all years with most means between 1.5 and 2.0. This suggests that across years, the item means were falling in a similar range. Standard deviations were not similarly evaluated as typically there is little variation in standard deviations from Likert scale items.

Though the location of all items is approximately equal, there may be systematic differences between particular items. To see this, the differences in means for each item was considered for all pairs of years and similarly plotted, see Fig. 2. In almost all cases, the means of most items from all years are within 0.5 of each other with only three items in Year 1 having a mean about 0.6 lower than the corresponding mean in *Y2*.

In order to see if the items as a whole behaved coherently, the mean scores were correlated across all years. The expectation would be that items that score low in one year would tend to score low in other years and, similarly, items that scored highly would also do so across all years. As anticipated, the correlations of item means between all three years were all above $r = 0.8$ and scatterplots showed a strong linear relationship. This indicated that the items were behaving similarly across the three years of using the UES and provided confidence that the three datasets could be merged for validating the UES-SF.

7.3. Validation of the UES and UES-SF with SBS data

The first step to validation was to ensure that the full UES was performing in a way expected given the previous analysis. An exploratory 1-factor analysis (see Table 7) gave a good single factor that accounted for 38% of the variance in the UES data, comparable to *eShopping1* but lower than *eShopping2*. However, it should also be noted that PU.5 and PU.6 failed to load on this factor and PU.8 only loaded weakly (0.24). These items relate to participants' feelings of being able to do the task given to them. It may be that in this study, the task was not too demanding and so did not come to bear upon engagement.

A confirmatory 4-factor analysis was conducted using the factor structure of FA, PU, AE and RW as proposed in the previous section

[2] The bookbag was a specific system feature for storing selections for later review and reference; similar to the "shopping cart" used by e-commerce sites.

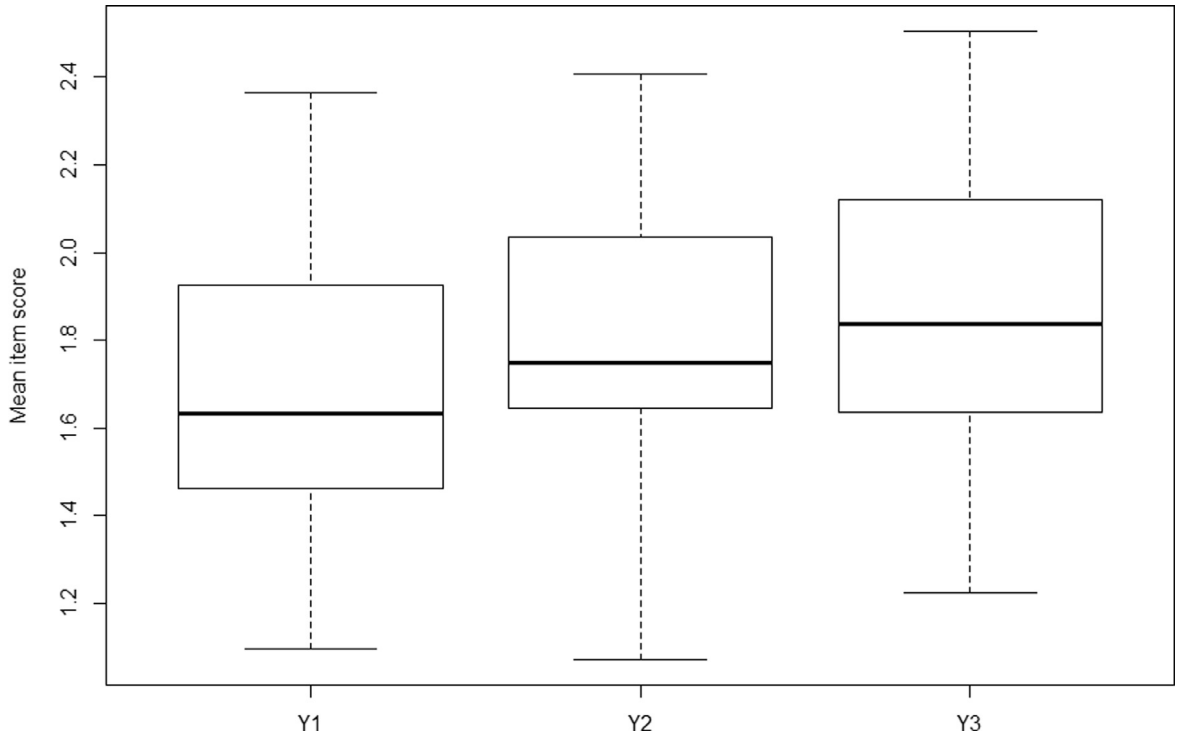

Fig. 1. The boxplot of means for each UES item in each year of the SBS data.

