Received 1 June 2015
Revised 2 October 2015
Accepted 23 October 2015

A study of the use of simulated work task situations in interactive information retrieval evaluations

A meta-evaluation

Pia Borlund
Royal School of Library and Information Science, University of Copenhagen, Aalborg, Denmark

Abstract

Purpose – The purpose of this paper is to report a study of how the test instrument of a simulated work task situation is used in empirical evaluations of interactive information retrieval (IIR) and reported in the research literature. In particular, the author is interested to learn whether the requirements of how to employ simulated work task situations are followed, and whether these requirements call for further highlighting and refinement.

Design/methodology/approach – In order to study how simulated work task situations are used, the research literature in question is identified. This is done partly via citation analysis by use of Web of Science®, and partly by systematic search of online repositories. On this basis, 67 individual publications were identified and they constitute the sample of analysis.

Findings – The analysis reveals a need for clarifications of how to use simulated work task situations in IIR evaluations. In particular, with respect to the design and creation of realistic simulated work task situations. There is a lack of tailoring of the simulated work task situations to the test participants. Likewise, the requirement to include the test participants' personal information needs is neglected. Further, there is a need to add and emphasise a requirement to depict the used simulated work task situations when reporting the IIR studies.

Research limitations/implications – Insight about the use of simulated work task situations has implications for test design of IIR studies and hence the knowledge base generated on the basis of such studies.

Originality/value – Simulated work task situations are widely used in IIR studies, and the present study is the first comprehensive study of the intended and unintended use of this test instrument since its introduction in the late 1990's. The paper addresses the need to carefully design and tailor simulated work task situations to suit the test participants in order to obtain the intended authentic and realistic IIR under study.

Keywords Interactive information retrieval study, IIR study, Test design, Simulated work task situations, Meta-evaluation

Paper type Research paper

1. Introduction

This paper examines how the test instrument of a simulated work task situation is used in the research literature. The major challenge of its use lies in the design of authentic and applicable simulated work task situations, which are relevant and realistic to the test participants who are to apply the situations for reliable interactive information retrieval (IIR). In that light, it is interesting to examine how the test instrument is used in previous IIR evaluation studies, and what we can learn from those evaluations and that use. Hence, the paper reports a meta-evaluation of the use of simulated work task situations. The present paper follows up on a paper by Borlund and Schneider (2010), in which preliminary results were presented.

Journal of Documentation
Vol. 72 No. 3, 2016
pp. 394-413

0022-0418
DOI 10.1108/JD-06-2015-0068

This type of meta-evaluation and further development of approaches for IIR evaluation is motivated by Belkin in his 2008 ECIR Keynote (Belkin, 2008). In the keynote, he particularly addresses the need for alternative evaluation approaches to the Cranfield model, or TREC style as he formulates it (Belkin, 2008, p. 52). This need still stands. Belkin explicitly highlights the IIR evaluation model by Borlund (e.g. Borlund, 2003) as such an attempt (Belkin, 2008, p. 52). He further points out how the contradictions between the necessity for realism and the desire for comparability and generalisation have not yet been solved (Belkin, 2008, p. 52). The issues of realism and comparability are dealt with by the test instrument of a simulated work task situation inherent the IIR evaluation model. In order to compare search behaviours and performance results and hereby generate a reliable knowledge base, the employed simulated work task situations must be realistic to the test participants.

In brief, the concept of a simulated work task situation was introduced in 1997 via a feasibility study of the use of cover-stories functioning as scenarios for user-authentic evaluation of IR effectiveness and user satisfaction with retrieved information (Borlund and Ingwersen, 1997). In 2000 the IIR evaluation model was developed (Borlund, 2000a, 2003); an evaluation model for the evaluation of IR interaction that includes the application of simulated work task situations according to a set of empirically based requirements of how to use this test instrument (Borlund; 2000a, b, 2003).

The overall objectives of the present study are to identify how simulated work task situations are used, and for what types of evaluations they are used. In particular, we want to learn about the intentional, and the unintentional use of simulated work task situations in order to clarify and improve the requirements for the application of simulated work task situations. The idea is to learn from previous research and improve future research. The overall long-term ambition is to increase the knowledge base of empirical IIR studies with refined requirements about how to use simulated work task situations. The insight gained will also help to set out directions for future meta-evaluations of the use of simulated work task situations.

The remainder of the paper is structured as follows: Section 2 introduces the concept and test instrument of a simulated work task situation, and summarises the most basic requirements of how to use the instrument. Section 3 presents the methodological approach taken to examine the use of simulated work task situations as reported in the research literature. Section 4 reports on the results of the study concerning how the evaluation instrument is used in previous evaluation studies. On this basis, directions for future empirical studies that validate and increase our understanding of how to use simulated work task situations are outlined. The paper closes with concluding statements in Section 5.

2. The test instrument of a simulated work task situation

This section introduces the test instrument of a simulated work task situation and the corresponding requirements regarding its use. The theoretical assumptions underlying this instrument are described in Borlund and Ingwersen (1997) and Borlund (2000a, b, 2003).

A simulated work task situation is a short textual description that presents a realistic information requiring situation that motivates the test participant to search the IR system (Borlund, 2003). A simulated work task situation serves two main functions: it causes a "simulated information need" by allowing for user interpretations of the simulated work task situation, leading to cognitively individual information need interpretations as in real life; and it is the platform against which situational relevance

is judged by the test participant (Borlund and Ingwersen, 1997, pp. 227-228). More specifically it helps to describe to the test participants:

- the source of the information need;
- the environment of the situation;
- the problem which has to be solved; and it also
- serves to make the test participants understand the objective of the search (Borlund and Ingwersen, 1997, pp. 227-228).

As such the simulated work task situation is a stable concept, i.e. the given purpose and goal of the IR system interaction. Box 1 depicts a classic example of a simulated work task situation tailored to university students.

Further, by being the same for all the test participants experimental control is provided, and the search interactions are comparable across the group of test participants for the same simulated work task situation. As such, the use of simulated work task situations ensures the IIR study both realism and control.

The issue of realism of the descriptions of the simulated work task situations is essential in order for the prompted search behaviour and relevance assessments of the test participants to be as genuine as intended. The simulated work task situations create simulated information needs that are to replicate genuine information needs. Therefore realism is emphasised in the requirements of how to employ simulated work task situations (Borlund, 2003). In brief, the requirements are as follows:

(1) To tailor the simulated work task situation to the test participants:
   - a situation the test participants can relate to and identify themselves with;
   - a situation the test participants find topically interesting and/or of relevance to them; and
   - a situation that provides enough context in order for the test participants to be able to apply the situation.

(2) To include test participants' personal information needs as baseline.

(3) To rotate the order of simulated work task situation and personal information needs (counterbalancing).

(4) To pilot test prior to actual testing (often more than once).

Requirement no. 1 concerns the tailoring of the simulated work task situations. If the evaluation takes place by the involvement of university students, then the simulated work task situation should describe a situation they can relate to, which they can identify themselves as being in, and present a topic of searching they find interesting. The described situation should be authentic, relevant, and realistic to the university

Box 1. Example of a simulated work task situation (e.g. Borlund, 2003)

Simulated work task situation: after your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.

students – males and females – so that it leads to realistic interpretations and interactions with the simulated information needs. The requirement to tailor the simulated work task situations entails a certain degree of homogeneity of the group of test participants. They need to have something in common, which can form the foundation for the design, tailoring, and use of the simulated work task situations. This further requires a relatively high abstraction level of the wording of the simulated work task situations for them to be applicable by the test participants.

