Received 28 April 2015
Revised 5 August 2015
Accepted 22 August 2015

Order effect in interactive information retrieval evaluation: an empirical study

Melanie Landvad Clemmensen and Pia Borlund
University of Copenhagen, Copenhagen, Denmark

Abstract

Purpose – The purpose of this paper is to report a study of order effect in interactive information retrieval (IIR) studies. The phenomenon of order effect is well-known, and it is the main reason why searches are permuted (counter-balanced) between test participants in IIR studies. However, the phenomenon is not yet fully understood or investigated in relation to IIR; hence the objective is to increase the knowledge of this phenomenon in the context of IIR as it has implications for test design of IIR studies.

Design/methodology/approach – Order effect is studied via partly a literature review and partly an empirical IIR study. The empirical IIR study is designed as a classic between-groups design. The IIR search behaviour was logged and complementary post-search interviews were conducted.

Findings – The order effect between groups and within search tasks were measured against nine classic IIR performance parameters of search interaction behaviour. Order effect is seen with respect to three performance parameters (website changes, visit of webpages, and formulation of queries) shown by an increase in activity on the last performed search. Further the theories with respect to motivation, fatigue, and the good-subject effect shed light on how and why order effect may affect test participants' IR system interaction and search behaviour.

Research limitations/implications – Insight about order effect has implications for test design of IIR studies and hence the knowledge base generated on the basis of such studies. Due to the limited sample of 20 test participants (Library and Information Science (LIS) students) inference statistics is not applicable; hence conclusions can be drawn from this sample of test participants only.

Originality/value – Only few studies in LIS focus on order effect and none from the perspective of IIR.

Keywords Evaluation, Research methods, Information retrieval, User studies, Searching, Information searches

Paper type Research paper

1. Introduction

It is common practise to permute, or to rotate, search tasks between test participants in interactive information retrieval (IIR) studies, so that no, or a minimum of test participants conduct the assigned search tasks in the same order (e.g. Kelly, 2009; Tague-Sutcliffe, 1992). The reasons for permuting (counter-balancing) are among others to avoid the possible effect of learning (e.g. topical domain knowledge and/or system knowledge) and fatigue during testing. Though permutation of search tasks is recommended in IIR studies, the phenomenon of order effect is not yet fully understood or investigated in relation to IIR, which calls for further research and motivates the present study and paper. Hence the objective of this paper is to study the phenomenon of order effect of search tasks and hereby increase our knowledge of order effect in the context of IIR. The study is undertaken partly in the form of a review of related work

The authors are grateful to Kaja H. Brandt for her share in this work. Special thanks are due to the test participants. This work has been supported in part by the Danish Ministry of Culture (REX project/TAKT2011-072).

Journal of Documentation
Vol. 72 No. 2, 2016
pp. 194-213

0022-0418
DOI 10.1108/JD-04-2015-0051

and partly as an empirical IIR study. In the empirical IIR study of order effect, we are interested in whether the order effect can be identified, and if, how and why it occurs. We are also interested in whether the test participants during the IIR study become more accustomed and comfortable, whether their commitment changes, and whether a fatigue effect occurs.

The remainder of the paper is structured as follows: Section 2 presents the review of related work on studies of order effect; within these studies different theories are suggested to explain order effect. These explanatory theories are further introduced in Section 3. Section 4 outlines the methodology of the empirical IIR study that is characterised as a classic between-groups design. Section 5 presents the results of search task order between groups and within search tasks, which are discussed in Section 6. Section 7 closes with concluding remarks and puts forward reflections and suggestions for future work.

2. Related work

This section presents a review of studies of order effect within different fields, starting with a review of the few identified studies on order effect in Library and Information Science (LIS). The purpose of this section is to provide a general view of how order effect has been recognised, and even utilised in different ways.

2.1 Order effects in LIS

Few studies in LIS have dealt with order effect, though they focus on how the presentation order of documents effects the relevance assessments (e.g. Eisenberg and Barry, 1988; Huang and Wang, 2004; Parker and Johnson, 1990; Xu and Wang, 2008). These studies provide us with potential explanations as to why order effect occurs. Eisenberg and Barry (1988) build on the work by Eisenberg (1986) when they address the issue of whether the order the documents are presented to test participants affect the judged relevance. Their study consists of two different experiments using two different measuring scales. In the first experiment, a category rating scale ranging from 1-7 was used. In the second experiment, a dynamic estimation scale was used. Half the test participants in both experiments were presented with the documents ranking from low to high relevance, and the other half presented with the documents ranking from high to low relevance. The result of the study indicated an order effect in that the test participants, who were presented with the documents ranking from high to low, had a tendency to underestimate the relevance of the individual documents. The opposite tendency appeared when the test participants were presented with the documents ranking from low to high. Both tendencies are strongest in the first experiment (using the fixed category scale), but are also evident in the second experiment, although not in a significant way (Eisenberg and Barry, 1988). According to Eisenberg and Barry (1988, p. 297), the explanation of the documented relevance assessment behaviour may be that test participants are not likely to use the very top or very low part of the scales to begin with, as they do not know how the following documents will rank (Eisenberg and Barry, 1988, p. 297).

Parker and Johnson (1990) examined a similar hypothesis to that of Eisenberg and Barry (1988). The main difference between the two studies is the estimation scale used to judge relevance. Parker and Johnson used a three-point scale; 1 indicating "relevant", 2 indicating "relevance not known", and 3 indicating "not relevant". The conclusion of their study is that when test participants were presented with less than 15 documents

no order effect is demonstrated (Parker and Johnson, 1990, p. 494). When presented with 15 or more documents, indications show that the latest presented documents are less likely to be judged relevant (Parker and Johnson, 1990, p. 494).

Huang and Wang (2004) investigated the correlation between the number of judged documents and order effect. In other words, how many documents test participants must relevance assess before order effect becomes visible and affect the performance. In this study, a seven-point scale (rating 0-6) was used and document sets of: 5, 15, 30, 45, 60, and 75 documents were judged. In line with the results of Parker and Johnson (1990), no order effect was visible when less than 15 documents were being assessed. But a significant indication of order effect was seen when sets of 15 and 30 documents were judged. The indication was still present when sets of 45 and 60 documents were judged, although not in a significant way. When dealing with a set of 75 documents, the order effect is no longer visible. According to Huang and Wang (2004, p. 974), this may be a result of the fatigue effect that hinders the test participants from making accurate judgements.