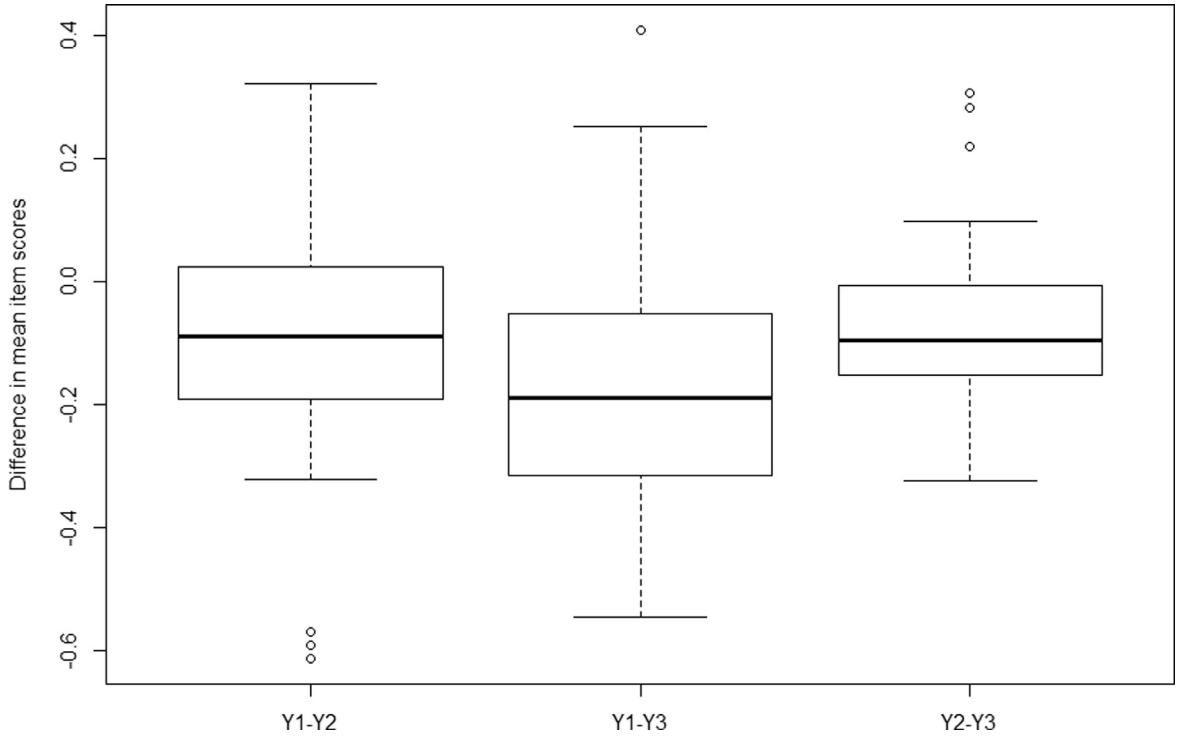

Fig. 2. The boxplot of differences in mean for each UES item between each year of the SBS data.

(see Table 7). This showed an underlying single factor for all items with all but two items loading on the full items scale (loading less than 0.2). The two items that did not load were PU.5 and PU.6, as seen in single dimension exploratory analysis. The individual factors for each component showed good loadings for the appropriate items except for two items in the RW factor. The items that did not load well on the RW factor were EN.3 and EN.5 though EN.5 had a weak loading of 0.24. This may be an artefact of the study because EN.3 is about the task not working out as planned, and it is possible that, in this context, participants did not have particular plans for the interaction. The other item is about recommending the system to others, given the limited availability of the experimental system, people did not see how they could recommend it to others to use.

Table 7
Factor loadings of the models with 1 factor and bifactor with 4 specific factors for the *SBS* dataset. Loadings of magnitude less than 0.3 are omitted for clarity and note that the bifactor model is confirmatory so FA, PU, AE and RW are constrained to only load on their corresponding factors.