Requirement no. 2 is about how to employ a combination of simulated work task situations and the test participants' personal information needs – both when pilot testing and when carrying out the actual evaluation. This means that the test participants should prepare in advance a personal information need, which they bring with them and search as part of the evaluation. Hence, personal information needs function as a baseline against the simulated information needs. The purpose of this is to serve as a control of the reliability of the search interaction derived from the searching of the simulated work task situations. In addition, the personal information needs provide information about the system's effect on genuine information needs. The inclusion of personal information needs is also useful in the pilot test (requirement no. 4), because personal information needs can inspire to realistic and user-adaptable simulated work task situations.

Requirement no. 3 informs to rotate the order of simulated work task situations and the test participants' personal information needs between the test participants, so that a minimum of test participants search the simulated work task situations and personal information needs in the same order. This is to neutralise order effect on the results in terms of bias of search interaction and relevance assessment behaviour of the test participants.

Requirement no. 4 concerns the ever good test practice of pilot testing prior to actual evaluation. When pilot testing the test setting, the test requisites (e.g. protocol, tutorials, pre-search questionnaire, the simulated work task situations, and post-search interview guide), the test procedure, and the collected data are evaluated, and adjusted accordingly if required. In particular, the suitability of the simulated work task situations and the test participants' perceptions of the simulated work task situations. As mentioned, it is most useful to instruct the pilot-test participants to contribute with real, personal information needs, as these needs can inspire to simulated work task situations that are ideal for the group of test participants. If that is the case, subsequent pilot testing is necessary in order to evaluate the test participants' opinion of the new simulated work task situation(s).

With this introduction to the test instrument of a simulated work task situation and the empirically based requirements, we close Section 2. Section 3 presents how the sample of IIR evaluation studies that use simulated work task situations is identified and analysed.

3. Methodology

In order to investigate how simulated work task situations are used, the research literature in question is identified. This is done partly via citation analysis by use of Web of Science® and partly by systematic search of online repositories of mainly published conference proceedings, e.g. the ACM Digital Library.

The citation analysis was carried out with a citation time window of ten years (1998-2008) on the basis of the six publications authored by Borlund (Borlund and Ingwersen, 1997, 1998, 1999; Borlund, 2000a, b, 2003). A citation window of ten years is a big window that does take into consideration the time consuming process of planning, conducting, data coding, analysing, and reporting of IIR studies. Further, this particular

citation window provides a suitable and manageable sample of 67 papers for illustration and illumination of the most typical intentional and unintentional use of simulated work task situations. The total citation count was 193 for the six publications. On the basis of the citation analysis, 157 individual publications were identified. Another 41 individual publications were identified via the searching of online repositories. In total, 198 individual publications cite one or more of the six publications by Borlund. Paper copies were made of every single publication. One publication is represented only by an abstract in English, as the main text is in Japanese (Sagara, 1999).

The 198 identified papers were divided into two categories of publications either reporting on evaluation studies using simulated work task situations, or not. In total 67 papers report on empirical evaluation by use of simulated work task situations and constitute the sample of our study (Airio, 2008; Arapakis *et al.*, 2008; Bell and Ruthven, 2004; Bierig and Goker, 2006; Blomgren *et al.*, 2004; Brajnik *et al.*, 2002; Craven, 2003; Fachry *et al.*, 2008; Freund and Toms, 2006; Haggerty *et al.*, 2003; Hammer-Aebi *et al.*, 2006; Harper and Kelly, 2006; Harper *et al.*, 2003; Joho and Jose, 2006a, b, 2008; Joho *et al.*, 2008; Jose *et al.*, 1998; Kim, 2007; Kules and Shneiderman, 2008; Langer and Frøkjær, 2008; Larsen and Ingwersen, 2001; Larsen, 2002; Nielsen, 2004; Pansanato and Fortes, 2007; Papaeconomou *et al.*, 2008; Petrelli and Auld, 2008; Petrelli, 2008; Petrelli *et al.*, 2002, 2004, 2006; Pharo and Nordlie, 2005; Pharo, 2008; Price *et al.*, 2007; Qu and Furnas, 2008; Rodden *et al.*, 2001; Ruthven *et al.*, 2002, 2003; Sihvonen and Vakkari, 2004; Skov and Ingwersen, 2008; Suomela and Kekäläinen, 2005a, b; Szlavik *et al.*, 2006a, b, 2007; Tombros *et al.*, 2003, 2005; Toms *et al.*, 2007; Villa *et al.*, 2008; Wen *et al.*, 2006a, b, 2007; White *et al.*, 2001a, b, 2002, 2003, 2005a, b, 2006, 2007, 2008; White and Marchionini, 2006a, b, 2007; White and Ruthven, 2006; Wu *et al.*, 2004; Xu and Yin, 2008; Xu and Wang, 2008; Yuan, 2006; Yuan and Belkin, 2007). The sampled papers are analysed with a particular focus on the methodological aspects of the reported evaluations. More precisely, the 67 papers are read with the aim of answering the following questions and hereby to shed light on the types of evaluations simulated work task situations are used for, and how they are used with respect to the stipulated requirements:

(1) What is evaluated?

(2) Are the simulated work task situations tailored?

(3) Are personal information needs included as baseline?

(4) Are the simulated work task situations rotated?

(5) Is pilot testing carried out to tune simulated work task situations?

The results of the analysis of the 67 papers are presented below in Section 4.

4. Results, discussion, and future work

The five questions are answered one by one. The answer to the first question serves with background information for the further four questions, which concern the previously listed methodological requirements one to four.

4.1 What is evaluated?

In addition to answering what the target of evaluation is in the 67 studies, this sub-section also looks into who the test participants are, how many test participations that are involved, and how many simulated work task situations are used.

The target of evaluation can roughly be divided into three categories: evaluation of systems performance; evaluation of systems facilities and functionalities; and search behaviour. Examples of systems performance studies are evaluation of cross-language retrieval (e.g. Airio, 2008; Petrelli *et al.*, 2002), relevance feedback (e.g. Ruthven *et al.*, 2002), XML retrieval (e.g. Pharo and Nordlie, 2005; Szlavik *et al.*, 2006a, b), image retrieval (e.g. Jose *et al.*, 1998), and citation overlaps (e.g. Larsen, 2002). Examples of evaluation studies of systems facilities and functionalities include thesaurus assisted searching (Sihvonen and Vakkari, 2004), adaptive TV News (Haggerty *et al.*, 2003), retrieval of domain-specific documents (Price *et al.*, 2007), and different types of interfaces developed to support and facilitate IIR (e.g. Brajnik *et al.*, 2002; Harper *et al.*, 2003, Joho and Jose 2006a, b, 2008; Langer and Frøkjær, 2008). The third category concerns studies of search behaviour, for example, the role of emotion in IIR (Arapakis *et al.*, 2008), access to electronic resources by visually impaired people (Craven, 2003), or access to archival material (Fachry *et al.*, 2008), enterprise search behaviour of software engineers (Freund and Toms, 2006), and users' web search tactics (Kules and Shneiderman, 2008). The 67 studies are fairly evenly distributed among the three categories: 33 per cent (22) of the studies are categorised as systems performance studies, 37 per cent (25) as studies of systems facilities and functionalities, and 30 per cent (20) of the studies are associated the category of search behaviour.

The impact of the internet is clearly visible as quite many studies focus on web retrieval, web information access, or use the internet as "search system" (e.g. Airio, 2008; Bell and Ruthven, 2004; Craven, 2003; Haggerty *et al.*, 2003; Joho and Jose, 2008; Kim, 2007; Kules and Shneiderman, 2008; Langer and Frøkjær, 2008; Pansanato and Fortes, 2007; Papaeconomou *et al.*, 2008; Pharo and Nordlie, 2005; Pharo, 2008; Rodden *et al.*, 2001; Skov and Ingwersen, 2008; Tombros *et al.*, 2003, 2005; Villa *et al.*, 2008; White *et al.*, 2001a, b, 2003, 2007, 2008.