Inspired by the previous studies, Xu and Wang (2008) also investigated the order effect of relevance judgments, but they paid more attention to order effects forming mechanisms and proposed a set of forming mechanisms. The main three mechanisms are: learning effect, the sub-need scheduling effect, and the cursoriness effect. In their definition, learning effect refers to the test participants gaining more knowledge about the topic. Therefore, a document presented later might be regarded with a lower degree of relevance, because the information need it fulfils has already been satisfied. The sub-need scheduling effect is an extension of the learning effect. It is based on the notion that an information need often is divided into sub-needs. These sub-needs are partially overlapping, but at the same time managed sequentially. A sub-need scheduling effect arises when a document is presented prior or later than the sub-need it is most relevant to. As a result the document will be judged higher or lower, depending on the time of presentation (Xu and Wang, 2008, p. 1267). The third proposed mechanism is the cursoriness effect. This effect refers to the cognitive capacity and motivation of the test participants. When the first part of an information need has been fulfilled, motivation will reduce and lead to a focus on more peripheral things, the cognitive capacity will decrease and fatigue will set in. In other words, the judgments will be based on more primitive impressions (Xu and Wang, 2008, p. 2168). Xu and Wang (2008, p. 1274) concluded that the observed order effect is caused by a combination of these three mechanisms. One example is that learning effect, both with respect to getting familiar with the test setting and input of information, makes the test participants more informed about the subject of the study (Xu and Wang, 2008). Another effect mentioned by Huang and Wang (2004) is mental fatigue resulting in a decrease in performance.

Although the number of studies of order effect in LIS (and IIR) is limited, the awareness of this effect as a potential bias is present.

2.2 Order effects in marketing and survey research

Order effect is a known phenomenon in different fields. Within the field of marketing and commercials, an order of entrance effect has been studied and the knowledge used to get ahead of competition (Kardes and Kalyanaram, 1992). Studies have shown that the order of entrance effect of a product launched first have the advance of being preferred by consumers as a result of novelty (Kardes and Kalyanaram, 1992). The effect is so strong that products launched later, even though they might have better features, are not preferred simply as a result of the order of entrance effect. Brunel and Nelson (2003) studied the order effect in relation to commercials and found a difference

in perception depending on the order the information was given. These findings can help to determine where and when to place an advert and best get the message across to the consumers (Brunel and Nelson, 2003). Within the field of risk assessments (e.g. Cushing and Ahlawat, 1996; Monroe and Ng, 2000), order effect has been investigated with respect to the belief-adjustment model developed by Hogarth and Einhorn (1992). The belief-adjustment model presents a general framework for people's ability to reconsider their own perception, and focus on the order effect in the process of reconsideration. In brief, the model is based on the assumption that a current state of perception potentially will adjust with the incoming of new information; in other words, a recency effect will potentially occur (Hogarth and Einhorn, 1992). The model has also been applied in LIS research in relation to relevance assessment (e.g. Huang and Wang, 2004; Xu and Wang, 2008). The belief-adjustment model bears a strong resemblance to the work of Belkin (1977, 1980) and Brookes (1977) on information processing throughout the search session.

In marketing, order effect is roughly divided into two types: conditioning effect and positioning effect. The conditioning effect is defined as an effect that is biased by a previous event, for example, a previous question (Laird Landon, 1971; Perreault, 1975; Gibson *et al.*, 1978). This means that a previous question will give certain associations and expectations, and will create the foundation for the future responses or attitude towards the survey. In some cases, the respondents may even begin to adjust their replies in accordance to earlier replies, in order to maintain a consistent response pattern (Perreault, 1975). Effects that can be defined as positioning effect are primacy effect, recency effect, and fatigue effect. Some studies show that primacy effect is more likely to occur when dealing with written forms. This is because participants have a higher likelihood of choosing a response alternative from the top of a list. The effect is less notable if the respondents are familiar with the topic of inquiry (Duffy, 2003). The recency effect is more likely to occur when the response alternatives are read out loud (Krosnick and Alwin, 1987; Duffy, 2003). A third effect is that of fatigue. The fatigue hypothesis was introduced by Clancy and Wachsler (1971). The hypothesis states that respondents during a survey can become fatigued, and that the result of this is a yea-saying response. This is a response style where respondents are likely to agree independently of the content of a question (Clancy and Wachsler, 1971).

3. Explanatory theories of order effect

This section takes a closer look at some of the theories of order effect presented in the previous section. In our study, we have a broader perception of performance than that of the previous research on order effect in LIS where the focus has been on relevance assessments. Therefore, we also draw on experiences from studies of psychology.

3.1 The significance of motivation

Motivation is a factor that greatly influences people in both thought and action. "To be motivated means 'to be moved' to do something" (Ryan and Deci, 2000, p. 54). Research on motivation focuses on what makes people behave in certain ways (Ryan and Deci, 2000; Franken, 2002).

Motivation is a multi-faceted concept. Besides the obvious, that the level of motivation can vary, different types of motivation is also in play (Ryan and Deci, 2000). Motivation can roughly be divided into two main types: intrinsic and extrinsic motivation (Hidi, 2000). Intrinsic motivation is triggered by variables from within an

individual, for example, an activity that is interesting, enjoyable, or in other ways satisfying. An individual can also be motivated by the process itself (Crawford *et al.*, 2002, p. 771; Schiefle, 1998, p. 92). Extrinsic motivation is about reaching a goal, a reward, or to avoid punishment (Crawford *et al.*, 2002, p. 771; Ryan and Deci, 2000, p. 55). The general assumption is that motivation is a combination between intrinsic and extrinsic motivation (Crawford *et al.*, 2002; Hidi, 2000, p. 310).

One aspect to motivation that has been subject to investigation is the bearing of purpose and goal (Latham and Locke, 1979; Locke and Latham, 2006). Goal is perceived to be an important facilitator for motivation as goals prompt attention, effort, and action in a certain direction. High ambitions and concrete goals equal high motivation. Vague goals with low ambitions foster low motivation (Locke and Latham, 2006). Another variable of motivation is interest (Schiefle, 1998). Interest can relate to both the activity in connection to a given task, or to a specific topic. Interest can be driven by either already existing individual interests or by situational interest. The individual interest is of a static and sustainable nature. In contrast, the situational interest prompted by contextual factors is of a more temporary emotional state and dynamic in nature (Hidi and Baird, 1986; Schiefle, 1998). Interest is mainly associated with positive feelings, although not necessarily at all times. Within the concept of interest lies a willingness to make an effort. Novelty is closely related to interest. Because of the curious nature of humans, something new and exciting can prompt motivation. This type of motivation will decrease when there is no new information to process and its novelty is lost. Further, lack of novelty generates boredom. This is handled by either retrieving from the activity or exploring it further (Smith, 1981). How rapidly this happens depends on the level of complexity (Franken, 2002). Often it is the combination between individual and situational motivation that creates excitement and makes something interesting (Schiefle, 1998, pp. 93-94; Krapp, 1999).

In IIR we aim for motivation that will engage our test participants in dedicated search interaction and genuine relevance assessments, which we can then observe and study. This is also seen with the design criterion of interest with respect to simulated work task situations commonly used in IIR studies to initiate searching, which serve that exact purpose to create motivation (Borlund, 2000). In fact, the requirement to tailor the simulated work task situations to the group of test participants concerns motivation and realistic search interaction. This emphasises the need for careful tailoring of the simulated work task situations in that, by not being equally motivational, unintended order effect may occur.