Scale Name	One F1	Bifactor general	fa	pu	ae	rw
FA.1	0.45	0.26	0.71			
FA.2	0.61	0.37	0.70			
FA.3	0.47	0.29	0.43			
FA.4	0.60	0.44	0.59			
FA.5	0.56	0.21	0.67			
FA.6	0.67	0.42	0.39			
FA.7	0.44	0.26	0.42			
PU.1	−0.55	−0.52		0.58		
PU.2	−0.42	−0.39		0.55		
PU.3	−0.68	−0.58		0.46		
PU.4	−0.59	−0.48		0.60		
PU.5				0.63		
PU.6				0.55		
PU.7	−0.39	−0.35		0.34		
PU.8	−0.24	−0.27		0.46		
AE.1	0.68	0.56			0.56	
AE.2	0.57	0.47			0.69	
AE.3	0.62	0.45			0.54	
AE.4	0.72	0.56			0.61	
AE.5	0.63	0.48			0.59	
RW.1	0.88	0.71				0.40
RW.2	0.71	0.54				0.39
RW.3	0.45	0.36				
RW.4	0.70	0.51				0.39
RW.5	0.79	0.64				0.24
RW.6	0.57	0.36				0.35
RW.7	0.74	0.54				0.42
RW.8	0.73	0.51				0.53
RW.9	0.75	0.54				0.43
RW.10	0.64	0.47				0.45
RW.11	0.84	0.62				0.42

Table 8
The ω reliability estimate with 95% confidence interval of the UES-SF subscales and their correlation with remainder of the corresponding component using dataset *MH*.

Subscale	ω	95% ci of ω	Correlation with remainder (r)
FA-S	0.75	(0.69, 0.79)	0.69
PU-S	0.70	(0.63, 0.75)	0.80
AE-S	0.88	(0.85, 0.90)	0.85
RW-S	0.79	(0.74, 0.82)	0.82

The overall reliability of the bifactor model was $\omega = 0.88$; the model captures about 88% of the variance in the whole dataset. The general factor alone having $\omega_H = 0.45$ accounted for 45% of the variance in the data and thus just over half of the overall variance in the model.

Overall, the bifactor analysis suggested that there was a good single notion of engagement underlying all of the original UES items but also that there were specific uncorrelated effects captured by FA, PU, AE and RW. Thus, each of these separate factors contributed distinct insight into different aspects of engagement.

Having confirmed that the UES matched the expectations of how it should behave on the new dataset, we turned our consideration to the statistical validity of the UES-SF. The same reliability analysis was conducted as was done on dataset *eShopping2*. The results of this analysis are given in Table 8.

This analysis showed that on the previously unseen dataset *SBS*, the UES-SF was effective. Each subscale correlated well with the remaining items in the components of the UES. The subscale for FA correlated lower with its remaining FA items than seen in the *eShopping2* dataset, but still with a strong correlation.

Reliabilities were generally high, as to be hoped for in the short form of a questionnaire. The reliability of PU-S is notably lower than in the previous dataset. This may relate to the fact that some items of PU, including PU.5 which is in PU-S, did not load well on the general engagement factor due to the possible lack of contextual fit. Further it should be noted that internal reliabilities reflect the variation in data due to contextual effects such as the system used to gather the data.

Overall, the UES-SF was statistically reliable and relevant in estimating the full UES scores in a novel context. This holds promise for it being an effective short form of the UES to be used across a range of studies.

8. Guide to use of UES and UES-SF

8.1. General considerations

Questionnaires, in addition to being rigorously evaluated, must also be 'fit for purpose'. In other words, "a scale will be useful only if it fits the researcher's conceptualization of the variable of interest" (Devellis, 2003) (p. 187). In addition, questionnaires must be suited for populations of interest in terms of literacy levels or other developmental characteristics (Devellis, 2003). The UES has been developed and tested with Western adult populations in studies that evaluate digital technologies; it may not be effective in, for example, studies with children or non-digital technologies, such as print books.

If researchers elect to use the UES or UES-SF, then further decisions must be made regarding how it will be utilized. As we have demonstrated in our analysis, both the full UES and UES-SF are sufficiently robust for implementation; the short form may be more ideal in within-subject studies where participants are completing multiple tasks or trials, or comparing two or more HCI applications. Such study designs involve repetition and can be fatiguing, even when conditions/applications are counterbalanced. The length of time needed to complete the questionnaires will vary depending on the reading ability of the sample, but, in general, participants should be able to complete the UES in less than 15 minutes and the UES-SF in 5–10 minutes. Researchers may be interested in using only subscales of the UES, and this is encouraged provided that it is clear that engagement as a holistic construct is not being measured.

We have modified the wording of the UES slightly to enable its use in different HCI settings. (We have included both the UES and UES-SF as appendices to show how the items can be customized for a particular context.) However, it must be remembered that changing the wording to a great extent will nullify existing work that shows the UES to be a reliable and valid questionnaire; even altering the five-point rating scale has implications for comparing findings across multiple research studies (Kelly and Sugimoto, 2013; O'Brien and McCay-Peet, 2017). Adding new items to the UES would also be problematic, as these new additions would have "face validity only" (Kazdin, 2016) (p. 220) and would need to be examined on their own and in relation to other UES items for reliability. Researchers interested in translating the UES into a language other than English should determine if the meaning of the questions is appropriate for their population of interest; for example, whether the language of the items "make sense" in a non-English or non-North American context. In addition, rigorous evaluation of the reliability and validity would need to be undertaken upon data collection to establish the robustness of the translated tool.