In brief, simulated work task situations are used in a variety of IIR studies that all aim to ensure realism in the process of searching for information, to capture the users' interactions with information and system(s), as well as the users' perception and satisfaction with the retrieved information.

The variety in studies is also seen by the variety in groups of test participants. However, the majority of studies (54 per cent) recruit test participants from the academic environment. 22 studies (33 per cent) use university students and 14 studies (21 per cent) report using a combination of university students, research assistants, and faculty staff members. The number could be higher as, for instance Szlavik *et al.* (2006a, b, 2007) describe their test participants as having computer science background. White and colleagues categorise their test participants with respect to search experience only (White *et al.*, 2001a, b, 2002, 2003, 2005, 2006; White and Ruthven, 2006). Petrelli and her colleagues, who evaluate cross language retrieval, characterise their test participants as native language speakers (Petrelli *et al.*, 2002, 2006; Petrelli, 2008). There are five cases (7.5 per cent) where no information is provided about who the test participants are (e.g. Pharo, 2008; Tombros *et al.*, 2003; White *et al.*, 2007). Obviously, information about who the test participants are is crucial in order to assess whether the simulated work task situations are tailored to suit the test participants and hence function as intended. Usually, the intensive use of university students is considered inappropriate and unrepresentative as discussed in Henrich *et al.* (2010), but in this case is it highly recommendable as simulated work task situations, as test instrument is validated only for this particular group of test participants.

A total of 14 studies (18 per cent) represent a large variety of test participants such as software engineers (Freund and Toms, 2006), employees at Microsoft (White *et al.*, 2008),

medical doctors (Larsen and Ingwersen, 2001; Larsen, 2002; Price *et al.*, 2007), scientific staff from the medical industry (Nielsen, 2004), visually impaired and sighted people (Craven, 2003), graphic design professionals (Jose *et al.*, 1998), journalists (Blomgren *et al.*, 2004), food industry educated users (Suomela and Kekäläinen, 2005a, b), conference attendees (Rodden *et al.*, 2001[1]) museum online users (Skov and Ingwersen, 2008), and the study by Petrelli and colleagues that uses 1 business analyst, one journalist, three librarians, and five translators as test participants (Petrelli *et al.*, 2004). One way to compensate for the search behaviour of the groups of test participants not yet being validated with respect to the use of simulated work task situations is to include test participants' personal information needs as baseline.

On average, 26 test participants are involved in a study with a range from 1-133 test participants and with a median value of 24. Also, 24 is the number of test participants used most frequently in the analysed studies, namely 17 times. 3.8 simulated work task situations are searched per test participant per study. But typically more simulated work task situations in different versions are put forward for the test participants to choose from. The different versions aim to, e.g., replicate different levels of task complexity or types of information needs (e.g. Bell and Ruthven, 2004; Craven, 2003; Freund and Toms, 2006; Kim, 2007; Pansanato and Fortes, 2007). In such cases, it is also necessary to include test participants' personal information needs in comparable forms to the different versions used in the study. That is, to control the reliability of the obtained IIR results.

4.2 Are the simulated work task situations tailored?

This sub-section answers the question of whether the used simulated work task situations in the 67 reported studies are tailored to suit the test participants. As a result of dealing with this question this sub-section also looks into what types of simulated work task situations are used, which further raises the question of whether the used simulated work task situations are depicted in the reporting of the studies.

Of the 67 publications, 3 (4.5 per cent) have used tailored simulated work task situations (Larsen and Ingwersen, 2001; Larsen, 2002; Pansanato and Fortes, 2007). Though, they will all claim to have done so, which we will return to with examples later in this sub-section. The two publications by Larsen and Ingwersen (2001) and Larsen (2002) report the same overall (pilot) study with one test participant (medical doctor) and three simulated work task situations. The simulated work task situations are tailored and personalised to reflect the test participant's current research interests into osteoporosis (brittleness of the bones) (Larsen, 2002, p. 166). Obviously, it is easier to tailor and personalise the simulated work task situations in order to obtain genuine and realistic search interaction and relevance assessment (IIR), when only one test participant is involved. In the case with more test participants the trick is to increase the abstraction level of the simulated work task situations to allow for individual interpretation of the information need given the simulated work task situation. Pansanato and Fortes (2007, p. 815) work with a high level of abstraction in their study involving professors as test participants in that they do not specify, e.g., what learning material to what course to search for. They do, however, present their simulated work task situations with a minimum of context usually done to stimulate interest, relevance, and motivate searching. Their simulated work task situations are barely one-liner descriptions that further take the form of replicating different types of information needs (Pansanato and Fortes, 2007, p. 815).

In our sample, the attempts to tailor exist, but are for various reasons not convincing in their current forms. Let us recall the purpose of tailoring of the simulated work task

situations and hereafter consider some examples from the sample. The purpose of tailoring of simulated work task situations is to create text descriptions that trigger (simulated) information needs, which may as well have been formulated by the test participants themselves. The simulated information needs should be so realistic and of personal interest/relevance to the test participant that the simulated information needs can replace the test participants' own personal information needs. Otherwise the given search interaction and relevance assessment (IIR) is not reliable from the perspective of being realistic. We refer to a simulated work task situation as "simulated" because it is imposed the test participants by the evaluators in the form of simulated work task situations, hence ensuring experimental control and comparability of search behaviour and performance at the level of the searched simulated work situation. Simulated work task situations do not aim to be trivial imitations, but aim to be realistic simulations of information needs.

There are examples where university students are to imagine themselves being either members of an information broker company (Joho *et al.*, 2008, p. 96), or to imagine how they are to prepare for a talk on the subjects "tea" or "everyday drinks for old people" – a talk the test participants further are to imagine to give in two weeks' time at the neighbourhood library to an audience of ordinary people (Qu and Furnas, 2008, pp. 540-541). Though they are not information brokers or about to give a talk in due time. These examples are hardly realistic, given the test participants' personal situations as, e.g., computer science students, engineer students, or information science students. These examples depend on imagination rather than identifying situations from the test participants' point of view. Thus, these studies do not provide the optimal conditions for realistic IIR. Another study asks the test participants (mostly university students) to imagine they work as a researcher for a television company with the task to summarise and review major news stories, and in another to imagine they are studying part-time for a media studies degree at the Open University (Villa *et al.*, 2008, p. 498). Obviously, this is not optimal tailoring of the simulated work task situations to the test participants. In addition, non-tailored simulated work task situations suffer the risk of being of no interest to the test participants by being too restricted in terms of what to search for. A similar problem with lack of interest or relevance may be the consequence when simulated work task situations are framed around TREC topics (e.g. Arapakis *et al.*, 2008; Harper and Kelly, 2006; Joho *et al.*, 2008; Ruthven *et al.*, 2002). To search for tropical storms, declining birth rates, robotic technology, tourism increase, and tourist violence as in the study by Harper and Kelly (2006) may not be equally interesting or relevant to all test participants in a study. The issue of interest has been commented upon in an IIR study by Claypool *et al.* (2001, p. 34), who note that test participants do as instructed even if they do not find it interesting. The explanation for this behaviour is known as the good-subject effect, that refers to test participants' willingness to do what they think is desired or expected of them (e.g. Nichols and Maner, 2008). In other words, it is not difficult to get test participants to search, but it requires an effort to create realistic and hence reliable IIR.

Yet another often observed problem in the sample is the unintended built-in conditions with the purpose to create context that affect the realism of the simulated work task situations. To illustrate the problem is here an example of a simulated work task situation from a study of online museum guests of a military history museum (Skov and Ingwersen, 2008, p. 111):

> You went to the flea market last weekend and by coincidence you found an old powder horn. You bought the powder horn and was told, that it had been used in connection with hunting.