3.2 The concept of fatigue

Fatigue is a concept that is often suggested as a factor that affects human performance (e.g. Bar-Ilan *et al.*, 2009; Huang and Wang, 2004; Kantowitz *et al.*, 2001; Xu and Wang, 2008). The concept of fatigue covers both psychical and mental characteristics. Psychical fatigue occurs due to psychical strain and manifests as psychical discomfort and declining strength. Mental fatigue results in decreasing mental capacity and occurs due to mental strain. The result of this is a sense of fatigue, lower motivation, and decrease in performance (Barker and Nussbaum, 2011; Holding, 1983; van der Linden *et al.*, 2003). In the following, we primarily focus on mental fatigue, as this is the most relevant part of the concept in relation to IIR.

It is important to note that both psychical and mental aspects of the concept contribute to a potential decrease in performance. Some of the most distinct features of fatigue are decreasing commitment and effort (Holding, 1983; van der Linden *et al.*, 2003). This is, for

example, reported in the study by van der Linden *et al.* (2003) who focus on the impact of mental fatigue when performing a difficult task. The group of test participants were divided into two groups, and the first group were initially asked to perform a task that demanded much mental capacity. Then the two groups were given the same task and asked to perform it in accordance to a given example. The results of the study showed that the members of the group whom were initially asked to perform a demanding mental task were less systematic and focused on the second task. Also, they made more mistakes than the group who did not perform the initial task (van der Linden *et al.*, 2003).

As shown by, e.g. van der Linden *et al.* (2003), the effect of fatigue can affect the performance of test participants. But at the same time other factors are in play and can potentially abolish or at least downplay the effect. As Holding (1983) points out, the sense of being fatigued, does not necessarily lead to a deterioration of capacity or decrease in performance. Even though fatigue is a powerful effect, it is easily outmatched. Factors as, for example, enthusiasm or an emergency seem to some degree to neutralise the effect of fatigue (Barker and Nussbaum, 2011; Holding, 1983). This appears to be the case with the study presented by Barker and Nussbaum (2011). They tested the effect of fatigue on nurses' performance and anticipated a decrease in performance in the final part of the test. However, the opposite happened as the effort increased in the final part of the test (Barker and Nussbaum, 2011).

It has been shown that the concept of mental fatigue can have great bearing on the test participants in a study. This creates a potential bias that needs to be taken into account when analysing the results of a study. But, evidently, other factors can disrupt the effect of fatigue as shown by Barker and Nussbaum (2011).

From an IIR perspective, one way to handle fatigue is to consider the number of assigned search tasks (e.g. in the form of simulated work task situations) in an IIR studies. That is both with respect to the actual number of simulated work task situations and the complexity/difficulty of the simulated work task situations, which must be balanced to handle possible fatigue. The obligatory pilot testing prior to the main study ought to include a focus on how fatigue is handled.

3.3 The good-subject effect

The relationship between test participants and investigator(s) is a vital part of a test situation, and can potentially affect the outcome (Orne, 1962; Worchel *et al.*, 2000). When test participants knowingly take part in a study, one effect to consider is the good-subject effect. This effect speaks of test participants' willingness to do what they think is desired or expected of them (Nichols and Maner, 2008). The effect can emerge due to multiple reasons. The first reason is a "pleasing effect" seen by test participants aspiring to do well in order to please the investigator(s) by being cooperative. The second reason concerns that test participants do not want to appear ignorant, and therefore are willing to do what they think is expected (Nichols and Maner, 2008). One way to try to neutralise the good-subject effect is by not telling the test participants what the objective of the study is, or by not allowing them to figure it out. This prevents the test participants from systematically adjusting their behaviour and thereby distorting the results in a particular direction (Worchel *et al.*, 2000). Another closely related aspect is the unique relationship between test participants and investigator(s). There will be a certain amount of authority from the investigator's position and obedience can therefore also come into play (Worchel *et al.*, 2000).

There are several examples of studies illustrating the good-subject effect. Orne (1962) described a series of pilot tests that demonstrated the effect by asking test participants to

perform meaningless activities, which they carried out for long periods of time. For example, test participants were presented with a big stack of paper sheets and asked to calculate a large amount of numbers on each paper sheet. When finished, they were to rip the paper sheet into 32 pieces and continue with a new sheet of paper. Or they were presented with an endless stack of paper sheets and asked to continue ripping paper sheets into pieces without any explanation as to why. The test participants followed orders and performed the meaningless activity without asking questions (Orne, 1962, p. 777). The effect has also been illustrated using more extreme settings. A well-known experiment that demonstrates human willingness to follow order even in extreme circumstances is that of Milgram (1963, 1965). Milgram proved that it is possible to make a person administrate electricity going into another human being. The result of the experiment was that 65 per cent of the test participants administrated what they thought was 450 V. The remaining stopped between 300-375 V (Milgram, 1963, 1965). These examples are extreme and are in no way representative of the current study, but they do provide a picture of how powerful the desire to please is.

As illustrated, order effect can also be explained by the good-subject effect. With respect to minimising the good-subject effect in IIR studies, we are partly not to tell the test participants the objective of the study, or not to allow them to figure it out; and partly to acknowledge our perceived authoritative role. Claypool *et al.* (2001, p. 34) comment on the latter when they note how test participants, who are instructed to read articles, do so even if they do not find them interesting. This is a reminder about the issue of interest and the importance to design and tailor the simulated work task situations with great care in order to achieve reliable and realistic search interaction and relevance assessments, because the test participants will do as told and search no matter the quality of simulated work task situations.

4. Methodology

The empirical IIR study was carried out in spring 2012 and involved 20 students from the Royal School of LIS: 13 females and seven males, 21-43 years of age with an average age of 26.5 years. The group of students constitutes a convenience sample. Prior to this, a pilot test was conducted in order to test the tailoring of the three simulated work task situations assigned for searching, the logging equipment, the study protocol and test procedure, and the appropriateness of the collected data. The study was designed as a classic between-groups design. That is a study design where the test participants are randomly assigned to two groups with one group being the treatment group and the other the control group. The treatment group (1) conducted the information searching in a permuted order, and the control group (2) searched for information in a fixed order. Each group consisted of ten test participants. The distribution of gender was as equally divided as possible. The study was conducted in a semi-laboratory setting and the test participants each searched the three simulated work task situations and one self-prepared personal information need. The test participants prepared their personal information need in advance, which was to be on a topic related to leisure hence comparable to the simulated work task situations. For example, one test participant searched for wedding cake recipes, inspiration for wedding cakes, and experiences from making wedding cakes as the test participant was to get married that summer. Several test participants were looking for sightseeing spots, tours, and bi-cycling routes for the upcoming summer vacation. Another test participant was looking for information about how to keep a vegetable garden. The personal information need functions as the baseline of the test participant's search interaction. The simulated work task situations were

designed in accordance with the requirements by Borlund (2000) and were tailored to fit university students (see the list below). The three simulated work task situations used in the study, tailored for university students (translated from Danish to English):

(1) Simulated work task situation A – student discount: as a student you have the possibility of getting student discounts numerous places both on the internet and in your local shops. Therefore you would like to look into whether you are missing out on any of these discount possibilities in your everyday life.