When it comes to administering the UES (or any other questionnaire) there are two further considerations: mode of administration and contingent variables. Firstly, the mode in which people respond to self-report measures can affect responses. An excellent example of this is the work of Kelly and colleagues, who compared responses to questions asked orally, via pen and paper or via a computer interface. They demonstrated that there were no differences in the number of unique ideas conveyed in the responses, even though the people who responded orally provided longer answers; however, those who completed the computer-based administration of the questionnaire rated their attitudes and experiences

more favourably (Kelly et al., 2008). The UES has only been administered by the authors electronically, and lower mean scores for UES subscales could be observed for pen and paper administrations; this would be an interesting study to conduct.

Secondly, contingent variables are those contextual artefacts of a study, such as participant motivation, fatigue or response style, which impact/are impacted by what tasks and measures are used in a study and their order (Devellis, 2003). DeVellis cites an example of asking people to rate their mood or to rate their satisfaction with their possessions (e.g., car, house) before asking them about their aspirations, where the former measures induced a mood or frame of reference that influenced how they rated their aspirations. Factors such as mood and "cognitive sets" can negatively affect scale reliability and validity and alter the relationships between items (Devellis, 2003) (p. 190). Diligent study design, i.e., counterbalancing, randomizing, and so on, can mitigate contingent variables to some extent – or at least lessen their impact. However, researchers must consider the placement of the UES in their overall procedure, understanding that the order in which people are asked about their experiences, e.g., immediately following an interaction with a system versus after completing a knowledge retention test, will influence the results.

In terms of analyzing the UES or UES-SF, researchers may elect to perform factor analysis to verify the four dimensions. The use of multivariate statistics, however, must take the sample size into account, e.g., the ratio of participants to questionnaire items (Tabachnick and Fidell, 2013) and that factor analysis or principle components analysis (PCA) can be run (e.g., type of rotation used) and interpreted (e.g., structure vs. path matrix) in different ways (O'Brien and McCay-Peet, 2017). Further, exploratory factor analysis is a reductionist technique intended to reduce scale data, whereas confirmatory factor analysis is intended to show the reproducibility of the factor structure compared with the original. Therefore, exploratory factor analysis may not be an appropriate approach for analyzing the UES, especially given that we have now developed and validated a short form and that reducing the number of items within subscales may threaten reliability. Confirmatory factor analysis may be effective for assessing the robustness of the UES in general, and specifically in new settings, with new populations, or with different HCI applications.

8.2. Scoring the UES and UES-SF

In Appendices A and B we provide instructions for scoring the UES and UES-SF. The UES has different numbers of items on each of its subscales, while there are three items on each dimension of the UES-SF. With multidimensional questionnaires, such as the UES, it is not appropriate to discount the distinct dimensions. This is especially true when questionnaires contain different numbers of items on each of their subscales, because this weights the overall score toward the subscales with more items. We recommend that scores be calculated as means for each subscale; this approach has been used by the developers of other multidimensional questionnaires (Schutte and Malouff, 2007). For example, the Focused Attention subscale of the long form UES contains seven items; responses to all seven items would be added together and the sum divided by seven. This process can be followed for both the revised UES and UES-SF.

To calculate an overall engagement score, the average of each of the four subscales of the UES long form should be summed (Appendix A); in the case of the UES-SF, all items can be added together and the sum divided by twelve as all SF subscales contain three items and therefore weighting is not a concern (Appendix B). If the UES/UES-SF is completed more than once in a study, then scores for each iteration should be calculated and examined independently. These recommendations for scoring allow researchers to explore differences across participants' or differences in individuals' experiences with multiple tasks, interfaces, and so on.

8.3. Reporting findings from the administration of the UES and UES-SF

In addition to being an essential output in a research paper, reporting the findings of questionnaire administration in detail is key for questionnaire developers and other researchers. This is vital information for anyone interested in evaluating the performance of the measure across samples, studies, or, in our case, HCI applications, determining whether to adopt in their own work, or deciding whether improvements are needed (Kazdin, 2016; Kelly and Sugimoto, 2013). Reporting involves articulating what measures were utilized (or what components of whole measures) and whether modifications were made to the wording and scale, and reflecting on how these and other aspects of the study design may or may not have impacted responses. At a minimum, how the data was prepared for analysis (e.g., the presence and treatment of missing values) and basic descriptive information should be reported.