> However, one of your friends is certain it was used in the military. Now you are looking for different types of powder horns to try to decide on its use.

This simulated work task situation requires that each and every test participant in this study have visited the flea market last weekend, bought a powder horn, been told it was used for hunting, and to have a friend who is certain it was used in the military. Literally speaking, because this is how the simulated work task situations should be considered. In the case with the online museum guests, this is not very likely. Another example is from a study on daily life tasks on the web (Kim, 2007, p. 174), where one of the simulated work task situations read:

> You have recently moved to Boston and you are interested in buying a home. You have heard that most homes built before 1978 have some lead paint, but that their paint status is often reported as "unknown". You think you should learn about lead paint and housing. The Web seems like a good place to locate this information.

This particular simulated work task situation requires that the test participants (university students) have recently moved to Boston, are interested in buying a house, and are concerned/interested in lead paint. In the same study, the test participants were furthermore conditioned to have a cousin who is "a typical teenage girl", who has a friend who has started to smoke, and they were also instructed to plan a visit to San Francisco next week (Kim, 2007, p. 174).

The example of planning a visit, a trip, or a holiday is, in fact, quite common examples of simulated work task situations. For example, US university students who were to plan for a short trip to Rome, Italy (White and Marchionini, 2007, p. 689), or Danish university students who had to plan to visit the Olympic Games in Beijing (Papaeconomou *et al.*, 2008, p. 125). But asking test participants to plan for a trip is only realistic if they are going. If so, one should avoid to define the destination, unless it is known and in common to all test participants. One should rather work with a higher level of abstraction and leave it to the test participants to search for where, when, for how long, at what cost, or what to do. Most people will agree that more realistic searching takes place if one searches for one's own purpose or interest. Since IIR is about users' interaction with system and information, and their perception and satisfaction with retrieved information, realism is important, and so is tailoring of simulated work task situations to achieve it.

The majority of simulated work task situations in our sample are not tailored to ensure realistic IIR. Part of the explanation may be found in the publications by Borlund that promote this approach to IIR evaluation, because in those publications examples of her simulated work task situations are shown (e.g. Borlund, 2000b). To some, they may have served as inspiration. However, Borlund typically tests at a meta-evaluation level and combines good and bad simulated work task situations in order to identify effects of these. This means simulated work task situations used by her are not always examples of good simulated work task situations.

The analysis of whether simulated work task situations in the 67 studies are tailored in accordance with the stipulated requirements also revealed the use of different types of simulated work task situations with the objective to replicate different levels of task complexity (e.g. Bell and Ruthven, 2004; Freund and Toms, 2006; Joho and Jose, 2006b; Sihvonen and Vakkari, 2004; White *et al.*, 2005b; White and Ruthven, 2006) or types of information needs (e.g. Craven, 2003; Freund and Toms, 2006; Joho and Jose, 2006a; Kim, 2007; Langer and Frøkjær, 2008; Pansanato and Fortes, 2007; Tombros *et al.*, 2003, 2005; Toms *et al.*, 2007; White *et al.*, 2001a, b, 2003, 2005a, b, 2007, 2008; White and

Marchionini, 2007). This clearly illustrates the need to conduct IIR studies with different representations of information needs, and emphasises the incompleteness of simulated work task situations by so far only being validated for exploratory searching (e.g. Borlund, 2000b, 2003).

In order to analyse whether the used simulated work task situations were tailored, it is required that the simulated work task situations are depicted in the form presented to the test participants. In 31 per cent (21) of the studies all simulated work task situations were shown (e.g. Bell and Ruthven, 2004; Kim, 2007; Kules and Shneiderman, 2008; Larsen, 2002; Pharo, 2008; Rodden *et al.*, 2001; Ruthven *et al.*, 2002, 2003). In 25 per cent (17) of the studies, examples of used simulated work task situations were displayed (e.g. Fachry *et al.*, 2008; Haggerty *et al.*, 2003; Joho *et al.*, 2008; Langer and Frøkjær, 2008; Larsen and Ingwersen, 2001; Nielsen, 2004; Skov and Ingwersen, 2008; Szlavik *et al.*, 2007; Toms *et al.*, 2007). This is compared to 43 per cent (29) of the studies that in some cases did provide information about what was searched for, but did not depict the actual simulated work task situations used (e.g. Petrelli *et al.*, 2006; Price *et al.*, 2007; Qu and Furnas, 2008). Obviously, it is necessary with respect to the current analysis to be able to read the actual simulated work task situations, but so it is to all readers of the studies. The used simulated work task situations should to be depicted in the form presented to the test participants in order for the readers to understand and assess the reported result. This has made us realise that the requirement to depict the used simulated work task situations in the form presented to the test participants should be added to the list of requirements regarding the use of simulated work task situations.

4.3 Are personal information needs included as baseline?

The requirement to include test participants' personal information needs serves the purpose of validating the information search behaviour achieved via the searching of the simulated work task situations, and hence functions as a baseline of IIR. Hereby, the personal information needs become the tool to compare, interpret, and validate the test participants' interaction patterns achieved by use of the simulated work task situations. In our sample, only two studies (3 per cent) (Blomgren *et al.*, 2004; Pharo, 2008) include test participants' personal information needs, but they do not use them as baseline to compare the results to those of the simulated work task situations. In both cases, the personal information needs are used to evaluate the system's effect on genuine information needs. Blomgren *et al.* (2004) are, in addition, the only ones to have employed the entire IIR evaluation model in their study of journalists' information searching and satisfaction with search results. In their assessment on the use of the IIR evaluation model, they conclude:

> To sum up, we mean that the evaluation method used in this study is well suited for evaluations of operational systems, covering system, user and context. It aims to provide an overall view of how well the system suits its users and the system's role among other available information sources. The approach as such has functioned well and provided a solid methodological base. The measures used have yielded valuable information about the system from the users' point of view. These different measures functioned well and generated different types of information to complete each other (Blomgren *et al.*, 2004, p. 67).

More specifically they note "[…] that composing a simulated work task situation that offers a sufficient level of reality for all participants must be done with great care. Moreover, the importance of using at least one real work task cannot be overvalued" (Blomgren *et al.*, 2004, p. 66). The results and experiences by Blomgren *et al.* (2004)

confirm two things: first, it is important to also involve the test participants' personal information needs because the search behaviour on the basis of the personal, genuine information needs provides an indication of how (realistically) the information search behaviour of the simulated work task situations can be interpreted; and second, tailoring of the simulated work task situations is essential for engaged and reliable search interaction by the test participants.

4.4 Are the simulated work task situations rotated?

In total, 61 per cent (41) of the studies explicitly report to have rotated the simulated work task situations (and systems/interfaces under testing) to neutralise order effects (e.g. Arapakis *et al.*, 2008; Bierig and Goker, 2006; Fachry *et al.*, 2008; Freund and Toms, 2006; Haggerty *et al.*, 2003; Wu *et al.*, 2004). In total, 31 per cent (21) of the studies provide no information about rotation (e.g. Airio, 2008; Brajnik *et al.*, 2002; Craven, 2003; Kim, 2007; Pansanato and Fortes, 2007), few studies have no need of rotation as only one simulated work task situation is used (Larsen and Ingwersen, 2001; Larsen, 2002; Rodden *et al.*, 2001), and others presented the simulated work tasks in decreasing order of interest as expressed by the test participants (e.g. Harper *et al.*, 2003; Joho *et al.*, 2008). Joho *et al.* (2008, p. 92) explain how "[t]he three topics selected by participants were presented in decreasing order of their interest. In other words, they performed the least interesting topic (out of three) first and the most interesting topic last, to compensate for the fatigue effect with their topic interest".