(2) Simulated work task situation B – evening meal: the financial crisis is hard on Denmark and you feel it too. Nevertheless, you need food on the table. When the kitchen cabinets are empty and the money is tight, one must be creative. Consider what is in your kitchen/refrigerator right now and use the Internet to search for inspiration to create a meal on the basis of those ingredients.

(3) Simulated work task situation C – student job: studies show that it is easier to get a job after graduation if you have had a relevant student job alongside your studies. You would therefore like to check out whether there are any available student jobs that can help develop your qualifications, and thereby improve your chances of getting a job when you have earned your degree.

4.1 Data collection methods

In order to investigate if, how, and why order effects occur across the searching of the simulated work task situations, multiple data collection methods were employed. These were: pre-search questionnaire, search interaction logging of the searching, and semi-structured post-search interviews. The methods were employed in the listed order. The purpose of the pre-search questionnaire was solely to gather demographic data and information about search experience in order to describe the test participants and to explain possible cases of unexpected search behaviour. Search interaction logging allowed us to record the interaction between test participant and system during searching with respect to the nine performance parameters (see below). The logging software used was Morae version 3.2.1 (Morae usability testing software from TechSmith, www.techsmith.com/morae.html). The data collection closed with the post-search interview that allowed us to follow-up on behaviour observed during testing, and hereby to provide a deeper understanding of why the participants portrayed certain behaviours.

Nine classic IIR variables were used as performance parameters to measure the potential order effect. The parameters are as follows:

- Time spent searching.
- Number of webpage changes.
- Number of visited websites.
- Number of visited webpages.
- Number of visited webpages per website.
- Number of relevant webpages.
- Number of accumulated queries.
- Number of search terms used per query.
- Number of unique search terms.

The performance parameters were measured at the level of searching per simulated work task situation (A, B, and C) and the test participants' searching of their self-prepared, personal information needs (labelled "Own"). For a definition of the parameters, please consult Table II.

4.2 Test design

As mentioned, the test participants were randomly divided into two groups of ten participants. The test procedure differed only with respect to the order of searching. Table I depicts how group 1 conducted the searches in a permuted order, based on a Latin square design, and how group 2 searched in a fixed order.

All searches were performed online and there was no time restriction. The test participants were instructed to search for as long as it would take to satisfy the perceived information need. The test participants took part one at a time. Internet Explorer was the chosen browser, because it is the preferred browser of Morae. Further, the same browser ensures consistency and comparability of the collected data. The test participants were given free range to use any search engine of their preference. It happened that all, with no exception, searched via Google if not entering the specific URL directly.

4.3 Data analysis

The search interaction logs were analysed on the basis of the video recordings provided by Morae. Table II provides an overview of the collected data and informs how the performance parameters are defined and managed.

In addition to the automatic counts of webpage changes, Morae supplies a list of all visited webpages in chronological order. This list was used to double-check the manually counted webpages and websites.

The search interactions measured by the nine performance parameters are reported as descriptive statistics (in Section 5). Albeit, there is a common understanding and practice that data collected via this type of test design with test participants randomly assigned to control and treatment groups, even for restricted samples, can be analysed as inference statistics, this is incorrect. Fact is, our sample is too small to fulfil the requirements for valid and powerful inference statistic conclusions and hence these statistics are not applied (e.g. Schneider, 2013).

Participants	Permuted search order of group 1	Fixed search order of group 2
1	A B C Own	A Own B C
2	Own A B C	A Own B C
3	C Own A B	A Own B C
4	B C Own A	A Own B C
5	A Own C B	A Own B C
6	B A Own C	A Own B C
7	C B A Own	A Own B C
8	Own C B A	A Own B C
9	B Own A C	A Own B C
10	C A B Own	A Own B C

Table I.
Search order of group 1 and 2

Notes: A, simulated work task situation on student discount; B, simulated work task situation on preparing an evening meal; C, simulated work task situation on student job; and Own, test participant's personal self-prepared information need

Performance parameters	Definition	Method of management
Extent of time spend during the searching of A, B, C, and Own	Time is estimated in minutes	Manual indication of start and finish for the individual searching of A, B, C, Own is made by Morae
Number of webpage changes made during the searching of A, B, C, and Own	Number of times a shift between webpages has been made. Independent of how many times the same page was visited	Counted automatically by Morae
Number of visited websites during the searching of A, B, C, and Own	A website is defined by a collection of webpages confined in an overall domain	Counted manually
Number of visited webpages during the searching of A, B, C, and Own	A webpage is defined as a unique URL. Each website is only counted once. Independent of how many times it has been visited	Counted manually
Number of visited webpages per website during the searching of A, B, C, and Own	An average of all webpages over all websites visited during the searching	Counted manually and calculate as a fraction of the number of webpages over the number of websites
Number of relevant webpages assessed during the searching of A, B, C, and Own	Webpages saved as "favourites" in the Internet Explorer browser by the test participant as relevant to the perceived information need. With no regard to the degree of relevance	Counted manually
Number of queries accumulated during the searching of A, B, C, and Own	Includes all queries generated, both in general search engines and search engines on specific sites	Counted manually
Number of search terms used per query during the searching of A, B, C, and Own	All individual words are counted as search terms	Counted manually and calculated as a fraction of the number of individual search terms over the total number of queries
Number of unique search terms used during the searching of A, B, C, and Own	All individual words are counted as search terms	Counted manually

Table II. Definition and methods of management of the performance parameters

4.3.1 Quantitative data analysis. The interviews were focused and concerned the perception of the simulated work task situations and the personal self-prepared information needs with respect to interest, relevancy, motivation during searching, and realism of search behaviour. The interviews were an average length of nine minutes. The interviews were transcribed and uploaded to MAXQDA (MAXQDA: Qualitative Data Analysis Software, 1995-2011, www.maxqda.com) for handling and structuring.

5. Results

The presentation of results starts with a brief comparison of the test participants' search behaviour of their own information needs and the simulated work task situations A, B, and C. Hereafter, the results from the logged data and the interviews are presented.

5.1 Baseline

The establishing of a baseline allows us to compare how realistic the test participants' search and interaction of the searching of the assigned simulated work task situations

are to that of their own information needs. That is, whether the search behaviour of the simulated work task situations can be taken as an indication of the test participants' natural search behaviour. Table III presents the aggregated mean and standard deviation values of the performance parameters for the four searches (A, B, C, and Own).

At an overall level, the performance parameters depict an agreeable search behaviour of A, B, C, and Own that makes us rely on the search behaviour achieved from the searching of the simulated work task situations. However, there are minor deviations on different performance parameters independently of whether we are looking at the searching of the personal information needs (Own) or any of the simulated work task situations. When focusing on individual performance parameters relating to the searching of the personal information needs (Own), the test participants have spent the most time and have performed the fewest queries on this searching. The time spent might be because of interest in their own information need (e.g. Borlund *et al.*, 2012), and few queries might be a result of knowledge about the search topic in question.