It is also the role of the researcher

> "to *think* about one's findings. Especially if the results appear strongly counterintuitive or countertheoretical, the researcher must consider the possibility that the scale is invalid in the context of that particular study (if not more broadly). It may be that the extent to which the validity of the scale generalizes across populations, settings, specific details of administration, or an assortment of other dimensions is limited (Devellis, 2003) (p. 191)".

This means that researchers need to examine the analysis in relation to the research questions, other variables measured, and the study's outcomes. An interesting example of this comes from a study conducted by Warnock and Lalmas (2015) where user engagement with two versions of a web interface was tested. The versions contained the same content, but aesthetic elements (colour, font, ads) were varied, such that one was aesthetically appealing and the other was not. The aesthetic appeal subscale failed to detect differences between these two interfaces, which was countertheoretical; the authors rightfully questioned the validity of the aesthetic appeal items. Upon closer examination of the study, however, we note that Amazon Mechanical Turk (AMT) workers recruited for the study were geographically based in the United States, while the two websites chosen for the experiment were Wikipedia and BBC.com, a UK based media source. Participants found Wikipedia *less* aesthetically appealing than BBC, which may speak to the fact that Wikipedia was more familiar and participants could therefore see that its aesthetic conventions were violated in the study. Unfortunately the paper did not report the means for the self-report scales used across websites and interface conditions, but previous work has noted the affect of familiarity on user engagement in online news settings based on the geography of participants (O'Brien and Cairns, 2015).

The ability to connect findings across studies to evaluate the robustness of a questionnaire should not be inflated with generalizability:

> The value and role of self-reports in the assessment of any specific construct is best evaluated on a domain-by-domain and study-by-study basis. Universal recommendations, even within a specific area, are unwise because various contextual factors can influence the validity and utility of self-reports. Thus, it is incumbent upon the users of self-report instruments to document that their particular measure behaves as it should, given current theories about the construct it measures (Krueger and Kling, 2000) (p. 222).

In the case of self-report measures, individuals may respond differently at different points in time, and individual difference variables, such as gender, can influence how people respond to an item (Devellis, 2003; Guttman, 1944). In the case of the UES, we are not seeking to collect scores across applications for the purpose of norming, or emphatically stating what "low" or "high" engagement should look like. Rather, it is up to the researcher to determine what an adequate level of engagement with a particular application should look like. Given that the UES is multidimensional, there may be circumstances where it is more cru-

cial for a specific dimension to be high. In past work we have created low, medium and high categories within the data set by examining the mean or median scores of the sample depending on whethe data are normally distributed. This enabled us to examine variations in engagement for participants in a specific context completing the same tasks and using the same computer-mediated system, and to observe how the range of scores varied for each subscale. For example, in one study, focused attention had a lower mean than perceived usability, and this indicated that overall focused attention was not affected by the usability of the system (O'Brien and Lebow, 2013).

9. Conclusion

One of the aims of this work was to confirm the observations of several studies that have questioned the six original factors of the User Engagement Scale (UES) (O'Brien, 2008; O'Brien and Toms, 2008), pointing instead to a four-factor structure (Banhawi and Ali, 2011; O'Brien and Cairns, 2015; O'Brien and Toms, 2013; Wiebe et al., 2014). We reanalyzed data collected in the development of the original UES (O'Brien, 2008; O'Brien and Toms, 2008) using state-of-the-art statistical techniques that were previously not widely accessible. Although the original six factors had some explanatory power, we confirmed a four-factor model. We proposed a revised UES where the items did not change, but how they were grouped as subscales was altered. The new dimensions of the UES are now aesthetic appeal, focused attention, perceived usability, and reward.

A second aim was to devise a short form of the UES (UES-SF), which we accomplished. This twelve item measure still captures the core concepts represented in the full form. Both the revised full-form and UES-SF were validated with novel data from the SBS project. In future, this work could be extended through experimental work that is tailored to manipulations of specific components of engagement; for instance, only aesthetics or only the reward component. While this may be slightly removed from the engineering of information systems, it would be a means to establish the value of the UES as a measure with wide applicability and relevance.

The final aim of this research was to guide researchers in the effective administration and scoring of the UES and UES-SF. We raised general considerations around the adoption of these measures in different research settings or with different populations, and discussed scoring and reporting procedures. It is our hope that the brevity of the UES-SF will encourage uptake of the entire questionnaire, and that this paper sheds light on the importance of considering reliability and validity in the use of self-report questionnaires to strengthen the overall findings of individual studies, and comparability across different studies.

Appendix A. User engagement scale long form: questionnaire items and instructions for scoring

Instructions for administrators: When administering the UES and UES-SF, all items should be randomized and dimension identifiers (e.g., "Focused Attention or FA") should not be visible to participants. Below we provide general instructions to participants than can be modified to suit the study context; the five-point rating scale should be used to allow for comparisons across studies/sampled populations. The wording of the questions may be modified to your context of use. For example, item PU.1 "I felt frustrated while using this Application X" may be reworded to "I felt frustrated while using this search engine."