4.5 Is pilot testing carried out to tune simulated work task situations?

The final questions concerns the requirement to use pilot testing to gain inspiration for simulated work task situations based on pilot-test participants' personal information need, and to evaluate the suitability of the already devised simulated work task situations. Many of the studies mention to having pilot tested, but unfortunately none with the purpose to tailor and evaluated the simulated work task situations. No doubt pilot testing of the simulated work task situations would be an excellent way to get informed about test participants' perceptions of the simulated work task situations and to adjust and tailor accordingly, if necessary. Our analysis shows room for improvement of tailoring of the simulated work task situations, hence this requirement should be strongly emphasised.

4.6 Further discussions and future work

The analysis shows a need for clarifications of the requirements of how to use simulated work task situations. This is in particular with respect to "realism" of the simulated work task situations seen by the lack of tailoring to suit the test participants. It is important to tailor to ensure that the employed simulated work task situations are realistic, relevant, and interesting from the test participants' point of view with the purpose to obtain realistic and engaged interactive search behaviour. In order to put power behind the emphasis of tailoring of simulated work task situations, and to explicitly verify this requirement, it would be useful to further investigate the effect of non-tailored simulated work task situations compared to tailored situations.

The issue of interest is considered in several studies, and are for example handled by letting the test participants choose between several simulated work task situations or topics framed in the same simulated work task situation (e.g. Airio, 2008; Arapakis *et al.*, 2008; Joho *et al.*, 2008). In the study by Airio (2008), the test participants got to

choose between ten simulated work task situations of which they selected four that they would like to search. In the case of Arapakis *et al.* (2008), the test participants were to search three simulated work task situations at each their level of difficulty. For each level of difficulty the test participants could choose between two scenarios of simulated work task situations. The study by Joho *et al.* (2008) is another example of how the test participants got to choose three topics out of 15 TREC topics to be inserted in the context of a simulated work situation. In addition, they normalised for the possible effect of "interestness" by letting the test participants search in decreasing order of their interest. It is a good idea to let test participants choose between simulated work task situations/topics of interest to them to ensure interest in searching. At the same time, one should be aware of the danger in terms of providing too many different simulated work task situations to choose between that generalisations of search interaction cannot be made across the test participants. We do not really know enough about how the issue of interest or relevancy of simulated work task situations affects the test participants' and their system interaction and relevance assessments. This lack of knowledge obviously calls for further studies. The normalisation of interest addresses a different phenomenon namely that of order effect, which is usually handled with the requirement of rotation of the employed simulated work task situations. The phenomenon of order effect is well-known, but is not yet fully understood or investigated in relation to IIR, thus this also calls for further studies and has recently been addressed by Clemmensen and Borlund (2016).

The analysis also revealed the incompleteness of simulated work task situations by so far only being validated for the user group of university students and exploratory searching. Obviously, validation studies with different user groups are required. From the sample, we know how several studies reflected different task complexity levels and types of information needs (e.g. Bell and Ruthven, 2004; Toms *et al.*, 2007; White *et al.*, 2001a, 2002, 2003, 2008). Thus further studies are needed that verify the search behaviour of different complexity levels in order to provide appropriate recommendations and guidelines for construction of these types of simulated work task situations. The study by Bell and Ruthven (2004) makes a first attempt to validate how task complexity/difficulty affects web IR. From a meta-evaluation point of view, this is interesting as the study validates task complexity of simulated work task situations. However, further validation is required in order to provide clear guidelines and recommendations for constructions of simulated work task situations with reference to task complexity/difficulty. The structure/complexity of tasks applied in Bell and Ruthven's study resembles the three types of information needs categorised by Ingwersen (2000, pp. 163-165) known as the verificative information need, conscious topical information need, and muddled topical information need. The first complexity-level corresponds to the verificative information need and is fact-oriented. The second complexity level, which is semantically more open, corresponds to the concept of a simulated work task situation, which again can be seen as a combination of Ingwersen's conscious and muddled topical information needs. The third level of complexity matches the muddled topical need when most vague in nature. These types of information needs are also represented as simulated work task situations in the studies by White and colleagues (e.g. White *et al.*, 2001a, 2002, 2003, 2008) with IIR carried out with respect to fact search, decision search, and background/exploratory search as well as by Toms and her group of colleagues (Toms *et al.*, 2007). The study by Toms and her colleagues compliments the work by Bell and Ruthven in terms of how different types of search tasks and task structures lead to different search efforts of the

test participants. Clearly further research is needed to get a deeper understanding of the sub-components of work tasks and their effects on the use and interaction with IR systems. As such, we are in line with Toms and co-workers (2007, p. 370) who conclude that: "[o]verall, our results demonstrate different levels of effort expended by participants relative to task types and structures. This underscores the need to understand the effects of task on search behaviour […]". The study by Borlund and Dreier (2014) that verifies the predicted search behaviour of Ingwersen's three types of information needs in the setting of everyday life information seeking is an attempt of such a study.

5. Concluding statements

The overall objectives of the present study were to learn for what types of IIR evaluations simulated work task situations are used, and how they are used in the IIR evaluation studies. In particular, we wanted to learn about the intentional, and the unintentional use of simulated work task situations in order to clarify and improve the requirements for the application of simulated work task situations in order to obtain the intended authentic and realistic IIR under study.

A sample of 67 papers that report on empirical IIR evaluations by use of simulated work task situations was analysed. The sample was obtained from a citation time window of ten years (1998-2008). At an overall level, the analysis aimed to answer the following five questions: (1) What was evaluated? Were the simulated work task situations tailored? Were personal information needs included as baseline? Were the simulated work task situations permuted? and Was pilot testing carried out to tune simulated work task situations? The analysis showed that the IIR evaluations studies could roughly be divided into three categories: evaluation of systems performance; evaluation of systems facilities and functionalities; and search behaviour. To sum up: the analysis showed a need for clarifications of how to use simulated work task situations in IIR evaluations. This paper may also be seen as an attempt to do that. The clarification is particularly needed with respect to the design and creation of realistic simulated work task situations. There was a lack of tailoring of the simulated work task situations to suit the test participants, and they were often formulated with built-in conditions that unintended minimised the realism of the simulated work task situation. The requirement to include the test participants' personal information needs was neglected, even though it would have qualified the interpretation of the IIR results in general, and validated the obtained results with respect to different types of information needs, complexity levels, and user groups. None of the analysed studies reported to have used pilot testing with the purpose to tune and refine the simulated work task situations. In brief, we can conclude that the test instrument of a simulated work task situation is accepted, but the requirements for its use are not followed. Further, the analysis made it clear that the list of requirements should be extended with the requirement to always depict the used simulated work task situations when reporting the IIR studies. As a result, the requirements for the use of simulated work task situations read:

(1) To tailor the simulated work task situation to the test participants:

- a situation the test participants can relate to and identify themselves with;
- a situation the test participants find topically interesting and/or of relevance to them; and
- a situation that provides enough imaginative context in order for the test participants to be able to apply the situation.

(2) To include test participants' personal information needs as baseline.

(3) To rotate the order of simulated work task situation and personal information needs (counterbalancing).

(4) To pilot test prior to actual testing (often more than once).

(5) To display the used simulated work task situations when reporting the study.

In addition, a number of future studies were identified that will validate and increase our understanding of how to use simulated work task situations. The identified studies address the effect of tailoring of simulated work task situations, the effect of interest and hence closely related to the study on effect of tailoring. A study of the phenomenon of order effect of simulated work task situations was identified, and so were studies that mirrored different types of information needs and/or task complexity levels of simulated work task situations and its use to different groups of users. It is our hope that these future studies will result in refined requirements about how to use simulated work task situations and increase the knowledge base of empirical IIR studies.