Of the remaining performance parameters, the searching of the simulated work task situations do not vary markedly compared to the searching of the personal information needs (Own). That is, with the exception of simulated work task situation B (evening meal) where the parameters of search terms per query and unique search terms are considerably higher than the others due partly to the variety of ingredients searched for (e.g. see Table V), and partly the extremely dedicated search behaviour of particularly test participant no. 1, group 1 (for an elaboration see Section 5.2). In other words, a condition built in the simulated work task situation B. Based on this, we feel comfortable that the search behaviour of the simulated work task situations reflects natural and realistic searching of the test participants and hence is reliable.

5.2 The performance parameters

This sub-section presents the results of the performance parameters used to measure the four searches by the ten test participants in each of the two groups. The results are presented in Tables IV and V. Table IV depicts the results with respect to the order of searching referred to as search 1, 2, 3, and 4, hereby not distinguishing between the simulated work task situations A, B, and C or the personal information need (Own). The purpose of Table IV is to visualise and highlight possible order effects of group 2

Table III. Comparison of performance parameters of personal and simulated information needs

	Simulated work task situation A (student discount)		Simulated work task situation B (evening meal)		Simulated work task situation C (student job)		Personal information need (Own)	
Performance parameters	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Minutes	15.6	6.9	15.1	4.9	18.5	8.7	22.3	7.4
Webpage changes	59.7	30.6	42.8	15.0	64.0	36.1	58.0	32.0
Websites	11.2	4.4	8.2	3.3	12.7	5.8	13.0	8.8
Webpages	35.4	18.4	25.9	8.2	41.3	21.8	37.3	19.4
Webpages per website	3.3	1.2	3.4	1.0	3.4	1.2	3.5	1.7
Relevant webpages	5.6	5.5	3.7	2.3	4.0	2.5	4.7	2.7
Number of queries	8.2	4.5	8.5	3.1	10.2	6.3	7.1	5.0
Search terms per query	2.1	0.8	5.0	5.1	2.6	2.1	2.3	1.1
Unique search terms	10.8	7.8	23.8	18.9	12.1	10.8	9.4	8.6

Performance parameters	Group	Search 1 Mean	Search 1 SD	Search 2 Mean	Search 2 SD	Search 3 Mean	Search 3 SD	Search 4 Mean	Search 4 SD
Minutes	Group 1	17.1	8.7	19.1	7.0	19.6	9.4	19.2	4.8
	Group 2	14.1	8.0	20.9	8.0	14.7	6.1	18.2	7.3
	Group 1 and 2	15.6	8.3	20.0	7.4	17.2	8.1	18.7	6.1
Webpage changes	Group 1	50.6	27.7	50.1	20.9	50.5	30.0	59.3	25.1
	Group 2	59.3	38.7	60.9	37.9	48.7	16.0	69.3	39.6
	Group 1 and 2	55.0	33.1	55.5	30.3	49.6	23.4	64.3	32.7
Websites	Group 1	10.2	6.2	8.6	4.8	11.1	5.3	11.2	5.8
	Group 2	10.6	4.8	14.6	10.3	9.5	3.7	14.2	5.7
	Group 1 and 2	10.4	5.4	11.6	8.4	10.3	4.5	12.7	5.8
Webpages	Group 1	30.8	16.1	30.3	16.2	29.9	15.1	35.1	12.4
	Group 2	37.1	23.8	40.7	23.0	30.1	9.1	45.7	24.2
	Group 1 and 2	34.0	20.1	35.5	20.1	30.0	12.2	40.4	19.5
Webpages per website	Group 1	3.4	1.2	3.7	1.0	3.0	1.6	3.5	1.2
	Group 2	3.6	1.5	3.1	1.3	3.4	1.1	3.4	1.7
	Group 1 and 2	3.5	1.3	3.4	1.2	3.2	1.4	3.4	1.4
Relevant webpages	Group 1	4.1	3.6	2.9	1.4	4.8	2.3	3.1	1.7
	Group 2	6.3	7.3	6.1	2.8	4.6	2.7	3.9	2.2
	Group 1 and 2	5.2	5.7	4.5	2.7	4.7	2.5	3.5	2.0
Number of queries	Group 1	6.8	3.6	8.7	5.3	8.9	6.8	12.3	4.1
	Group 2	7.1	3.8	5.8	2.9	8.8	3.5	9.4	6.4
	Group 1 and 2	7.0	3.6	7.3	4.4	8.9	5.3	10.9	5.4
Search terms per query	Group 1	3.3	4.3	4.6	6.2	2.1	1.1	2.9	1.4
	Group 2	2.1	0.6	2.2	0.9	3.8	1.8	3.1	2.8
	Group 1 and 2	2.7	3.1	3.4	4.5	3.0	1.7	3.0	2.1
Unique search terms	Group 1	13.6	16.5	19.5	23.5	10.9	8.0	17.3	9.9
	Group 2	8.7	6.0	7.5	4.6	20.2	12.0	13.6	14.0
	Group 1 and 2	11.2	12.4	13.5	17.6	15.6	11.0	15.5	12.0

Table IV.
Order of searches measured against nine performance parameters

Table V.
The specific searches measures against nine performance parameters

Performance parameters	Group	Simulated work task situation A (student discount) Mean	SD	Simulated work task situation B (evening meal) Mean	SD	Simulated work task situation C (student job) Mean	SD	Personal information need (Own) Mean	SD
Minutes	Group 1	17.1	5.7	15.4	3.6	18.8	10.3	23.7	6.9
	Group 2	14.1	8.0	14.7	6.1	18.2	7.3	20.9	8.0
	Group 1 and 2	15.6	6.9	15.1	4.9	18.5	8.7	22.3	7.4
Webpage changes	Group 1	60.0	21.7	36.8	11.7	58.6	33.3	55.1	26.5
	Group 2	59.3	38.7	48.7	16.0	69.3	39.6	60.9	37.9
	Group 1 and 2	59.7	30.6	42.8	15.0	64.0	36.1	58.0	32.0
Websites	Group 1	11.7	4.2	6.9	2.4	11.2	5.8	11.3	7.3
	Group 2	10.6	4.8	9.5	3.7	14.2	5.7	14.6	10.3
	Group 1 and 2	11.2	4.4	8.2	3.3	12.7	5.8	13.0	8.8
Webpages	Group 1	33.7	12.0	21.7	4.3	36.9	19.2	33.8	15.6
	Group 2	37.1	23.8	30.1	9.1	45.7	24.2	40.7	23.0
	Group 1 and 2	35.4	18.4	25.9	8.2	41.3	21.8	37.3	19.4
Webpages per website	Group 1	3.0	0.9	3.4	1.0	3.4	0.6	3.8	2.0
	Group 2	3.6	1.5	3.4	1.1	3.4	1.7	3.1	1.3
	Group 1 and 2	3.3	1.2	3.4	1.0	3.4	1.2	3.5	1.7
Relevant webpages	Group 1	4.8	3.2	2.8	1.5	4.0	2.9	3.3	1.7
	Group 2	6.3	7.3	4.6	2.7	3.9	2.2	6.1	2.8
	Group 1 and 2	5.6	5.5	3.7	2.3	4.0	2.5	4.7	2.7
Number of queries	Group 1	9.2	5.1	8.1	2.8	11.0	6.5	8.4	6.3
	Group 2	7.1	3.8	8.8	3.5	9.4	6.4	5.8	2.9
	Group 1 and 2	8.2	4.5	8.5	3.1	10.2	6.3	7.1	5.0
Search terms per query	Group 1	2.2	0.9	6.1	6.9	2.1	1.1	2.4	1.3
	Group 2	2.1	0.6	3.8	1.8	3.1	2.8	2.2	0.9
	Group 1 and 2	2,1	0.8	5.0	5.1	2.6	2.1	2.3	1.1
Unique search terms	Group 1	12.9	9.0	26.7	24.2	10.5	6.7	11.2	11.3
	Group 2	8.7	6.0	20.2	12.0	13.6	14.0	7.5	4.6
	Group 1 and 2	10.8	7.8	23.5	18.9	12.1	10.8	9.4	8.6