Instructions for respondents: The following statements ask you to reflect on your experience of engaging with Application X or "this study". For each statement, please use the following scale to indicate what is most true for you.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1	2	3	4	5

User Engagement Scale Long Form (UES-LF).

FA.1	I lost myself in this experience.
FA.2	I was so involved in this experience that I lost track of time.
FA.3	I blocked out things around me when I was using Application X.
FA.4	When I was using Application X, I lost track of the world around me.
FA.5	The time I spent using Application X just slipped away.
FA.6	I was absorbed in this experience.
FA.7	During this experience I let myself go.
PU.1	I felt frustrated while using this Application X.
PU.2	I found this Application X confusing to use.
PU.3	I felt annoyed while using Application X.
PU.4	I felt discouraged while using this Application X.
PU.5	Using this Application X was taxing
PU.6	This experience was demanding.
PU.7	I felt in control while using this Application X.
PU.8	I could not do some of the things I needed to do while using Application X.
AE.1	This Application X was attractive
AE.2	This Application X was aesthestically appealing
AE.3	I liked the graphics and images of Application X.
AE.4	Application X appealed to be visual senses.
AE.5	The screen layout of Application X was visually pleasing.
RW.1	Using Application X was worthwhile
RW.2	I consider my experience a success.
RW.3	This experience did not work out the way I had planned.
RW.4	My experience was rewarding.
RW.5	I would recommend Application X to my family and friends
RW.6	I continued to use Application X out of curiosity.
RW.7	The content of Application X incited my curiosity.
RW.8	I was really drawn into this experience.
RW.9	I felt involved in this experience.
RW.10	This experience was fun.

A1. Scoring the UES-LF

Instructions for administrators: When administering the UES and UES-SF, all items should be randomized and dimension identifiers (e.g., "Focused Attention or FA") should not be visible to participants. Below we provide general instructions to participants than can be modified to suit the study context; the five-point rating scale should be used to allow for comparisons across studies/sampled populations. The wording of the questions may be modified for one's context of use. For example, item PU.1 "I felt frustrated while using this Application X" may be reworded to "I felt frustrated while using this search engine."

Instructions for respondents: The following statements ask you to reflect on your experience of engaging with Application X or "this study". For each statement, please use the following scale to indicate what is most true for you.

1. Reverse code the following items: PU-1, PU-2, PU-3, PU-4, PU-5, PU-6, PU-8, and RW-3.
2. Scale scores are calculated for each participant by summing scores for the items in each of the four subscales and dividing by the number of items:
   - Sum FA-1, FA2, ... FA7 and divide by seven.
   - Sum PU-1, PU-2, ... PU-8 and divide by eight.
   - Sum AE-1, AE-2, AE-3, AE-4, and AE-5 and divide by five.
   - Sum RW-1, RW-2, ... RW-10 and divide by ten.
3. If participants have completed the UES more than once as part of the same experiment, calculate separate scores for each iteration. This will enable the researcher to compare engagement within participants and between tasks/iterations.
4. An overall engagement score can be calculated by adding the average of each subscale as per #2.

Appendix B. User engagement scale short form: questionnaire items and instructions for scoring

FA-S.1	I lost myself in this experience.
FA-S.2	The time I spent using Application X just slipped away.
FA-S.3	I was absorbed in this experience.
PU-S.1	I felt frustrated while using this Application X.
PU-S.2	I found this Application X confusing to use.
PU-S.3	Using this Application X was taxing.
AE-S.1	This Application X was attractive.
AE-S.2	This Application X was aesthetically appealing.
AE-S.3	This Application X appealed to my senses.
RW-S.1	Using Application X was worthwhile.
RW-S.2	My experience was rewarding.
RW-S.3	I felt interested in this experience.

B1. Scoring the UES-SF

- Reverse code the following items: PU-S1, PU-S2, PU-S3.
- If participants have completed the UES more than once as part of the same experiment, calculate separate scores for each iteration. This will enable the researcher to compare engagement within participants and between tasks/iterations.
- Scores for each of the four subscales can be calculated by adding the values of responses for the three items contained in each subscale and dividing by three. For example, "Aesthetic Appeal" would be calculated by adding AE-S1, AE-S2, and AE-S3 and dividing by three.
- An overall engagement score can be calculated by adding all of the items together and dividing by twelve.

References

Arapakis, I., Lalmas, M., Cambazoglu, B.B., Marcos, M.-C., Jose, J.M., 2014. User engagement in online news: under the scope of sentiment, interest, affect, and gaze. J. Assoc. Inf. Sci. Technol. 65 (10), 1988–2005.