Note

1. Rodden *et al.* (2001) report two experiments, one experiment using conference attendees, and another experiment using university students. Hence this publication is counted twice, once in each category.

References

Airio, E. (2008), "Who benefits from CLIR in web retrieval?", *Journal of Documentation*, Vol. 64 No. 5, pp. 760-778.

Arapakis, I., Jose, J.M. and Gray, P.D. (2008), "Affective feedback: an investigation into the role of emotion in the information seeking process", in Myaeng, S.H., Oard, D.W. and Sebastiani, F. (Eds), *Proceedings of the 31st ACM Sigir Conference on Research and Development of Information Retrieval (SIGIR 2008)*, ACM Press, Singapore, pp. 395-402.

Belkin, N.J. (2008), "Some(what) grand challenges for information retrieval", *ACM SIGIR Forum*, Vol. 42 No. 1, pp. 47-54.

Bell, J.D. and Ruthven, I. (2004), "Searcher's assessment of task complexity for web searching", in McDonald, S. and Tait, J. (Eds), *Proceeding of the 26th European Conference on Information Retrieval (ECIR 2004)*, Springer Verlag and University of Sunderland, Berlin and Heidelberg, pp. 57-71.

Bierig, R. and Goker, A. (2006), "Time, location and interest: an empirical and user-centred study", in Ruthven, I., Borlund, P., Ingwersen, P., Belkin, N., Tombros, A. and Vakkari, P. (Eds), *Proceedings of the 1st International Symposium on Information Interaction in Context (IIiX 2006)*, ACM Press, Copenhagen, pp. 79-87.

Blomgren, L., Vallo, H. and Bystrom, K. (2004), "Evaluation of an information system in an information seeking process", in Heery, R. and Lyon, L. (Eds), *Proceedings of the 8th International European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2004), Bath*, Springer Verlag, Berlin and Heidelberg, pp. 57-68.

Borlund, P. (2000a), "Evaluation of interactive information retrieval systems", doctoral thesis, Åbo Akademi University, Åbo Akademi University Press, Turku.

Borlund, P. (2000b), "Experimental components for the evaluation of interactive information retrieval systems", *Journal of Documentation*, Vol. 56 No. 1, pp. 71-90.

Borlund, P. (2003), "The IIR evaluation model: a framework for evaluation of interactive information retrieval systems", *Information Research*, Vol. 8 No. 3, available at: www.informationr.net/ir/8-3/paper152.html (accessed 26 May 2015).

Borlund, P. and Dreier, S. (2014), "An investigation of the search behaviour associated with Ingwersen's three types of information needs", *Information Processing & Management*, Vol. 50 No. 4, pp. 493-507.

Borlund, P. and Ingwersen, P. (1997), "The development of a method for the evaluation of interactive information retrieval systems", *Journal of Documentation*, Vol. 53 No. 3, pp. 225-250.

Borlund, P. and Ingwersen, P. (1998), "Measures of relative relevance and ranked half-life: performance indicators for interactive IR", in Croft, B.W., Moffat, A., van Rijsbergen, C.J., Wilkinson, R. and Zobel, J. (Eds), *Proceedings of the 21st ACM Sigir Conference on Research and Development of Information Retrieval (SIGIR 1998)*, ACM Press and York Press, Melbourne, pp. 324-331.

Borlund, P. and Ingwersen, P. (1999), "The application of work tasks in connection with the evaluation of interactive information retrieval systems: empirical results", in Draper, S.W., Dunlop, M.D., Ruthven, I. and van Rijsbergen, C.J. (Eds), *Mira '99: Evaluating Interactive Information Retrieval*, British Computer Society, Glasgow, pp. 1-18.

Borlund, P. and Schneider, J.W. (2010), "Reconsideration of the simulated work task situation: a context instrument for evaluation of information retrieval interaction", *IIiX'10 Proceeding of the Third Symposium on Information Interaction in Context, ACM, New York, NY* and *New Brunswick, NJ*, pp. 155-164.

Brajnik, G., Mizzaro, S. and Tasso, C. (2002), "Strategic help in user interfaces for information retrieval", *Journal of the American Society for Information Science and Technology*, Vol. 53 No. 5, pp. 343-358.

Claypool, M., Le, P., Waseda, M. and Brown, D. (2001), "Implicit interest indicators", *Proceedings of ACM Intelligent User Interfaces Conferences (UIU '01), Santa Fe*, pp. 33-40.

Clemmensen, M.L. and Borlund, P. (2016), "Order effect in interactive information retrieval evaluation: an empirical study", *Journal of Documentation*, Vol. 72 No. 2, pp. 194-213.

Craven, J. (2003), "Access to electronic resources by visually impaired people", *Information Research*, Vol. 8 No. 4, available at: http://informationr.net/ir/8-4/paper156.html (accessed 26 May 2015).

Fachry, K.N., Kamps, J. and Zhang, J. (2008), "Access to archival material in context", in Borlund, P., Schneider, J.W., Tombros, A., Lalmas, M., Feather, J., Kelly, D. and de Vries, A. (Eds), *Proceeding of the Second IIiX Symposium on Information Interaction in Context (IIiX 2008)*, ACM Press, London, pp. 102-109.

Freund, L. and Toms, E.G. (2006), "Enterprise search behavior of software engineers", in Efthimiadis, E.N., Dumais, S.T., Hawking, D. and Järvelin, K. (Eds), *Proceedings of the 29th ACM Sigir Conference on Research and Development of Information Retrieval (SIGIR 2006)*, ACM Press, Seattle, WA, pp. 645-646.

Haggerty, A., White, R.W. and Jose, J.M. (2003), "NewsFlash: adaptive TV news delivery on the web", in Nurnberger, A. and Detyniecki, M. (Eds), *Proceedings of the 1st International Workshop on Adaptive Multimedia Retrieval, Hamburg (AMR 2003)*, Springer Verlag, Berlin and Heidelberg, pp. 72-86.

Hammer-Aebi, B., Christensen, K.W., Lund, H. and Larsen, B. (2006), "Users, structured documents and overlap: interactive searching of elements and the influence of context on search behavior", in Ruthven, I., Borlund, P., Ingwersen, P., Belkin, N., Tombros, A. and Vakkari, P. (Eds), *Proceedings of the 1st International Symposium on Information Interaction in Context (IIiX 2006)*, ACM Press, Copenhagen, pp. 46-55.

Harper, D.J. and Kelly, D. (2006), "Contextual relevance feedback", in Ruthven, I., Borlund, P., Ingwersen, P., Belkin, N., Tombros, A. and Vakkari, P. (Eds), *Proceedings of the 1st International Symposium on Information Interaction in Context (IIiX 2006)*, ACM Press, Copenhagen, pp. 129-137.

Harper, D.J., Koychev, I. and Sun, Y.X. (2003), "Query-based document skimming: a user-centred evaluation of relevance profiling", in Sebastiani, F. (Ed.), *Proceeding of the 25th European Conference on Information Retrieval (ECIR 2003), Pisa*, Springer Verlag, Berlin/Heidelberg, pp. 377-392.

Henrich, J., Heine, S.J. and Norenzayan, A. (2010), "The weirdest people in the world?", *Behavioral and Brain Sciences*, Vol. 33 Nos 2/3, pp. 61-135.

Ingwersen, P. (2000), "Users in context", in Agosti, M., Crestani, F. and Pasi, G. (Eds), *Lectures on Information Retrieval. Third European Summer-School, ESSIR*, Springer-Verlag, Varenna and Heidelberg, pp. 157-178.