compared to group 1. Table V shows the results explicitly for the searched simulated work task situations (A, B, and C) and the personal information needs (Own) in order to point out whether possible order effects may be due to the searching of A, B, C, or Own.

Table IV shows that there is a larger deviation among the number of "minutes" measured as time spent on the individual searches in group 2 (range: 14.1-20.9 minutes), compared to group 1 where the time spent searching is more similar between the four searches (range: 17.1-19.6 minutes). The latter may be an illustration of the effect of permutation with reference to neutralisation, and hence an indirect illustration of order effect of group 2. In comparison, Table V shows how the figures reporting time spent are relatively similar across the simulated work task situations hence indicating they are not influential of time spent. In both groups, the most time spent is on searching the personal information need (Own) followed by C, B, and A.

When looking at the parameter "webpage changes" (Table IV), both groups have performed the most webpage changes on the final search 4 compared to the other three searches – group 1 59.3 and group 2 69.3 webpage changes. Table I depicts how the final search of group 1 is constituted of 2 searchers of simulated work task situation A, 2 searches of simulated work task situation B, 3 searches of simulated work task situation C, and 3 searchers of the personal information need (Own) compared to the searching of the simulated work task situation C (student job) of group 2. This indicates that the order of searching has an effect on the number of webpage changes. So does the comparison of group 1 and 2 (Table IV), which exhibits larger amounts of variation of webpage changes between the searches of group 2 (range: 48.7-69.3) with searched order compared to group 1 where the searches were permutated (range: 50.1-59.3). From Table V we learn how especially the simulated work task situation B (evening meal) has caused notably fewer webpages changes for both groups, though this does not explain the patterns of Table IV.

With respect to number of visited "websites", Table IV shows how search 2 in group 1 stands out with fewer websites being visited (8.6) in contrast to group 2, where more websites are visited during the searching of search 2 (14.6). In group 2, search 4 (simulated work task situation C (student job)) reveals similar numbers of visited websites to that of search 2 (personal information needs, (Own)). With the exception of search 2, group 1 demonstrates relatively equal numbers of visited websites for the remaining three searches. In Table V, group 1 shows a similar pattern, here with the simulated work task situation B (evening meal) as the outsider with an average of 6.9 visited websites compared to 11.7, 11.2, and 11.3 of the simulated work task situations A, C, and the personal information needs (Own), respectively. When comparing group 1 and 2 in Table V, it becomes apparent that order effects takes place seen by the varying numbers of visited websites of group 2 compared to those of group 1.

In regards to the performance parameter of "webpages" visited we see in Table IV a similar trend to that of the parameter webpage changes in that both groups have visited most webpages on the final search 4 compared to the other three searches. Further, Table V informs us that both groups have made more webpage visits in the searching of the simulated work task situation C (students job), which is the final search of group 2. But it does not entirely explain the result of group 1 with the simulated work task situation C being searched only three times out of ten.

When looking at the performance parameter of "webpages per website" with reference to search order (Table IV), we see minor variations for both groups ranging from 3.0-3.7 for group 1 and 3.1-3.6 for group 2. Similarly, Table V shows no obvious differences when comparing the two groups with respect to the specific searching of the simulated work task situations (A, B, and C) and the personal information needs (Own).

Generally, group 2 identified more "relevant webpages" compared to group 1 (Tables IV and V). Group 2 further assessed the most relevant for search 1 (6.3) followed in decreasing order by search 2 (6.1), search 3 (4.6), and search 4 (3.9) (Table IV). These are the most distinct patterns in regards to the performance parameter of relevant webpages. Table V shows how the simulated work task situation A (student discount) – first search of group 2 – results in the highest amount of relevant webpages from both group 1 and 2 with 4.8 and 6.3 relevant webpages, respectively. Group 2 assessed the fewest relevant webpages with respect to the simulated work task situation C (student job) with 3.9, which is very similar to the 4.0 of group 1. Group 1 assessed the fewest relevant webpages with respect to the simulated work task situation B (evening meal) with 2.8 compared to 4.6 of group 2. Finally, group 1 assessed 3.3 relevant webpages in response to the personal information needs (Own) in contrast to 6.1 of group 2.

Table IV depicts how both groups formulated the highest "number of queries" for the final search 4 (Table IV). Group 1 formulated the highest number with 12.3 queries compared to 9.4 of group 2. Further, we see an interesting and unexplainable pattern within group 1 (Table IV), in that the number of queries increases during searching with 6.8 in search 1, 8.7 in search 2, 8.9 in search 3, and 12.3 in search 4. Table V (group 1) shows that the simulated work task situations A (student discount) and C (student job) are top scores with for 9.2 and 11.0 query formulations succeeded by the personal information needs (Own) with 8.4 and simulated work task situation B (evening meal) with 8.1 query formulations.

The performance parameters of "search terms per query" and total number of "unique search terms" present no obvious patterns with the exception of group 1 using relatively more search terms per query (4.6) and a much higher number of unique search terms (19.5) (Table IV) for the second search compared to group 2. While group 2 uses more search terms per query (3.8) and more unique search terms (20.2) on search 3 (Table IV). At first glance it appears to be caused by the simulated work task situation B (evening meal) (Table V). No doubt the invitation to search on food ingredients stimulated the use of a higher number of search terms, as also mention in Section 5.1. However, with respect to group 1, it is mainly due to the search formulations of three test participants (test participant no.1, 2, and 7) and in particular the extreme number of search terms employed by test participant 1. Test participant no. 1 and 7 searched the simulated work task situation B as their second search as depicted in Table I, and test participant no. 2 searched the simulated work task situation A (student discount) as search 2. Test participant no. 1 and 7 used each on average 22.1 and 3.7 search terms per query and a total of 82 and 22 unique search terms, respectively. Test participant no. 7 used in the searching on simulated work task situation A an average of 3.7 search terms per query and a total of 29 unique search terms.

The logged data revealed order effect on three out of nine performance parameters. This is manifested as an increase in activity on the last performed search. The test participants performed more webpage changes, visited more webpages, and formulated more queries on the final search. The data also indicated that the nature of simulated work task situations and the personal information needs (Own) might influence the searching. Further, the data reflected individual search behaviour and preferences of the test participants. This leads us to the qualitative data gathered from the complementary interviews.