Arguello, J., Wu, W.-C., Kelly, D., Edwards, A., 2012. Task complexity, vertical display and user interaction in aggregated search. In: Proceedings of the Thirty-Fifth International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 435–444.

Banhawi, F., Ali, N.M., 2011. Measuring user engagement attributes in social networking application. In: Proceedings of the International Conference on Semantic Technology and Information Retrieval (STAIR). IEEE, pp. 297–301.

Bellot, P., Bogers, T., Geva, S., Hall, M.M., Huurdeman, H., Kamps, J., Kazai, G., Koolen, M., Moriceau, V., Mothe, J., Preminger, M., SanJuan, E., Schenkel, R., Skov, M., Tannier, X., Walsh, D., 2014. Overview of INEX 2014. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M.M., Hanbury, A., Toms, E. (Eds.), Information Access Evaluation. Multilinguality, Multimodality, and Interaction. In: Lecture Notes in Computer Science, 8685. Springer International Publishing, pp. 212–228. doi:10.1007/978-3-319-11382-1_19.

Cairns, P., 2013. A commentary on short questionnaires for assessing usability. Interact. Comput. 25 (4), 312–316.

Chalmers, R.P., et al., 2012. Mirt: a multidimensional item response theory package for the R environment. J. Stat. Softw. 48 (6), 1–29.

Devellis, R.F., 2003. Scale Development: Theory and Applications (Applied Social Research Methods), second ed. Sage Publications, Inc.

Dunn, T.J., Baguley, T., Brunsden, V., 2014. From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. Br. J. Psychol. 105 (3), 399–412.

Edwards, A., 2015. Engaged or Frustrated? Disambiguating Engagement and Frustration in Search. University of North Caroline at Chapel Hill Ph.D. thesis.

Embretson, S.E., Reise, S.P., 2013. Item Response Theory. Psychology Press.

Fredricks, J.A., Blumenfeld, P.C., Paris, A.H., 2004. School engagement: potential of the concept, state of the evidence. Rev. Educ. Res. 74 (1), 59–109.

Galbraith, J., Moustaki, I., Bartholomew, D.J., Steele, F., 2002. The Analysis and Interpretation of Multivariate Data for Social Scientists. CRC Press.

Grafsgaard, J.F., 2014. Multimodal Affect Modeling in Task-Oriented Tutorial Dialogue. North Carolina State University Ph.D. thesis.

Guttman, L., 1944. A basis for scaling qualitative data. Am. Sociol. Rev. 9 (2), 139–150.

Hall, M.M., Toms, E., 2013. Building a common framework for iir evaluation. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, Berlin, Heidelberg, pp. 17–28.

Jacques, R.D., 1996. The Nature of Engagement and its Role in Hypermedia Evaluation and Design. South Bank University Ph.D. thesis.

Kazdin, A.E., 2016. Selecting Measures for Research Investigations. American Psychological Association.

Kelly, D., Harper, D.J., Landau, B., 2008. Questionnaire mode effects in interactive information retrieval experiments. Inf. Process. Manag. 44 (1), 122–141.

Kelly, D., Sugimoto, C.R., 2013. A systematic review of interactive information retrieval evaluation studies, 1967–2006. J. Am. Soc. Inf. Sci. Technol. 64 (4), 745–770.

Koolen, M., Bogers, T., Gäde, M., Hall, M., Hendrickx, I., Huurdeman, H., Kamps, J., Skov, M., Verberne, S., Walsh, D., 2016. Overview of the CLEF 2016 Social Book Search Lab. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer International Publishing, pp. 351–370.

Koolen, M., Bogers, T., Gäde, M., Hall, M.M., Huurdeman, H., Kamps, J., Skov, M., Toms, E., Walsh, D., 2015. Overview of the CLEF 2015 Social Book Search Lab. In: Mothe, J., Savoy, J., Kamps, J., Pinel-Sauvagnat, K., Jones, G.J.F., San Juan, E., Capellato, L., Ferro, N. (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. In: Lecture Notes in Computer Science, 9283. Springer International Publishing, pp. 545–564. doi:10.1007/978-3-319-24027-5_51.

Krueger, R.F., Kling, K.C., 2000. Self-report. In: Kazdin, A.E. (Ed.), Encyclopedia of psychology, 7. Oxford University Press, pp. 220–224.

Lalmas, M., O'Brien, H., Yom-Tov, E., 2014. Measuring user engagement. Synth. Lect. Inf. Concepts Retr. Serv. 6 (4), 1–132.