Joho, H. and Jose, J.M. (2006a), "A comparative study of the effectiveness of search result presentation on the web", in Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T. and Yavlinsky, A. (Eds), *Proceedings of the 28th European Conference on IR Research (ECIR 2006), London*, Springer Verlag, Berlin and Heidelberg, pp. 302-313.

Joho, H. and Jose, J.M. (2006b), "Slicing and dicing the information space using local context", in Ruthven, I., Borlund, P., Ingwersen, P., Belkin, N., Tombros, A. and Vakkari, P. (Eds), *Proceedings of the 1st International Symposium on Information Interaction in Context (IIiX 2006)*, ACM Press, Copenhagen, pp. 66-74.

Joho, H. and Jose, J.M. (2008), "Effectiveness of additional representations for the search result presentation on the web", *Information Processing & Management*, Vol. 44 No. 1, pp. 226-241.

Joho, H., Hannah, D. and Jose, J.M. (2008), "Comparing collaborative and independent search in a re-call-oriented task", in Borlund, P., Schneider, J.W., Tombros, A., Lalmas, M., Feather, J., Kelly, D. and de Vries, A. (Eds), *Proceeding of the Second IIiX Symposium on Information Interaction in Context (IIiX 2008)*, ACM Press, London, pp. 89-96.

Jose, J.M., Furner, J. and Harper, D.J. (1998), "Spatial querying for image retrieval: a user-oriented evaluation", in Croft, B.W., Moffat, A., van Rijsbergen, C.J., Wilkinson, R. and Zobel, J. (Eds), *Proceedings of the 21st ACM Sigir Conference on Research and Development of Information Retrieval (SIGIR 1998)*, ACM Press and York Press, Melbourne, pp. 223-240.

Kim, J. (2007), "Task as a context of information seeking: an investigation of daily life tasks on the web", *Libri*, Vol. 58 No. 3, pp. 172-181.

Kules, B. and Shneiderman, B. (2008), "Users can change their web search tactics: design guidelines for categorized overviews", *Information Processing & Management*, Vol. 44 No. 2, pp. 463-484.

Langer, L. and Frøkjær, E. (2008), "Improving web search transparency by using a venn diagram interface", in Tollmar, K. and Jönsson, B. (Eds), *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges*, ACM Press, Lund, pp. 249-256.

Larsen, B. and Ingwersen, P. (2001), "Synchronous and diachronous citation analysis for information retrieval: generating a boomerang effect form the network of scientific papers", *8th International Conference on Scientometrics and Informetrics Proceedings: ISSI – 2001, Sydney, 16-20 July*, pp. 355-368.

Larsen, B. (2002), "Exploiting citation overlaps for information retrieval: generating a boomerang effect from the network of scientific papers", *Scientometrics*, Vol. 54 No. 2, pp. 155-178.

Nichols, A.L. and Maner, J.K. (2008), "The good-subject effect: investigating participant demand characteristics", *The Journal of General Psychology*, Vol. 135 No. 2, pp. 151-165.

Nielsen, M.L. (2004), "Task-based evaluation of associative thesaurus in real-life environment", in Schamber, L. and Barry, C.L. (Eds), *The Proceedings of the 67th ASIS&T Annual Meeting*, Information Today, Inc., Medford, NJ, pp. 437-447.

Pansanato, L.T.E. and Fortes, R.P.M. (2007), "System description: an orienteering strategy to browse semantically-enhanced educational wiki pages", in Franconi, E., Kifer, M. and May, W. (Eds), *Proceedings of the 4th European Semantic Web Conference: Innsbruck*, Springer Verlag, Berlin and Heidelberg, pp. 809-818.

Papaeconomou, C., Zijlema, A. and Ingwersen, P. (2008), "Searchers' relevance judgments and criteria in evaluating web pages in a learning style perspective", in Borlund, P., Schneider, J.W., Tombros, A., Lalmas, M., Feather, J., Kelly, D. and de Vries, A. (Eds), *Proceeding of the Second IIiX Symposium on Information Interaction in Context (IIiX 2008)*, ACM Press, London, pp. 123-132.

Petrelli, D. (2008), "On the role of user-centred evaluation in the advancement of interactive information retrieval", *Information Processing & Management*, Vol. 44 No. 1, pp. 22-38.

Petrelli, D. and Auld, D. (2008), "An examination of automatic video retrieval technology on access to the contents of an historical video archive", *Program Electronic Library and Information Systems*, Vol. 42 No. 2, pp. 115-136.

Petrelli, D., Levin, S., Beaulieu, M. and Sanderson, M. (2006), "Which user interaction for cross-language information retrieval? Design issues and reflections", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 5, pp. 709-722.

Petrelli, D., Demetriou, G., Herring, P., Beaulieu, M. and Sanderson, M. (2002), "Exploring the effect of query translation when searching cross-language", in Peters, C., Braschler, M., Gonzalo, J. and Kluck, M. (Eds), *Proceedings of the 3rd Workshop of the Cross-Language Evaluation Forum (CLEF 2002): Rome*, Springer Verlag, Berlin and Heidelberg, pp. 430-445.

Petrelli, D., Beaulieu, M., Sanderson, M., Demetriou, G., Herring, P. and Hansen, P. (2004), "Observing users, designing clarity: a case study on the user-centered design of a cross-language information retrieval system", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 10, pp. 923-934.

Pharo, N. (2008), "The effect of granularity and order in XML element retrieval", *Information Processing & Management*, Vol. 44 No. 5, pp. 1732-1740.

Pharo, N. and Nordlie, R. (2005), "Context matters: an analysis of assessments of XML documents", in Crestani, F. and Ruthven, I. (Eds), *Proceedings of the 5th International Conference on Conceptions of Library and Information Sciences (CoLIS 2005): Glasgow*, Springer Verlag, Berlin and Heidelberg, pp. 238-248.

Price, S.L., Nielsen, M.L., Delcambre, L.M.L. and Vedsted, P. (2007), "Semantic components enhance retrieval of domain-specific documents", in Silva, M.J., Laender, A.H.F., Baeza-Yates, R.A., McGuinness, D.L., Olstad, B., Olsen, Æ.H. and Falcão, A.O. (Eds), *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM 2007)*, ACM Press, Lisbon, pp. 429-438.

Qu, Y. and Furnas, G.W. (2008), "Model-driven formative evaluation of exploratory search: a study under a sensemaking framework", *Information Processing & Management*, Vol. 44 No. 2, pp. 534-555.

Rodden, K., Basalaj, W., Sinclair, D. and Wood, K. (2001), "Does organisation by similarity assist image browsing?", *Proceedings of the SIG-CHI on Human Factors in Computing Systems, ACM Press, Seattle, WA, 31 March -5 April*, pp. 190-197.

Ruthven, I., Lalmas, M. and van Rijsbergen, K. (2002), "Ranking expansion terms with partial and ostensive evidence", in Bruce, H., Fidel, R., Ingwersen, P. and Vakkari, P. (Eds), *The Proceedings of the 4th International Conference on Conceptions of Library and Information Science (CoLIS 4). Seattle, WA, July 21-25*, Libraries Unlimited, Greenwood Village, CO, pp. 199-219.

Ruthven, I., Lalmas, M. and van Rijsbergen, K. (2003), "Incorporating user search behavior into relevance feedback", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 6, pp. 529-549.

Sagara, Y. (1999), "The implications of the non-keyword information on the user's judgment of relevance in information retrieval", *Japanese Journal of Library and Information Science*, Vol. 42, pp. 23-46.

Sihvonen, A. and Vakkari, P. (2004), "Subject knowledge improves interactive query expansion assisted by a thesaurus", *Journal of Documentation*, Vol. 60 No. 6, pp. 673-690.