Overall, the test participants conveyed a relaxed and comfortable approach to the test situation. Though some of the test participants expressed that they had been

nervous prior to the test and ascribed this to uncertainty as to what was expected of them. But they also said that they quickly became relaxed, and they felt they exhibited a more natural behaviour. They confirmed to have searched as they usually would when searching for information. The test participants perceived the simulated work task situations as realistic to their current situation as university students by being relevant and of personal interest to them. The simulated work task situations A (student discount) and C (student job) were generally preferred over the simulated work task situation B (evening meal), but also favoured by others. For example test participant no. 4, who explained she liked C because she always tries to use everything in the refrigerator to avoid food waste and to save money. Similarly, test participant no. 16 and 20 liked simulated work task C because it resembled their own personal information needs on meal plans and recipes.

The test participants expressed unwanted attention towards the aspect of time. As mentioned in the methodology section (sub-section 4.2) there was no time restriction with respect to searching. Nevertheless, the test participants still focused on time, which informs us about their attention to the fact that they are taking part in a test. When asked to further explain their uncertainty, they expressed a concern of appearing ignorant or doing something wrong, hereby illustrating the good-subject effect.

In relation to motivation, the test participants stated multiple factors that affect their motivation in a positive or negative direction. This is not surprising, as described in sub-section 3.1, motivation is a complex concept. The test participants state interest, relevance, topicality, frustration, novelty, and the goal of searching as factors that affect their motivation. The test participants also expressed motivation in relation to the entire process, hence across the searches. The test participants explained that they made an effort even when tired and bored with the test situation.

In relation to fatigue, several of the participants expressed that the intense searching for information for long periods of time made them tired and affected their motivation. This is, however, not visible from the log data.

As indicated in the previous sections (sub-sections 3.1-3.3), the significance of motivation and the existence of the fatigue and good-subject effect affect the behaviour of the test participants and have strong correlations. This is also the case in our study with interacting variables that affect each other. For example, it may appear that the effect of fatigue is hard to document because it is overruled by the desire to make an effort, caused by the good-subject effect.

6. Discussions

The aim of the study has been to identify and attempt to explain why order effects occur in IIR evaluations. Results of the empirical study show the test participants made more website changes, visited more webpages, and formulated more queries on the final search. There is no indication of similar clear patterns with respect to the remaining performance parameters. Characteristics of the individual searches (A, B, C, and Own) seem to have a certain degree of influence on the performance parameters. Further, we cannot eliminate that both characteristics and presentation order interact in weakening or strengthening the test participants' performance.

From the interviews, we know some of the participants expressed that they were nervous prior to, and in the beginning of, the test because of uncertainty about the process. They were concerned about appearing ignorant or doing something wrong. In other words, they demonstrate the good-subject effect (Orne, 1962; Worchel *et al.*, 2000). As described in sub-section 3.3 the good-subject effect can result in modification

of behaviour of the test participants who would like to appear knowledgeable and want to please the investigator(s). The test participants stated that they made an effort even when tired and bored with the test situation. The desire to please couples the good-subject effect to motivation.

Motivation was another influential aspect during testing. The test participants explained how interest, relevance, topicality, frustration, and the goal of searching were factors that affected their motivation. These concepts are portrayed in the literature as important factors in relation to motivation (e.g. Latham and Locke, 1979, Locke and Latham, 2006; Schiefle, 1998). Also novelty affects motivation (Smith, 1981). Novelty was expressed when a new information search of either A, B, C, or Own was to take place, and the test participants explained they gained renewed energy from this. At the same time fatigue was present as well. As previously noted, several of the test participants expressed that they became tired during testing and it decreased their motivation. This led us to anticipate a decrease in activity as the test proceeded, but the exact opposite happened. This is a similar trend as the one seen in the study by Barker and Nussbaum (2011), though not dealing with nurses in the emergency room. In our case, an increase in activity occurred on the final search with respect to three out of nine performance parameters. We have not been able to demonstrate an effect of fatigue with respect to the performance parameters. In line with other studies, the effect does not appear (e.g. Barker and Nussbaum, 2011; Holding, 1983). Though one could assume that the high number of queries of search 4 could be due to misspellings of query terms, this is not the case (e.g. van der Linden *et al.*, 2003). This means that fatigue is present, but the effect of it is absent. There is no apparent answer to why the increase in activity occurs. As described previously, fatigue can be put out of play due to excitement. One may speculate that the effect of fatigue is simply overruled by motivation caused by novelty, interest, or the case of testing. The effect of the good-subject may also have an impact. The test participants might simply have pulled themselves together in the end of the study in order to finish on a good note. Statements made by the test participants when asked directly whether they thought a different presentation order would have changes their behaviour were contradictive. This further indicates that potential effects are individual, and may make one test participant use more time, search harder, and more dedicated compared to another, thereby shading potential effects.

7. Concluding remarks

The study revealed order effect on three out of nine performance parameters, manifested as an increase in activity on the final search. The interviews showed that many variables were in play, which interacted and affected the test participants' attitude, effort, and performance. In general, the patterns were weak and contradictive, and we were unable to separate the variables that caused the changes in performance, and hence not able to explain those changes and patterns. In line with past research on order effect, we do also conclude that further research is required (e.g. Eisenberg and Barry, 1988; Parker and Johnson, 1990; Xu and Wang, 2008). For example, a larger sample of test participants ought to provide clearer patterns of search behaviour. Further research is also needed in order to shed light on the impact order effect has on IIR studies with reference to the knowledge base generated in these studies. The phenomenon of order effect is not yet fully understood, and therefore the requirement to permute search tasks in IIR studies stands (e.g. Borlund, 2003; Kelly, 2009; Tague-Sutcliffe, 1992).

References

Bar-Ilan, J., Keenoy, K., Levene, M. and Yaari, E. (2009), "Presentation bias is significant in determining user preference for search results – a user study", *Journal of the American Society for Information Science and Technology*, Vol. 60 No. 1, pp. 135-149.

Barker, L.M. and Nussbaum, M.A. (2011), "The effects of fatigue on performance in simulated nursing work", *Ergonomics*, Vol. 54 No. 9, pp. 815-829.

Belkin, N.J. (1977), "Internal knowledge and external information", in Belkin, N.J., De Mey, M., Pinxten, R., Poriau, M. and Vandamme, F. (Eds), *International Workshop on the Cognitive Viewpoint*, University of Ghent, Ghent, pp. 187-194.

Belkin, N.J. (1980), "Anomalous states of knowledge as the basis for information retrieval", *Canadian Journal of Information Science*, Vol. 5, pp. 133-143.

Borlund, P. (2000), "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 27 April 2015).

Borlund, P., Dreier, S. and Byström, K. (2012), "What does time spent on searching indicate?", *Proceedings of the 4th Sympoium on Information Interaction in Context Symposium (IIiX' 12). Association for Computing Machinery*, pp. 184-193.