O'Brien, H., Cairns, P., 2015. An empirical evaluation of the user engagement scale (UES) in online news environments. Inf. Process. Manag. 51 (4), 413–427.

O'Brien, H., Cairns, P., 2016. Why Engagement Matters: Cross-Disciplinary Perspectives and Innovations on User Engagement with Digital Media. Springer Publishing Company, Incorporated.

O'Brien, H.L., 2008. Defining and Measuring Engagement in User Experiences with Technology. Dalhousie University Ph.D. thesis.

O'Brien, H.L., 2016a. Theoretical perspectives on user engagement. In: Why Engagement Matters: Cross-Disciplinary Perspectives and Innovations on User Engagement with Digital Media. Springer, pp. 1–26.

O'Brien, H.L., 2016b. Translating theory into methodological practice. In: Why Engagement Matters: Cross-Disciplinary Perspectives and Innovations on User Engagement with Digital Media. Springer, pp. 27–52.

O'Brien, H.L., 2017. Antecedents and learning outcomes of online news engagement. J. Assoc. Inf. Sci. Technol. 66 (12), 2809–2820.

O'Brien, H.L., Lebow, M., 2013. Mixed-methods approach to measuring user experience in online news interactions. J. Am. Soc. Inf. Sci. Technol. 64 (8), 1543–1556.

O'Brien, H.L., McCay-Peet, L., 2017. Asking good questions: questionnaire design and analysis in interactive information retrieval research. In: Proceedings of the Conference on Conference Human Information Interaction and Retrieval. ACM, pp. 27–36.

O'Brien, H.L., McKay, J., 2016. What makes online news interesting? Personal and situational interest and the effect on behavioral intentions. Proc. Assoc. Inf. Sci. Technol. 53 (1), 1–6.

O'Brien, H.L., Toms, E.G., 2008. What is user engagement? A conceptual framework for defining user engagement with technology. J. Am. Soc. Inf. Sci. Technol. 59 (6), 938–955.

O'Brien, H.L., Toms, E.G., 2010a. The development and evaluation of a survey to measure user engagement. J. Am. Soc. Inf. Sci. Technol. 61 (1), 50–69.

O'Brien, H.L., Toms, E.G., 2010b. Is there a universal instrument for measuring interactive information retrieval?: The case of the user engagement scale. In: Proceedings of the Third Symposium on Information Interaction in Context. ACM, pp. 335–340.

O'Brien, H.L., Toms, E.G., 2013. Examining the generalizability of the user engagement scale (UES) in exploratory search. Inf. Process. Manag. 49 (5), 1092–1107.

Oh, J., Sundar, S.S., 2015. How does interactivity persuade? An experimental test of interactivity on cognitive absorption, elaboration, and attitudes. J. Commun. 65 (2), 213–236.

Reise, S.P., 2012. The rediscovery of bifactor measurement models. Multivar. Behav. Res. 47 (5), 667–696.

Rowe, D.W., Sibert, J., Irwin, D., 1998. Heart rate variability: indicator of user state as an aid to human-computer interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM Press/Addison-Wesley Publishing Co., pp. 480–487.

Schutte, N.S., Malouff, J.M., 2007. Dimensions of reading motivation: development of an adult reading motivation scale. Read. Psychol. 28 (5), 469–489.

Sutcliffe, A., 2009. Designing for user engagement: aesthetic and attractive user interfaces. Synth. Lect. Hum. Cent. Inf. 2 (1), 1–55.

Tabachnick, B., Fidell, L., 2013. Using Multivariate Statistics (6th International Edition. Pearson, Boston, [Mass.]. (cover). ed.).

Taub, M., Mudrick, N.V., Azevedo, R., Millar, G.C., Rowe, J., Lester, J., 2017. Using multichannel data with multi-level modeling to assess in-game performance during gameplay with crystal island. Comput. Hum. Behav 76, 641–655.

Vakkari, P., 2001. A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. J. Doc. 57 (1), 44–60.

Warnock, D., Lalmas, M., 2015. An exploration of cursor tracking data arXiv preprint arXiv:1502.00317.

Webster, J., Ahuja, J.S., 2006. Enhancing the design of web navigation systems: the influence of user disorientation on engagement and performance. Mis. Q. 30 (3), 661–678.

Webster, J., Ho, H., 1997. Audience engagement in multimedia presentations. ACM SIGMIS Datab. 28 (2), 63–77.

Wiebe, E., Sharek, D., 2016. elearning. In: Why Engagement Matters. Springer, pp. 53–79.

Wiebe, E.N., Lamb, A., Hardy, M., Sharek, D., 2014. Measuring engagement in video game-based environments: investigation of the user engagement scale. Comput. Hum. Behav. 32, 123–132.