Skov, M. and Ingwersen, P. (2008), "Exploring information seeking behavior in a digital museum context", in Borlund, P., Schneider, J.W., Tombros, A., Lalmas, M., Feather, J., Kelly, D. and de Vries, A. (Eds), *Proceeding of the Second IIiX Symposium on Information Interaction in Context (IIiX 2008)*, ACM Press, London, pp. 110-115.

Suomela, S. and Kekäläinen, J. (2005a), "Ontology as a search-tool: a study of real users' query formulation with and without conceptual support", in Losada, D.E. and Fernández-Luna, J.M. (Eds), *Proceedings of the 27th European Conference on IR Research (ECIR 2005), Santiago de Compostela*, Springer Verlag, Berlin and Heidelberg, pp. 315-329.

Suomela, S. and Kekalainen, J. (2005b), "User evaluation of ontology as query construction tool", *Information Retrieval*, Vol. 9 No. 4, pp. 455-475.

Szlavik, Z., Tombros, A. and Lalmas, M. (2006a), "Investigating the use of summarization for interactive XML retrieval", *Proceedings of the 21st Annual ACM Symposium on Applied Computing (SAC 2006). ACM Press*, *Dijon*, *23-27 April*, pp. 1068-1072.

Szlavik, Z., Tombros, A. and Lalmas, M. (2006b), "The use of summaries in XML retrieval", in Gonzalo, J., Thanos, C., Verdejo, M.F. and Carrasco, R.C. (Eds), *Proceedings of 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006), Alicante*, Springer Verlag, Berlin and Heidelberg, pp. 75-86.

Szlavik, Z., Tombros, A. and Lalmas, M. (2007), "Feature- and query-based table of contents generation for XML documents", in Amati, G., Carpineto, C. and Romano, G. (Eds), *Proceedings the 29th European Conference on IR Research (ECIR 2007), Rome*, Springer Verlag, Berlin and Heidelberg, pp. 456-467.

Tombros, A., Ruthven, I. and Jose, J.M. (2003), "Searchers' criteria for assessing web pages", *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), ACM Press*, *Toronto*, *28 July-1 August*, pp. 385-386.

Tombros, A., Ruthven, I. and Jose, J.M. (2005), "How users assess web pages for information seeking", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 4, pp. 327-344.

Toms, E.G., O'Brien, H., Mackenzie, T., Jordan, C., Freund, L., Toze, S., Dawe, E. and MacNutt, A. (2007), "Task effects on interactive search: the query factor", in Fuhr, N., Kamps, J., Lalmas, M. and Trotman, A. (Eds), *Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX), Dagstuhl*, Springer Verlag, Berlin and Heidelberg, pp. 359-372.

Villa, R., Gildea, N. and Jose, J.M. (2008), "Facetbrowser: a user interface for complex search tasks", in El-Saddik, A., Vuong, S., Griwodz, C., Del Bimbo, A., Candan, K.S. and Jaimes, A. (Eds), *Proceedings of the 16th International Conference on Multimedia 2008 (MM '08)*, ACM Press, Vancouver, pp. 489-498.

Wen, L., Ruthven, I. and Borlund, P. (2006), "The effects on topic familiarity on online search behavior and use of relevance criteria", in Lalmas, M., MacFarlane, A., Rüger, S., Tombros, A., Tsikrika, T. and Yavlinsky, A. (Eds), *Proceedings of the 28th European Conference on IR Research (ECIR 2006), London*, Springer Verlag, Berlin and Heidelberg, pp. 456-459.

White, R., Bilenko, M. and Cucerzan, S. (2007), "Studying the use of popular destinations to enhance web search interaction", in Kraaij, W., de Vries, A.P., Clarke, C.L.A., Fuhr, N. and Kando, N. (Eds), *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, ACM Press, Amsterdam, pp. 159-166.

White, R., Bilenko, M. and Cucerzan, S. (2008), "Leveraging popular destinations to enhance web search interaction", *ACM Transactions on the Web*, Vol. 2 No. 3, p. 30.

White, R., Ruthven, I. and Jose, J.M. (2001a), "Web document summarisation: a task-oriented evaluation", in Tjoa, A.T. and Wagner, R. (Eds), *Proceedings of the 12th International Conference on Database and Expert Systems Applications (DEXA 2001), Munich*, Springer Verlag, Berlin and Heidelberg, pp. 551-555.

White, R.W. and Ruthven, I. (2006), "A study of interface support mechanisms for interactive information retrieval", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 7, pp. 933-948.

White, R.W. and Marchionini, G. (2006), "A study of real-time query expansion effectiveness", in Efthimiadis, E.N., Dumais, S.T., Hawking, D. and Järvelin, K. (Eds), *Proceedings of the 29th ACM Sigir Conference on Research and Development of Information Retrieval (SIGIR 2006)*, ACM Press, Seattle, WA, pp. 715-716.

White, R.W. and Marchionini, G. (2007), "Examining the effectiveness of real-time query expansion", *Information Processing & Management*, Vol. 43 No. 3, pp. 685-704.

White, R.W., Jose, J.M. and Ruthven, I. (2001b), "Query-biased web page summarization: a task-oriented evaluation", in Croft, W.B., Harper, D.J., Kraft, D.H. and Zobel, J. (Eds), *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, ACM Press, New Orleans, LA, pp. 412-413.

White, R.W., Jose, J.M. and Ruthven, I. (2003), "A task-oriented study on the influencing effects of query-biased summarisation in web searching", *Information Processing & Management*, Vol. 39 No. 5, pp. 707-733.

White, R.W., Jose, J.M. and Ruthven, I. (2005a), "Using top-ranking sentences to facilitate effective information access", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 10, pp. 1113-1125.

White, R.W., Jose, J.M. and Ruthven, I. (2006), "An implicit feedback approach for interactive information retrieval", *Information Processing & Management*, Vol. 42 No. 1, pp. 166-190.

White, R.W., Ruthven, I. and Jose, J.M. (2002), "Finding relevant documents using top-ranking sentences: an evaluation of two alternative schemes", in Beaulieu, M., Baeza-Yates, R., Myaeng, S.H. and Järvelin, K. (Eds), *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, ACM Press, Tampere, pp. 57-64.

White, R.W., Ruthven, I. and Jose, J.M. (2005b), "A study of factors affecting the utility of implicit relevance feedback", in Baeza-Yates, R.A., Ziviani, N., Marchionini, G., Moffat, A. and Tait, J. (Eds), *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, ACM Press, Salvador, pp. 35-42.

Wu, M., Muresan, G., McLean, A., Tang, M., Wilkinson, R., Li, Y., Lee, H. and Belkin, N.J. (2004), "Human versus machine in the topic distillation task", in Sanderson, M., Järvelin, K., Allan, J. and Bruza, P. (Eds), *Proceedings of the 27th Conference on Research and Development in Information Retrieval (SIGIR 2004)*, ACM Press, Sheffield, pp. 385-392.

Xu, Y.J. and Yin, H. (2008), "Novelty and topicality in interactive information retrieval", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 2, pp. 201-215.

Xu, Y.J. and Wang, D. (2008), "Order effect in relevance judgment", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 8, pp. 1264-1275.

Yuan, X. (2006), "Supporting multiple information-seeking strategies in a single system framework", in Moore, R.C., Bilmes, J.A., Chu-Carroll, J. and Sanderson, M. (Eds), *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, The Association for Computational Linguistics, New York, NY, pp. 147-249.

Yuan, X. and Belkin, N.J. (2007), "Supporting multiple information-seeking strategies in a single system framework", in Kraaij, W., de Vries, A.P., Clarke, C.L.A., Fuhr, N. and Kando, N. (Eds), *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, ACM Press, Amsterdam, pp. 247-254.

Corresponding author
Pia Borlund can be contacted at: pia.borlund@hum.ku.dk