Brookes, B.C. (1977), "The foundation of information science", in Belkin, N.J., De Mey, M., Pinxten, R., Poriau, M. and Vandamme, F. (Eds), *International Workshop on the Cognitive Viewpoint*, University of Ghent, Ghent, pp. 195-203.

Brunel, F.F. and Nelson, M.R. (2003), "Message order effects and gender differences in advertising persuasion", *Journal of Advertising Research*, Vol. 43 No. 3, pp. 330-341.

Clancy, K.J. and Wachsler, R.A. (1971), "Positional effects in shared-cost surveys", *The Public Opinion Quarterly*, Vol. 35 No. 2, pp. 258-265.

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

Crawford, L.E., Luka, B. and Cacioppo, J.T. (2002), "Social behavior", in Pashler, H. and Gallistel, R. (Eds), *Stevens' Handbook of Experimental Psychology. Volume 3: Learning, Motivation, and Emotion*, John Wiley & Sons Inc, New York, NY, pp. 737-799.

Cushing, B.E. and Ahlawat, S.S. (1996), "Mitigation of recency bias in audit judgment: the effect of documentation", *Auditing*, Vol. 15 No. 2, pp. 110-122.

Duffy, B. (2003), "Response order effects: how do people read?", *International Journal of Market Research*, Vol. 45 No. 4, pp. 457-466.

Eisenberg, M. (1986), "Magnitude estimation and the measurement of relevance", PhD thesis, Syracuse University, Syracuse, NY.

Eisenberg, M. and Barry, C. (1988), "Order effects: a study of the possible influence of presentation order on user judgments of document relevance", *Journal of the American Society for Information Science*, Vol. 39 No. 5, pp. 293-300.

Franken, R.E. (2002), *Human Motivation*, 5th ed., Wadsworth/Thomson Learning, Belmont, CA.

Gibson, C.O., Shapiro, G.M., Murphy, L.R. and Stanko, G.J. (1978), "Interaction of survey questions as it relates to interviewer-respondent bias", *Proceedings of the Survey Research Methods Section*, pp. 251-256.

Hidi, S. (2000), "An interest researcher's perspective: the effects of extrinsic and intrinsic factors on motivation", in Sansone, C. and Harackiewicz, J.M. (Eds), *Intrinsic and Extrinsic Motivation: The Search for Optimal Motivation and Performance*, Academic Press, San Diego, CA, pp. 309-372.

Hidi, S. and Baird, W. (1986), "Interestingness: a neglected variable in discourse processing", *Cognitive Science*, Vol. 10 No. 2, pp. 179-194.

Hogarth, R.M. and Einhorn, H.J. (1992), "Order effects in belief updating: the belief-adjustment model", *Cognitive Psychology*, Vol. 24 No. 1, pp. 1-55.

Holding, D.H. (1983), "Fatigue", in Hockey, R. (Ed.), *Stress and Fatigue in Human Performance*, John Wiley & Sons, Chichester, pp. 145-167.

Huang, M. and Wang, H. (2004), "The influence of document presentation order and number of documents judged on user's judgments of relevance", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 11, pp. 970-979.

Kantowitz, B.H., Roediger, H.L. III and Elmes, D.G. (2001), *Experimental Psychology. Understanding Psychological Research*, 7th ed., Brooks/Cole/Thomson Learning, Belmont, CA.

Kardes, F.R. and Kalyanaram, G. (1992), "Order-of-entry effects on consumer memory and judgment: an information integration perspective", *Journal of Marketing Research*, Vol. 29 No. 3, pp. 343-357.

Kelly, D. (2009), "Methods for evaluating interactive information retrieval systems with users", *Foundations and Trends in Information Retrieval*, Vol. 3 Nos 1-2, pp. 1-224.

Krapp, A. (1999), "Interest, motivation and learning: an educational-psychological perspective", *European Journal of Psychology of Education*, Vol. 14 No. 1, pp. 23-40.

Krosnick, J.A. and Alwin, D.F. (1987), "An evaluation of a cognitive theory of response-order effects in survey measurement", *Public Opinion Quarterly*, Vol. 51 No. 2, pp. 201-219.

Laird Landon, E. Jr (1971), "Order bias, the ideal rating, and the semantic differential", *Journal of Marketing Research*, Vol. 8 No. 3, pp. 375-378.

Latham, G.P. and Locke, E.A. (1979), "Goal setting: a motivational technique that works", *Organizational Dynamics*, Vol. 8 No. 2, pp. 68-80.

Locke, E.A. and Latham, G.P. (2006), "New directions in goal-setting theory", *Current Directions in Psychological Science*, Vol. 15 No. 5, pp. 265-268.

Milgram, S. (1963), "Behavioral study of obedience", *Journal of Abnormal and Social Psychology*, Vol. 67 No. 4, pp. 371-378.

Milgram, S. (1965), "Some conditions of obedience and disobedience to authority", *Human Relations*, Vol. 18 No. 1, pp. 57-76.

Monroe, G.S. and Ng, J. (2000), "An examination of order effects in auditors' inherent risk assessments", *Accounting and Finance*, Vol. 40 No. 2, pp. 153-168.

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.

Orne, M.T. (1962), "On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications", *American Psychologist*, Vol. 17 No. 11, pp. 776-783.

Parker, L.M. and Johnson, R.E. (1990), "Does order of presentation affect users judgment of documents?", *Journal of the American Society for Information Science*, Vol. 41 No. 7, pp. 493-494.

Perreault, W.D. (1975), "Controlling order-effect bias", *The Public Opinion Quarterly*, Vol. 39 No. 4, pp. 544-551.

Ryan, R.M. and Deci, E.L. (2000), "Intrinsic and extrinsic motivations: classic definitions and new directions", *Contemporary Educational Psychology*, Vol. 25 No. 1, pp. 54-67.

Schiefle, U. (1998), "Individual interest and learning: what we know and what we don't know", in Hoffmann, L., Krapp, A., Renninger, K.A. and Baumert, J. (Eds), *Interest and Learning*, Institut für die Pädagogik der Naturwissenschaften an der Universität Kiel, Kiel, pp. 91-101.

Schneider, J.W. (2013), "Caveats for using statistical significance tests in research assessments", *Journal of Informetrics*, Vol. 7 No. 1, pp. 50-62.

Smith, R.P. (1981), "Boredom: a review", *Human Factors*, Vol. 23 No. 3, pp. 329-340.

Tague-Sutcliffe, J. (1992), "The pragmatics of information retrieval experimentation, revisited", *Information Processing & Management*, Vol. 28 No. 4, pp. 467-490.

van der Linden, D., Frese, M. and Sonnentag, S. (2003), "The impact of mental fatigue on exploration in a complex computer task: rigidity and loss of systematic strategies", *Human Factors*, Vol. 45 No. 3, pp. 483-494.

Worchel, S., Cooper, J., Goethals, G.R. and Olson, J.M. (2000), *Social Psychology*, Wadsworth, Belmont, CA.

Xu, Y. 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.

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